INTERNATIONALE HOCHSCHULE

Masterarbeit – ISWM01

an der IU Internationale Hochschule

im Wintersemester 2025/26

zum Thema:

Kontexttreue maschinelle Übersetzung alttestamentlicher Texte

Eine Analyse der Leistungsfähigkeit neuronaler Übersetzungssysteme bei der Verarbeitung altgriechischer und althebräischer Bibeltexte am Beispiel ausgewählter Perikopen aus dem Alten Testament.

vorgelegt von	Josie Baldeweg Steingasse 14a 91094 Langensendelbach josie.baldeweg@iu-study.org
Studiengang	Informatik (M. Sc.) 3. Fachsemester IU14115908
Abgegeben	Montag, 19. Januar 2026
Betreuerin	Dr.-Ing. Maja Popović

ABSTRACT

This thesis examines the context-faithfulness of contemporary neural translation systems when applied to Old Testament texts, with context understood not only as lexical and syntactic adequacy but also as hermeneutical and theological interpretability. The study addresses how translation quality differs between classical neural machine translation (NMT) systems and large language model (LLM) approaches, and which model characteristics systematically affect structural fidelity, semantic precision, and the preservation of interpretive openness that is often constitutive for biblical texts. Methodologically, the thesis employs a triangulated evaluation design that combines quantitative and qualitative approaches. Quantitative assessment uses reference-based similarity measures and additional structural indicators to capture expansion tendencies, omissions, and non-source-grounded additions, thereby describing how closely outputs track the form and information structure of the source. Qualitative analysis complements this by focusing on genre-sensitive phenomena and the handling of theologically salient expressions, assessing whether model outputs remain compatible with philological and hermeneutical standards. To avoid restricting evaluation to modern translation norms alone, the framework includes diachronic reference points as interpretive checkpoints alongside established contemporary translations. Results show that translation behaviour is strongly genre-dependent: narrative texts tend to yield more stable and formally aligned outputs, whereas prophetic and especially poetic texts increase divergence, prompting more frequent restructuring and interpretive explicitation. Across models, quantitative scores prove to be informative for identifying relative tendencies and error profiles, but insufficient as stand-alone quality indicators, since high formal fluency can coincide with hermeneutically consequential shifts, and certain deviations reflect necessary target-language explicitation rather than straightforward error. At the level of model classes, NMT systems generally exhibit greater structural consistency, whereas LLM-based systems display higher variability and a stronger propensity toward explicative, interpretive renderings. Overall, the thesis concludes that context-faithful translation in a theological sense cannot be guaranteed by any single system type without methodological safeguards. However, when deployed transparently and under hermeneutical supervision, AI-based translation can support theological research – particularly for orientation, comparative analysis, and hypothesis generation – if reproducibility, documentation, and critical source-text control remain central.

Keywords

Neural Machine Translation | Large Language Models

Hebrew Bible | Septuagint | Vulgate

Hermeneutics | Theology

INHALTSVERZEICHNIS

ABBILDUNGS- UND TABELLENVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

AcI ..... *Accusativus cum infinitivo*
API ..... *Application Programming Interface*
BERT ..... *Bidirectional Encoder Representations from Transformers*
BHQ ..... *Biblia Hebraica Quinta*
BHS ..... *Biblia Hebraica Stuttgartensia*
BLEU ..... *Bilingual Evaluation Understudy*
BOS ..... *Beginning-of-Sentence*
BPE ..... *Byte-Pair-Encoding*
chrF ..... *character F*
DeepL ..... *Deep Learning*
EOS ..... *End Of Sequence*
GPT ..... *Generative Pre-trained Transformer*
GPU ..... *Graphics Processing Unit*
gRPC ..... *Google Remote Procedure Calls*
HTTP ..... *Hypertext Transfer Protocol*
IBM ..... *International Business Machines Corporation*
KI ..... *Künstliche Intelligenz*
KJV ..... *King James Version*
LLaMA ..... *Large Language Model Meta AI*
LLM ..... *Large Language Model*
LoRA ..... *Low-Rank Adaption*
LXX ..... *Septuaginta - "Übersetzung der Siebzig"*
MoE ..... *Mixture-of-Expoerts-Ansatz*
MT ..... *Machine Translation*
NLLB ..... *No Language Left Behind*
NLP ..... *Natural Language Processing*
NLT ..... *New Living Translation*
NMT ..... Neural Machine Translation
NTG ..... Novum Testamentum Greace
PSO ..... *Prädikat – Subjekt – Objekt*
RBMT ..... *Rule-Based Machine Translation*
RNN ..... *Rekurrentes neuronales Netz*
RoPE ..... *Rotary Position Embedding*
SDK ..... *Software Development Kit*
SMT ..... *Statistical Machine Translation*
SPO ..... *Subjekt – Prädikat – Objekt*
TFR ..... *Token Fragmentation Rate*
UTF ..... *Unicode Transformation Format*

1 Einleitung

1.1 Motivation und Relevanz des Themas

Biblische Texte sind nicht nur historische Quellen, sondern zugleich theologische Deutungstexte, deren Aussagegehalt wesentlich über Sprache vermittelt wird. Gerade im Alten Testament sind zentrale Begriffe, Metaphern und syntaktische Strukturen eng an die Eigenlogik des biblischen Hebräisch gebunden und in ihrer Wirkungsgeschichte oft durch frühere Übersetzungen geprägt. Hinzu kommt, dass das Alte Testament kein monolithischer Urtext ist, sondern in einer komplexen Überlieferungssituation vorliegt, die verschiedene Textzeugen und Sprachfassungen einschließt. Bereits die maßgeblichen wissenschaftlichen Editionen basieren auf konkreten Handschriften und Traditionslinien. Zugleich zeigen frühe Übersetzungen wie die Septuaginta, dass Übersetzung nicht nur sprachliche Übertragung, sondern immer auch interpretierende Auswahl und Kontextualisierung ist. Parallel zu diesen philologischen und hermeneutischen Anforderungen hat sich maschinelle Übersetzung in den letzten Jahren von regelbasierten Verfahren hin zu neuronalen Systemen entwickelt und ist durch große Sprachmodelle (LLMs) nochmals in der Breite verfügbar geworden. Für viele Forschungskontexte ist es dadurch naheliegend geworden, maschinelle Übersetzungen als „erste Annäherung“ einzusetzen – etwa zur Orientierung in Texten, zur schnellen Erschließung von Varianten oder zur Vorbereitung weiterer Analyseschritte. Gerade hier entsteht jedoch ein Spannungsfeld: Biblische Texte sind häufig mehrdeutig formuliert, arbeiten mit dichter Bildsprache – insbesondere in poetischen Texten – und enthalten theologisch hoch aufgeladene Schlüsselexeme, bei denen eine scheinbar kleine Übersetzungsentscheidung einen erheblichen Interpretationsrahmen eröffnen oder schließen kann. Aus technischer Perspektive verschärfen historische Sprachstufen diese Problemlage zusätzlich. Für biblisches Hebräisch oder Koine-Griechisch stehen im Vergleich zu modernen Sprachen nur begrenzte, nicht immer sauber parallelisierte Daten zur Verfügung; zugleich müssen Tokenizer und Analysewerkzeuge mit Besonderheiten umgehen, die in gegenwartssprachlichen Settings selten relevant sind (z. B. fehlende Worttrennung, komplexe Flexion, textkritische Divergenzen). Dadurch entsteht die Gefahr, dass Modelle glätten, ergänzen oder Bedeutungen konstruieren, anstatt sie textnahe zu übertragen – und dass Fehler nicht als solche erkennbar sind, weil sie grammatisch plausible Zieltexte erzeugen. Vor diesem Hintergrund ist die Frage nach der Kontexttreue maschineller Übersetzungen für alttestamentliche Texte sowohl praktisch als auch wissenschaftlich relevant: praktisch, weil Übersetzungstools längst im Arbeitsalltag genutzt werden; wissenschaftlich, weil sich hier exemplarisch zeigt, wo die Grenzen rein metrischer Qualitätsmessung liegen und warum hermeneutische Kriterien notwendig bleiben. Aus der Verbindung beider Perspektiven – der informatisch-technischen und der hermeneutisch-theologischen – ergibt sich das zentrale Anliegen dieser Arbeit: Maschinelle Übersetzung nicht nur als „funktionierendes Tool“, sondern als interpretatives System zu verstehen, dessen Leistung in einem philologisch sensiblen Feld überprüfbar gemacht werden muss.

1.2 Zielsetzung und Forschungsfrage

Ziel dieser Arbeit ist es, die Kontexttreue maschinelle Übersetzung bei alttestamentlichen Texten systematisch zu untersuchen und dabei die Leistungsunterschiede zwischen verschiedenen Modellklassen sichtbar zu machen. Im Mittelpunkt steht nicht die Frage, ob maschinelle Übersetzung grundsätzlich möglich ist, sondern wie verlässlich neuronale Systeme mit sprachlicher Mehrdeutigkeit, gattungstypischen Strukturen und theologischen Schlüsselstellen umgehen – und welche Fehlerprofile sich dabei wiederkehrend beobachten lassen. Die Arbeit knüpft damit an die Ausgangsannahme an, dass die Qualität maschineller Übersetzung in diesem Feld nicht allein über technische Metriken erfasst werden kann, sondern eine hermeneutisch reflektierte Bewertungsebene benötigt.

Aus dieser Zielsetzung ergibt sich als leitende Forschungsfrage:

> **In welchem Umfang erzeugen neuronale Übersetzungssysteme kontexttreue Übersetzungen alttestamentlicher Texte – und wie unterscheiden sich NMT-Systeme und LLM-basierte Modelle dabei hinsichtlich Strukturtreue, semantischer Präzision und hermeneutisch-theologischer Anschlussfähigkeit?**

Zur Beantwortung dieser Frage werden vier Systeme als repräsentative Vergleichspunkte herangezogen: zwei NMT-Ansätze (DeepL als kommerzielles System und NLLB-200 als offenes Forschungsmodell) sowie zwei LLM-basierte Ansätze (Gemini als großskaliges, cloudbasiertes Modell und LLaMA 3 als lokal betreibbares offenes Basismodell). Damit lassen sich Unterschiede entlang zweier Dimensionen untersuchen: (1) Trainingsstrategie (spezialisiert vs. generisch) und (2) Implementierungsstrategie (cloudbasiert vs. lokal). Methodisch verfolgt die Arbeit einen triangulären Ansatz, der Übersetzung nicht als bloße Relation zwischen Quell- und Zielsprache versteht, sondern eine dritte Vergleichsebene einbezieht: die Evaluierungssprache. Diese wird durch die Septuaginta und die Vulgata repräsentiert und dient sowohl der diachronen Kontextualisierung als auch als hermeneutische Prüfgröße. So werden maschinelle Outputs nicht nur gegen moderne Referenzübersetzungen, sondern zugleich gegen die Wirkungsgeschichte zentraler Übersetzungsentscheidungen gespiegelt. Die Zielsetzung ist damit doppelt: Erstens eine nachvollziehbare, mehrdimensionale Bewertung der Modelloutputs (quantitativ-technisch und qualitativ-theologisch), und zweitens eine Einordnung, welche Formen von Nutzbarkeit sich für theologische Forschung tatsächlich ergeben – inklusive klar benennbarer Grenzen und Risiken.

1.3 Aufbau der Arbeit

Die Arbeit ist so aufgebaut, dass sie zunächst die notwendigen Grundlagen klärt, anschließend die methodische Umsetzung transparent macht und schließlich die Ergebnisse in zwei komplementären Analyseperspektiven auswertet, bevor eine zusammenführende Diskussion erfolgt. Inhaltlich gliedert sich die Untersuchung in fünf Hauptkapitel. Kapitel 2 legt den theoretischen Rahmen. Zunächst werden die biblischen Ursprachen und die textgeschichtliche Situation des Alten Testaments skizziert, um die spezifischen Anforderungen an Übersetzung in diesem Feld deutlich zu machen. Dazu gehört die Einordnung des masoretischen Textes als wissenschaftliche Bezugsgröße ebenso wie die Rolle früher Übersetzungen (Septuaginta) für Rezeptionsgeschichte und Textvarianz. Anschließend werden Grundlagen der maschinellen Übersetzung und neuronaler Sprachverarbeitung dargestellt, einschließlich der besonderen Herausforderungen historischer Sprachen (z. B. Datenknappheit, abweichende Sprachstrukturen, textkritische Divergenzen). Kapitel 3 beschreibt die Methodik. Hier werden Auswahl und Aufbereitung der untersuchten Perikopen nachvollziehbar begründet (sprachliche, theologische und praktisch-methodische Kriterien) sowie die Auswahl der vier Modelle erläutert. Daran schließen sich die Durchführung der Übersetzungen und die Evaluationsmethoden an. Zentral ist dabei die Kombination aus einer quantitativen, technischen Analyse und einer qualitativen, hermeneutisch orientierten Auswertung, die die inhaltliche und theologische Bedeutung der Übersetzungsentscheidungen in den Blick nimmt. Kapitel 4 präsentiert die Analyse und Ergebnisse. Zunächst erfolgt ein Vergleich der maschinellen Ausgaben mit etablierten Referenzübersetzungen. Darauf aufbauend werden die Outputs metrisch und strukturell ausgewertet (z. B. Stabilität, formale Fehlerphänomene, Robustheit), bevor in einem zweiten Schritt eine qualitative Detailanalyse ausgewählter Schlüsselbegriffe und eine kontextuelle, hermeneutische Reflexion folgen. Damit werden technische Befunde und theologische Implikationen gezielt aufeinander bezogen. Kapitel 5 führt die Ergebnisse zusammen und diskutiert sie in technischer und theologischer Perspektive. Es werden Stärken und Schwächen der einzelnen Modelle herausgearbeitet, mögliche Einsatzfelder von KI in der theologischen Forschung eingeordnet und schließlich die zentralen Ergebnisse im Blick auf die Forschungsfrage zusammengefasst. Ergänzend enthält der Anhang die erstellten Übersetzungen und Vergleichstabellen, sodass die Argumentation an konkreten Textbeispielen überprüfbar bleibt.

2 Theoretischer Hintergrund

2.1 Grundlagen der biblischen Ursprachen

Die Bibel ist das Ergebnis eines jahrhundertelangen Entstehungs- und Überlieferungsprozesses, der sich in mehreren Sprachen und Kulturräumen vollzogen hat. Insbesondere das Alte Testament wurde nicht in einer einheitlichen Sprache verfasst, sondern ist in einer Kombination aus biblischem Hebräisch, Aramäisch und Altgriechisch (Koine) überliefert. Die komplexe Textgeschichte dieser Schriften ist ein zentraler Aspekt des Bibelverständnisses und der theologischen sowie sprachwissenschaftlichen Forschung.

Obwohl die Bibel zu den ältesten überlieferten Schriften der Menschheitsgeschichte zählt, stammen die ersten, vollständig erhaltenen Handschriften des Alten Testaments aus dem Mittelalter. Ein gemeingültiger Urtext existiert daher nicht. Vielmehr liegt das Alte Testament in einer Vielzahl von Textzeugen und Sprachfassungen vor, die sich in Nuancen, Struktur und teilweise auch dem Inhalt unterscheiden. „So bietet die BHS (Biblia Hebraica Stuttgartensia) den Text des bisher sogenannten Codex Leningradensis … aus dem Jahr 1008“ (Rösel, 2006). Einzelne Abschnitte des Alten Testaments wurden bereits in der Antike nicht auf Hebräisch, sondern auf Aramäisch verfasst – etwa Teile der Bücher Daniel und Esra. Später kamen auch griechische Übersetzungen und Überlieferungen hinzu, vor allem im Rahmen der jüdischen Diaspora in hellenistischer Zeit.
Die heute maßgebliche hebräische Textform des Alten Testaments basiert auf dem masoretischen Text, der zwischen dem 7. und 10. Jahrhundert n. Chr. von jüdischen Schriftgelehrten (den Masoreten) überliefert und mit einem ausgefeilten System von Vokalzeichen versehen wurde. Dieser Text liegt der BHS und der Biblia Hebraica Quinta (BHQ) – einer von Grund auf neu überarbeiteten Nachfolgerin der BHS – zugrunde und stellt den aktuellen Standard in der wissenschaftlichen Arbeit dar. Bereits mehrere Jahrhunderte vor der masoretischen Textfixierung entstand im hellenistischen Ägypten eine griechische Übersetzung des Alten Testaments: die Septuaginta (Rösel, 2006). Diese wurde im 3. bis 2. Jahrhundert v. Chr. für jüdische Gemeinden in der Diaspora angefertigt, die kein Hebräisch mehr sprachen. Beachtenswert ist, dass die Septuaginta nicht nur sprachlich, sondern auch inhaltlich von der hebräischen Überlieferung abweicht. In vielen Büchern finden sich erweiterte, gekürzte oder anders strukturierte Passagen. Besonders deutlich wird dies z. B. im Buch Jeremia, dessen griechische Fassung etwa ein Siebtel kürzer ist als der masoretische Text und die Kapitel in anderer Reihenfolge enthält.

Solche Unterschiede zeigen, dass es in der Zeit vor der endgültigen Textstabilisierung keine einheitliche Fassung der biblischen Bücher gab. Vielmehr waren mehrere Versionen desselben Textes im Umlauf, deren Abweichungen auf redaktionelle Eingriffe, unterschiedliche Übersetzungstraditionen oder regionale Textvarianten zurückzuführen sind. Diese parallelen Textformen sind nicht als „richtig“ oder „falsch“ zu bewerten, sondern zeugen von einem lebendigen und vielschichtigen Überlieferungsprozess.

2.1.1 Althebräisch und Altgriechisch

(Alt-)Hebräisch

Das Biblische Hebräisch ist die primäre Sprache des Alten Testaments und stellt damit einen wesentlichen Untersuchungsgegenstand der alttestamentlichen Exegese und Sprachwissenschaft dar. Es gehört zur nordwestsemitischen Untergruppe der semitischen Sprachfamilie und ist eng mit dem Phönizischen, Moabitischen und Aramäischen verwandt (Jenni, 1981). Seine historische Verbreitung konzentrierte sich auf das antike Palästina, insbesondere auf die Gebiete des ehemaligen Königreichs Juda. In der Forschung wird das Biblische Hebräisch oft in verschiedene Entwicklungsstufen unterteilt. Insbesondere wird zwischen einem sogenannten Frühbiblischen und einem Spätbiblischen Hebräisch unterschieden, die sich in Syntax, Lexik und stilistischen Mitteln teils deutlich voneinander abheben. Nach Jenni (1981, S. 15) lässt sich die Geschichte der hebräischen Sprache in drei Stufen gliedern. Sowohl das Alt- als auch das Mittelhebräisch sind für biblische Urtexte von Bedeutung. Während das Althebräisch in den ältesten Texten des Alten Testaments – aus der Königszeit, 10.–6. Jh. v. Chr. stammend – Anwendung findet, entwickelte sich das Mittelhebräisch zur Sprache der Mischna, die ersten mündlichen Überlieferungen der Tora, um Christi Geburt mit darin enthaltenen griechischen und aramäischen Lehnworten. Das Neuhebräisch (auch als Ivrith bekannt) stellt mit einer gewollt vereinfachten Grammatik seit der Gründung im Jahr 1948 die offizielle Sprache des Staates Israel dar. Die ursprünglichen Textzeugen des Alten Testaments liegen ausschließlich in alt- und mittelhebräischer Sprachform vor, sodass im Folgenden, sofern von „Hebräisch“ die Rede ist, in erster Linie diese beiden historischen Sprachstufen gemeint sind.

Eine charakteristische Eigenschaft des Biblischen Hebräisch ist sein stark wurzelbasiertes Lexikon. Die überwiegende Mehrheit der Wörter lässt sich auf sogenannte triliterale Wurzeln zurückführen, also Konsonantenverbindungen mit drei Grundkonsonanten – auch Radikale genannt. Aus diesen lassen sich durch eine Vielzahl morphologischer Muster sowohl Verben als auch Substantive, Adjektive und weitere Wortarten ableiten (Jenni, 1981). Beispielsweise liegt die Wurzel כתב (k-t-b) nicht nur dem Verb כָּתַב (*kātav* – „er schrieb“) zugrunde, sondern auch מִכְתָּב (*miktāv* – „Inschrift“), כָּתַבְתִּי (*katavti* – „ich schrieb“) und כָּתוּב (*katuv* – „geschrieben“). Dieses System verleiht der Sprache eine hohe interne Kohärenz bei gleichzeitiger Ausdrucksvielfalt.
Die Verblehre des Hebräischen unterscheidet sich grundlegend von den Tempussystemen, die in indogermanischen Sprachen verwendet werden. Arabische Sprachen, wozu auch das Hebräische zählt, orientieren sich daher nicht an einem chronologischen Zeitverlauf, sondern an der Art der Handlung bzw. subjektiven Aspekten (Jenni, 1981). Dabei wird zusätzlich zu einer infiniten Verbform zwischen *perfekten* (abgeschlossenen) und *imperfekten* (unabgeschlossenen) *Handlungen* unterschieden. Das Perfekt (Afformativkonjugation) „drückt alle Formen der Vergangenheit aus, das Perfekt, das Präteritum, das Plusquamperfekt und das Futur II, alle werden … das ‚Vollendete‘ genannt“ (Meghouche, 2025). Die damit übrigbleibenden Zeitstufen Präsens sowie Futur sind noch nicht abgeschlossen und werden daher durch das Imperfekt (Präformativkonjugation) gebildet. Die

konjugierten Verbformen können dabei „… durch Prä- und Afformative erweitert werden. … Damit meint man grammatische Morpheme, die vor … oder nach … den drei Radikalen einer Wurzel erscheinen“ (Neef, 2018). Diese Morpheme bestimmen damit Person, Genus, Numerus und Tempus einer Verbform. Die genauere Wahl der übersetzten Zeitform basiert somit häufig auf dem Kontext und kann aufgrund der unterschiedlichen Tempussysteme nicht geradlinig übertragen und übernommen werden. Die Syntax des Hebräischen ist meist verbinitial. Im Allgemeinen unterscheidet sich die Wortfolge hebräischer Verbalsätze vom Deutschen in der Stellung des Verbs. Im Deutschen beginnt ein Satz in der Regel mit einem Subjekt, das mit einem verbindenden Verb weitere Objekte anschließt. Daraus resultiert eine allgemeine Wortstellung: Subjekt – Prädikat – Objekt (SPO). Davon differierend werden hebräische Sätze in der Regel nach dem Schema: Prädikat – Subjekt – Objekt (PSO) gebildet werden. So entsteht nach Neef (2018) ein Verbalsatz als einfachste Form der Satzbildung im Hebräischen. Hierbei ist das (pronominale) Subjekt bereits in der Verbform enthalten. Im Gegensatz zu Nominalsätzen, die Zustände und Verhältnisse wiedergeben, beschreiben Verbalsätze Handlungen und Vorgänge. Als Sätze ohne finites Verb stehen Nominalsätze im Hebräischen auch ohne „… die im Deutschen notwendigen Kopula (Verbum ‚sein‘) …“ (Jenni, 1981). Für eine lesbare deutsche Übersetzung muss ein kontextpassendes kopulatives Verb eingefügt werden. Ebenso sind Nominalsätze keiner bestimmten Zeitstufe zugeordnet, „die Zeitsphäre der Aussage kann nur aus dem Zusammenhang geschlossen werden“ (Neef, 2018).
Ein weiterer Aspekt betrifft den reduzierten Wortschatz des Biblischen Hebräisch. Die Sprache arbeitet mit einem relativ kleinen, aber funktional effizienten Vokabular, das durch das erwähnte Wurzelsystem in seiner Anwendung stark erweitert werden kann. Die semantische Breite einzelner Lexeme ist daher oft kontextabhängig und stellt bei der Übersetzung sowie der lexikalischen Analyse eine besondere Herausforderung dar.
Die semantische Auslegung biblisch-hebräischer Texte erfordert stets eine sorgfältige Berücksichtigung des literarischen und historischen Kontexts. Viele hebräische Wortformen besitzen ein breites Bedeutungsspektrum, das sich nicht auf eine eindeutige, feststehende Übersetzung reduzieren lässt (Jenni, 1981). Nach Becker (2021) müssen bei der Übersetzung biblischer Texte Ausgangs- und Zielsprache aufeinander abgestimmt werden, wobei je nach Zielsetzung zwischen wörtlicher, philologisch genauer, sinngetreuer und freier Übersetzung unterschieden wird. Hinzu kommt, dass die formale Struktur des Hebräischen in einzelnen Passagen eine Mehrdeutigkeit zulässt, die in narrativen besonders aber in prophetischen und poetischen Texten bewusst eingesetzt wird.

Die Entwicklung des Biblischen Hebräisch lässt sich auch anhand sprachgeschichtlicher Unterschiede innerhalb des kanonischen Textes nachzeichnen. In den späteren biblischen Schriften zeigen sich deutliche Einflüsse des Aramäischen, sowohl im syntaktischen Aufbau als auch im Wortschatz. Zudem treten orthografische Vereinfachungen und abweichende Formen gegenüber älteren Textschichten auf.

(Alt-)Griechisch

Das Altgriechische gehört, wie das Deutsche, aber entgegen dem Hebräischen, zu den indogermanischen Sprachen. Somit weist das Griechische deutlichere Parallelitäten zum Deutschen als zum Hebräischen auf. Das Griechische umfasst Sprachstufen vom frühen literarischen Griechisch bis zur Spätantike – etwa 800 v. Chr. bis 550 n. Chr. (Siebenthal, 2008). Innerhalb dieser Tradition gilt das attisch-klassische Griechisch – vertreten durch Autoren wie Platon, Thukydides oder Sophokles – als literarischer Maßstab und Höhepunkt. Mit der Ausbreitung der hellenistischen Kultur, bedingt durch die Eroberungen Alexanders des Großen, bildete sich jedoch eine vereinfachte und vereinheitlichte Sprachform heraus, die sogenannte Koine („Gemeinsprache"). Sie diente als Lingua franca des östlichen Mittelmeerraums und prägte die kulturelle sowie religiöse Kommunikation in der gesamten Region (Decker, 2014).
Von zentraler Bedeutung ist dabei, dass die griechische Übersetzung des Alten Testaments, die Septuaginta (kurz: LXX), im 3. Jh. v. Chr. in dieser Koine verfasst wurde. Ebenso wurden die Schriften des Neuen Testaments – bekannt als Novum Testamentum Greace (NTG), durchweg in Koine-Griechisch abgefasst, wodurch diese Sprachform unmittelbaren Eingang in die biblische Überlieferung fand. Damit bildet die Koine die verbindende sprachliche Grundlage zwischen der jüdischen Tradition in der Diaspora und der frühchristlichen Bewegung.

Das Altgriechische ist eine stark flektierende Sprache, deren Formenreichtum sich besonders in der Nominal- und Verbalmorphologie zeigt. Substantive, Adjektive und Artikel werden nach Kasus, Numerus und Genus dekliniert. Während das klassische Griechisch noch über fünf Kasus (Nominativ, Genitiv, Dativ, Akkusativ, Vokativ) und zum Teil auch den Dual verfügt, beschränkt sich die Koine im alltäglichen Gebrauch zunehmend auf vier Kasus. Die Kongruenz zwischen Artikel, Substantiv und Adjektiv ist dabei ein zentrales grammatisches Strukturmerkmal (Kühner, 1890). Das Verbalsystem ist hoch differenziert: Verben werden nach Person, Numerus, Modus (Indikativ, Konjunktiv, Optativ, Imperativ), Diathese (Aktiv, Medium, Passiv) sowie Tempus und Aspekt konjugiert. Entscheidend ist die Unterscheidung von Aspekten – ähnlich zum Hebräisch: Der Aorist beschreibt punktuelle oder abgeschlossene Handlungen, das Präsens laufende oder wiederholte Vorgänge, während das Perfekt einen gegenwärtigen Zustand mit vergangener Ursache bezeichnet (Decker, 2014). Die Koine weist Vereinfachungstendenzen gegenüber dem klassischen Griechisch auf. So verliert der Optativ weitgehend an Bedeutung, komplexe synthetische Formen werden seltener gebraucht und durch einfachere oder periphrastische Konstruktionen ersetzt (Siebenthal, 2008). Gleichwohl bleibt die morphologische Vielfalt bestehen und erlaubt eine präzise Differenzierung, die im biblischen Kontext für die Auslegung von Texten eine besondere Rolle spielt.
Die Syntax des Altgriechischen ist von einer vergleichsweise freien Wortstellung geprägt, die durch die Kasusendungen ermöglicht wird. Grundsätzlich überwiegt die Abfolge nach Schema SPO, doch können Satzglieder je nach Betonung oder stilistischer Absicht an den Satzanfang oder in andere Positionen gestellt werden (Kühner, 1890). Charakteristisch sind Partizipialkonstruktionen, die häufig Nebensätze ersetzen und zusätzliche Informationen in den Satz integrieren. Besonders

bedeutsam ist der Accusativus cum infinitivo (AcI), der die Einbettung abhängiger Aussagen erlaubt und im narrativen wie argumentativen Stil häufig vorkommt (Siebenthal, 2008). Hinzu tritt der Genitivus absolutus, eine syntaktisch selbstständige Konstruktion, die eine adverbiale Nebensatzfunktion übernimmt und so komplexe Zusammenhänge in verdichteter Form ausdrückt.
Das klassische Griechisch ist für seine verschachtelten Perioden bekannt, in denen Haupt- und Nebensätze kunstvoll miteinander verbunden sind. In der Koine ist dagegen eine gewisse Vereinfachung zu beobachten: Längere Perioden treten seltener auf, Nebensätze werden reduziert oder durch parataktische Strukturen ersetzt. Diese Entwicklung macht die Sprache einerseits zugänglicher, bewahrt aber zugleich die charakteristische Ausdrucksvielfalt, die sich gerade in der Verbindung von Nebensatzstrukturen und Partizipialkonstruktionen zeigt (Decker, 2014). Die Semantik des Altgriechischen ist stark kontextabhängig und erfordert eine genaue Analyse der jeweiligen Textumgebung. Viele Lexeme besitzen ein weites Bedeutungsspektrum, das nicht durch eine einzige Übersetzung erfasst werden kann. Die konkrete Bedeutung ergibt sich erst aus syntaktischem Zusammenhang, literarischer Gattung und pragmatischer Intention. Hinzu kommt, dass die Sprache durch ihre ausgeprägte Flexionsmorphologie und die Vielzahl an Partizipial- und Nebensatzkonstruktionen zusätzliche Interpretationsspielräume eröffnet. So können etwa derselbe Kasus oder dieselbe Verbform in unterschiedlichen Kontexten verschiedene semantische Funktionen übernehmen (Kühner, 1890). Besonders in der Koine ist zu beobachten, dass semantische Nuancen oft durch den Kontext, nicht allein durch die Form, bestimmt werden. Der Rückgang komplexer Modusformen – insbesondere des Optativs – führte dazu, dass Bedeutung verstärkt durch Konjunktiv oder periphrastische Umschreibungen vermittelt wird (Decker, 2014). Auch beim Gebrauch des Artikels zeigt sich eine semantische Vielfalt: Er markiert nicht nur Bestimmtheit, sondern kann ganze Wortgruppen substantivieren oder abstrakte Begriffe konkretisieren (Siebenthal, 2008). Damit erweist sich die Semantik des Griechischen als hoch flexibel, zugleich aber abhängig von stilistischen, grammatischen und pragmatischen Faktoren, die in der Auslegung stets berücksichtigt werden müssen.

Das Altgriechische, insbesondere die Koine, bildet eine zentrale Grundlage der biblischen Überlieferung. Als Sprache der LXX und des NTG verbindet es die jüdische Tradition mit der griechisch-hellenistischen Welt und eröffnet so den Zugang der biblischen Texte über den engeren hebräischen Sprachraum hinaus. Seine morphologische Präzision und syntaktische Vielfalt haben nicht nur die literarische Gestalt der Schriften geprägt, sondern auch ihre theologische Rezeption und Auslegung bis in die Gegenwart hinein bestimmt.

2.1.2 Theologische Bedeutung von Übersetzung

Die Bibel ist als historisch gewachsenes Buch von Sprachen verschiedener Zeiten und Epochen geprägt. Übersetzungen bedingen die Aktualität und Bedeutung der biblischen Texte für die Glaubensgemeinschaften der Epochen entlang der biblischen Entstehungsgeschichte. So wurden schon früh die Schriften der Ursprachen in andere Sprachräume übertragen, um sie einer breiteren Leserschaft zugänglich zu machen. Dabei gilt der Grundsatz: Jede Übersetzung ist zugleich auch eine Interpretation – eine ideale Übersetzung kann es also nicht geben (Lebedewa, 2007). Eine Übersetzung lässt sich nie als iterative Wortübertragung in die Zielsprache umsetzen, sondern enthält immer Deutungen sowie Entscheidungen über Wortwahl, Satzbau und Bedeutungsunterschiede.

Von herausragender historischer Bedeutung ist die LXX. Sie eröffnete der jüdischen Diaspora den Zugang zu den heiligen Schriften und wurde zugleich zur maßgeblichen Bibel der ersten christlichen Gemeinden. Im Westen prägte die lateinische Vulgata, übersetzt von Hieronymus im 4. Jahrhundert n. Chr., über viele Jahrhunderte hinweg das Schriftverständnis der Kirche. Später setzte die Reformation mit der Lutherbibel einen entscheidenden Akzent: Sie verband eine genaue Rückbindung an die Ursprungstexte mit einer sprachlich eingängigen Gestaltung, die die Entwicklung der deutschen Sprache maßgeblich beeinflusste. In der Übersetzungswissenschaft lassen sich verschiedene Prinzipien unterscheiden. Becker (2021) listet die folgenden vier Unterscheidungen auf. Eine **wörtliche Übersetzung** orientiert sich stark an der Ausgangssprache. Nicht jede wörtliche oder grammatikalische Bedeutung der Ursprungssprache kann in der Zielsprache entsprechenden dargestellt werden. Dabei auftretende Unverständlichkeiten werden bewusst akzeptiert. Die **philologisch genaue (wissenschaftliche) Übersetzung** legt den Schwerpunkt auf die Grammatik und Wortbedeutungen der Zielsprache, um so eine Grundlage für sprachwissenschaftliche und exegetische Studien zu bieten. Beispielhaft lassen sich hier die Lutherübersetzungen der Bibel anführen. Verglichen damit folgt z. B. die BasisBibel eher der dritten Übersetzungspraxis – der **sinngetreuen Übersetzung**. Mit dieser wird das Ziel einer unmittelbaren, allgemeinverständlichen Übersetzung verfolgt. Der Text wird inhaltlich wiedergegeben und kann sich dadurch bedingt vom ursprünglichen Text in einem angemessenen Maß entfernen. Zuletzt bleibt die **freie Übersetzung**, die maßgeblich die Botschaft eines Textes transportiert, wobei die verschiedenen, kommunikativ orientierten Übertragungen an keinen zu übertragenden Text gebunden sind.

Sowohl das Griechisch als auch besonders das Hebräisch sind kontextsensitive Sprachen. Von besonderem Gewicht ist die Wortwahl, da einzelne Übersetzungsentscheidungen theologische Akzente setzen. Schon kleine Unterschiede können das Bibelverständnis prägen – bspw. lässt sich צְדָקָה (*ṣĕdāqāh*) als Barmherzigkeit, Verdienst, Gerechtigkeit oder Almosen übersetzen werden. Analog dazu liefert ebenfalls das griech. Wort λόγος (*lógos*) mit ‚Sprache', ‚Vernunft', ‚Lehre' oder auch ‚Rechnung' verschiedenste Übersetzungsmöglichkeiten.

Übersetzungen stehen immer in einem Spannungsfeld zwischen Ausgangstext, Zielkultur und intendierter Leserschaft – gerade darin gewinnen sie ihre bleibende Bedeutung.

2.1.3 Hermeneutik und Exegese

Das Verstehen biblischer Texte ist ein komplexer Prozess, der weit über das reine Lesen hinausgeht. Zwischen den Ursprungssprachen der Bibel – Hebräisch, Aramäisch und Griechisch – und den heutigen Leserinnen und Lesern liegen Jahrhunderte sprachlicher, kultureller und theologischer Entwicklung. Dieser Abstand macht eine methodische Reflexion des Verstehens notwendig. Hier setzt die Hermeneutik an, die als Theorie des Verstehens die Voraussetzungen und Bedingungen von Auslegung beschreibt. Die Exegese wiederum ist die konkrete Anwendung dieser Prinzipien: Sie bezeichnet die methodisch kontrollierte Analyse eines Textes mit dem Ziel, dessen ursprüngliche Aussage und Intention möglichst präzise zu erfassen. Beide Ebenen sind untrennbar miteinander verbunden, da jede Exegese implizit hermeneutisch bestimmt ist.

Unter Hermeneutik vom griech. ἑρμηνεύειν (*hermēneúein* – „erklären“, oder „übersetzen“) versteht man die Kunst und Wissenschaft der Interpretation. Sie geht davon aus, dass es kein unvoreingenommenes Lesen gibt, sondern jedes Verstehen durch Vorverständnisse geprägt ist – sprachliche, kulturelle oder auch theologische. Diese Vorverständnisse gilt es bewusst zu reflektieren. Hermeneutik fragt danach, wie ein Text in seinem ursprünglichen historischen Kontext verstanden wurde und wie er zugleich in neuen Situationen und Zeiten Bedeutung entfalten kann. Dabei geht es sowohl um die Differenz zwischen Text und Leser als auch um die Brücken, die zwischen beiden geschlagen werden. Besonders relevant ist dabei die Einsicht, dass Texte nicht statisch sind, sondern durch ihre literarische Gestalt, ihre Überlieferung und ihre Wirkungsgeschichte immer schon interpretativ geprägt sind.

Die Exegese vom griech. ἐξήγησις (*exēgesis* – „Erläuterung“) nimmt diese hermeneutischen Voraussetzungen auf und entfaltet daraus einen methodisch gegliederten Arbeitsprozess. Nach Uwe Becker (2021) umfasst die alttestamentliche Exegese mehrere Arbeitsschritte, die in einer klaren Abfolge angeordnet sind, zugleich aber in wechselseitigem Bezug zueinander stehen.
Den methodischen Anfang bildet die **Textkritik**. Sie vergleicht die verschiedenen Handschriften und Textzeugen – etwa den masoretischen Text, die Septuaginta oder die Qumranfunde – und versucht, den ursprünglichen Wortlaut so zuverlässig wie möglich zu rekonstruieren. Eng damit verbunden ist die eigene Übersetzung des Textes. Sie ist unverzichtbar, da sie nicht nur sprachliche Präzision verlangt, sondern auch schon eine erste interpretative Entscheidung darstellt: Welche Wortbedeutungen werden gewählt, wie wird Syntax übertragen, und wie wird der Sinn in der Zielsprache wiedergegeben? Ein weiterer grundlegender Schritt ist die Abgrenzung und Gliederung der zu untersuchenden Perikope. Hier wird bestimmt, wo ein Text beginnt und endet, wie er sich in den literarischen Kontext einfügt und welche innere Struktur er aufweist. Solche Beobachtungen sind entscheidend, um den logischen Aufbau und die literarische Eigenart eines Abschnitts zu erfassen. Darauf folgt die Sprach- und Wortanalyse, in der zentrale Begriffe und Formulierungen untersucht werden. Hierbei wird geklärt, welche Bedeutung ein Wort in verschiedenen Kontexten haben kann, welche semantischen Spielräume bestehen und wie grammatische Formen das

Verständnis prägen. Dieser Schritt macht sichtbar, dass Bedeutung oft nicht in einer festen lexikalischen Definition liegt, sondern durch Kontext und Verwendung bestimmt wird. Die anschließende **Literarkritik** und dazugehörigen Methoden eröffnen weitere Dimensionen. Die Formkritik fragt nach der literarischen Gattung eines Textes – ob es sich etwa um ein Gebet, ein Gesetz, eine Erzählung oder ein prophetisches Orakel handelt. Diese Bestimmung ist wichtig, da jede einzelne eigene Regeln, Funktionen und Aussageabsichten hat. Die Gattungsbestimmung erlaubt Rückschlüsse auf den ursprünglichen Sitz im Leben, also den konkreten lebensweltlichen Kontext, in dem ein Text gebraucht wurde. Darauf aufbauend fragt die **Überlieferungs**- und **Redaktionsgeschichte**, wie ältere Textschichten in den Prozess der Tradierung aufgenommen, erweitert oder in neuer Weise gedeutet wurden. Im Blick stehen dabei nicht nur die sprachlichen oder stilistischen Veränderungen, sondern auch die theologische Neuausrichtung, die durch redaktionelle Bearbeitung erfolgte. Damit wird deutlich, dass biblische Texte nicht als feste und unveränderliche Größen zu verstehen sind, sondern als Ergebnisse eines fortlaufenden Prozesses, indem überlieferte Traditionen bewahrt, angepasst und in immer neue Kontexte hinein aktualisiert wurden. Die **Formgeschichte** fragt nach den kleineren literarischen Einheiten innerhalb eines Textes und ordnet sie bestimmten Gattungen wie Erzählung, Gesetz, Prophetenspruch oder Psalm zu. Sie untersucht zudem deren ursprüngliche Funktion im mündlichen Gebrauch, dem genannten Sitz im Leben. Eng verbunden damit ist die **Traditionsgeschichte**, die nachzeichnet, wie solche Einheiten überliefert, weitergegeben und in unterschiedlichen historischen Kontexten gedeutet wurden. Beide Ansätze machen sichtbar, dass biblische Texte das Ergebnis eines längeren Prozesses sind, in dem überlieferte Stoffe bewahrt, verändert und in neue Zusammenhänge gestellt wurden.
Neben diesen klassischen Arbeitsschritten betont Becker (2021) die Sachanalyse. Hier werden die erarbeiteten Beobachtungen thematisch und inhaltlich gebündelt, um die zentrale Aussageabsicht des Textes herauszuarbeiten. Dabei geht es nicht nur um eine Zusammenfassung, sondern um die Klärung, welche theologische, ethische oder historische Botschaft im Zentrum steht. Schließlich führt dies zur theologischen Auswertung, in der die Bedeutung des Textes über seinen historischen Kontext hinaus reflektiert wird. Exegese endet also nicht bei der Rekonstruktion des Damals, sondern fragt auch, inwiefern der Text in heutige Kontexte hineinsprechen kann.

So verstanden ist Exegese ein vielschichtiger Prozess, der philologische Präzision, literarische Analyse und theologische Reflexion miteinander verbindet. Hermeneutik bildet den Rahmen, Exegese hingegen die methodische Umsetzung. Gemeinsam schaffen sie die Voraussetzung dafür, dass biblische Texte nicht nur in ihrer sprachlichen Gestalt verstanden, sondern auch in ihrer historischen Eigenart ernst genommen werden. Erst auf dieser Grundlage kann eine theologische Analyse erfolgen, die die Inhalte der Texte systematisch deutet und in aktuelle Fragestellungen hinein vermittelt.

2.2 Maschinelle Übersetzung und NLP

Die maschinelle Übersetzung stellt seit den Anfängen der Computerlinguistik eine der zentralen Anwendungen des Natural Language Processing (NLP) dar. Sie verbindet linguistische Theorien mit algorithmischen Verfahren und verdeutlicht exemplarisch die Entwicklung von der regelbasierten Sprachverarbeitung hin zu datengetriebenen, neuronalen Modellen. Während frühe Ansätze vor allem auf vordefinierten Regeln und Wörterbüchern basierten, ermöglichte der Fortschritt in Statistik, maschinellem Lernen und schließlich im Deep Learning immer leistungsfähigere Übersetzungssysteme. Zugleich bildet die maschinelle Übersetzung ein Schlüsselfeld, um die Möglichkeiten und Grenzen moderner, KI-gestützter Sprachverarbeitung sichtbar zu machen: Sie erfordert nicht nur die Analyse einzelner Wörter, sondern auch die Berücksichtigung von Syntax, Semantik und Kontext über ganze Texte hinweg. Im Folgenden wird daher die Entwicklung der maschinellen Übersetzung nachgezeichnet, die technischen Grundlagen neuronaler Modelle erläutert und die besonderen Herausforderungen bei der Verarbeitung historischer Sprachen herausgestellt.

2.2.1 Entwicklung der maschinellen Übersetzung

Die maschinelle Übersetzung (MÜ oder MT – *engl. machine translation*) ist keine Erscheinung des jüngsten KI-Zeitalters. Mit einer mehr als siebzigjährigen Forschungsgeschichte gehört sie zu den älteren KI-basierten Technologien, die eng mit den Entwicklungen der Computerlinguistik und der Künstlichen Intelligenz verbunden ist. Die Wurzeln der MT lassen sich an verschiedenen Zeitpunkten verorten. Stein (2009) sieht die Dechiffrierung feindlicher Funksprüche während des zweiten Weltkriegs als zentrale Rolle. So konnte eine Auswertung mittels statistischer Methoden auf den ersten Rechenmaschinen realisiert werden. Hutchins (2014) und Schwanke (1991) verweisen auf die gleichzeitige, aber unabhängige Entwicklung eines maschinengestützten Übersetzungssystems im Jahr 1933 von G. Artsrouni und P. Troyanskii. Beide Systeme verwendeten bestimmte Lexika, konnten allerdings nur unflektierte Wortformen – also in ihrer infiniten Form – berücksichtigen. Weitere, breit gestreute Forschung im Bereich der MT konnte mit dem Georgetown-IBM-Experiment aus dem Jahr 1954 gewonnen werden, bei dem ein Computer rund 60 russische Sätze korrekt ins Englische übertragen konnte. Dies geschah, obwohl das Experiment stark vorbereitet war und das System über nur etwa 250 Wörter und sechs grammatikalische Regeln verfügte (Hutchins, 2004). Die frühen Systeme setzten überwiegend auf direkte Wort-für-Wort-Übersetzung, wobei nur sehr einfache morphologische und syntaktische Regeln berücksichtigt wurden. Rasch zeigte sich, dass diese Ansätze an ihre Grenzen stießen: Sprachen sind hochgradig mehrdeutig, sodass die Ambiguität der zu übersetzenden Wörter das zentrale Problem der MT wurde (Stampf, 2012). Ohne tieferes Kontextverständnis konnten die Übersetzungen weder idiomatisch noch zuverlässig sein.

Als Reaktion wurden in den darauffolgenden Jahrzehnten **regelbasierte Systeme** (auch RBMT – *engl. Rule-Based Machine Translation*) entwickelt. Diese beruhen auf umfangreichen, von Linguistinnen und Linguisten erstellten Grammatiken und zweisprachigen Wörterbüchern (Arnold et al., 1994). Solche Systeme ermöglichten eine deutlich präzisere Kontrolle der Übersetzung, da die linguistischen Regeln transparent nachvollzogen werden konnten. Sie erwiesen sich vor allem in kontrollierten Fachdomänen oder für Sprachpaare mit gut erforschter Grammatik als nützlich. Der entscheidende Nachteil bestand jedoch in der geringen Skalierbarkeit: Für jedes neue Sprachpaar musste zahllose Regeln manuell erstellt und gepflegt werden. Damit war der Entwicklungsaufwand enorm und die Systeme reagierten empfindlich auf sprachliche Variation oder kreative Ausdrucksweisen.

In den 1990er-Jahren setzte sich daher zunehmend ein statistischer Ansatz durch. Die **statistische maschinelle Übersetzung** (SMT – *engl. Statistical Machine Translation*) beruht auf der Analyse großer zweisprachiger Korpora, aus denen Übersetzungswahrscheinlichkeiten berechnet werden. Wegweisend waren die von IBM entwickelten Modelle (Hutchins, 2014), die zunächst auf Wortebene arbeiteten. Später führte die Einführung von Phrase-Based SMT (Koehn et al., 2003) zu erheblichen Fortschritten, da nun auch längere Sprachbausteine berücksichtigt werden konnten. Der Vorteil dieses Ansatzes lag in der besseren Anpassungsfähigkeit an reale Sprachdaten: Systeme konnten durch das Einspielen neuer Korpora trainiert und für verschiedene Domänen optimiert werden. Gleichwohl blieben SMT-Systeme häufig bruchstückhaft: Grammatikalische Strukturen wurden nur unvollständig erfasst, die Übersetzungen wirkten oft nicht-idiomatisch, und ein hoher Bedarf an parallelen Textdaten stellte insbesondere für weniger verbreitete Sprachen eine Hürde dar (Koehn, 2010).

Einen grundlegenden Paradigmenwechsel brachte schließlich die Entwicklung der **neuronalen maschinellen Übersetzung** (NMT – *engl. Neural Machine Translation*) ab 2014. Mit Hilfe von Deep-Learning-Methoden gelang es erstmals, ganze Sätze als Sequenzen zu verarbeiten und kontextübergreifend zu übersetzen. Grundlage bildeten sogenannte Encoder-Decoder-Architekturen, die in Arbeiten u.a. von Sutskever et al. (2014) vorgestellt wurden. Besonders entscheidend war die Einführung von Attention-Mechanismen, die es dem Modell erlaubten, relevante Teile des Eingangssatzes flexibel zu gewichten. Einen weiteren Meilenstein stellte die Entwicklung der Transformer-Architektur dar (Vaswani et al., 2017), die die Verarbeitung längerer Abhängigkeiten innerhalb von Texten ermöglichte und bis heute den Standard in der maschinellen Übersetzung setzt. NMT-Systeme zeichnen sich durch flüssigere, kontexttreuere und insgesamt natürlichere Übersetzungen aus. Zugleich sind sie in der Lage, Generalisierungen über Sprachgrenzen hinweg vorzunehmen, was insbesondere für ressourcenschwache Sprachen neue Perspektiven eröffnet. Kritisch zu sehen ist allerdings die geringe Transparenz der Modelle: Während regelbasierte Systeme in jedem Schritt nachvollziehbar waren, bleibt NMT weitgehend eine Black Box. Zudem erfordert das Training große Datenmengen und erhebliche Rechenressourcen.

Seit den 2020er-Jahren werden NMT-Systeme zunehmend durch **Large Language Models (LLM)** ergänzt, die Übersetzung nicht mehr als isolierte Aufgabe, sondern als Teil eines allgemeinen Sprachverständnisses integrieren. Systeme wie GPT-5 sind in der Lage, Übersetzungen zu erzeugen, kontextuelle Informationen zu berücksichtigen und sogar stilistische Anpassungen vorzunehmen. Gleichzeitig haben sich spezialisierte Modelle etabliert, die für bestimmte Domänen – etwa Rechtstexte, medizinische Inhalte oder religiöse Schriften – trainiert werden. Ein weiterer wichtiger Forschungsstrang betrifft low-resource languages, für die wenig Daten verfügbar sind: Hier werden Methoden wie Transfer Learning oder die Kombination von neuronalen Ansätzen mit linguistischem Wissen erprobt. Diese Entwicklungen markieren einen Übergang von der klassischen Übersetzungstechnologie hin zu umfassenderen Sprachmodellen, die Textverstehen und Übersetzen miteinander verbinden.

2.2.2 Technische Grundlagen neuronaler Modelle

Die NMT bildet den gegenwärtigen Stand der Forschung im Bereich der automatisierten Sprachverarbeitung. Ihr Erfolg ist das Ergebnis einer kontinuierlichen Entwicklung, die auf den Defiziten früherer Ansätze, wie regelbasierter und statistischer Systeme, aufbaut und diese gezielt überwindet. RBMT versucht Übersetzungen auf Grundlage expliziter linguistischer Regeln abzubilden. Ein Satz wurde dabei in drei Schritten verarbeitet: Zunächst erfolgte die Analyse der Ausgangssprache (A), anschließend der Transfer in eine Zwischenrepräsentation (R) und schließlich die Synthese in der Zielsprache (S). Das Ergebnis ist die Übersetzung (T) eines Textes x. Formal lässt sich dieser Prozess wie folgt darstellen.

$$T(x) = S \circ R \circ A$$

Tabelle 2-1 *Beispieldarstellung: Übersetzung mit RBMT*

$\boldsymbol{x}$	**Haus**	**ist**	**groß**
$\boldsymbol{A}$	*Substantiv* Nominativ. Sing. Neutrum	*Verb* – „sein“ 3. Pers. Sing. Präsens	*Adjektiv* Prädikativ
$\boldsymbol{R}$	„Haus“ → „house“	„sein“ → „is“	„groß“ → „big“
$\boldsymbol{S}$	House	is	big
$\boldsymbol{T(x)}$	House is big		

Das Verfahren erlaubte zwar eine präzise Kontrolle der Übersetzung, war jedoch aufgrund des enormen Aufwands zur Regeldefinition kaum skalierbar (Arnold et al., 1994). Mit dem Beginn der SMT erfolgte ein Paradigmenwechsel. Übersetzung wurde als Wahrscheinlichkeitsproblem formuliert. Ziel war es, für eine Eingabe f – z. B. in Form eines Satzes die wahrscheinlichste Übersetzung $\hat{e}$ zu bestimmen. So gibt e eine von vielen möglichen Übersetzungen an, von denen die jeweiligen Wahrscheinlichkeiten P und durch $\arg\max - argumentum\ maximi$ der Maximalwert ermittelt wird.

$$\hat{e} = \arg\max_{e} P(e|f)$$

Die Zerlegung der Übersetzungswahrscheinlichkeit $P(e|f)$ in ein Sprachmodell $P(e)$ und ein Übersetzungsmodell $P(f|e)$ stellte einen wesentlichen Fortschritt dar, da sie erstmals die Nutzung großer Textkorpora ermöglichte. Während das Sprachmodell aus einsprachigen Daten typische Satzmuster der Zielsprache erlernte, konnten mit dem Übersetzungsmodell zweisprachige Korpora genutzt werden, um Entsprechungen zwischen Quell- und Zielsprache statistisch abzuleiten. Auf diese Weise ließen sich Übersetzungen datengetrieben erzeugen, ohne dass jede Regel explizit formuliert werden musste. Gleichwohl blieb der Ansatz in seiner Reichweite begrenzt, da n-Gramm-Modelle lediglich lokale Abhängigkeiten erfassen konnten und längere Satzstrukturen oft nur unzureichend modelliert wurden.

Neuronale Modelle setzen an diesem Punkt an. Sie modellieren Sprache nicht mehr als diskrete Symbolketten, sondern als Vektoren in kontinuierlichen Räumen, in denen semantische Beziehungen durch geometrische Nähe abgebildet werden. Ein einzelnes künstliches Neuron – die kleinste Recheneinheit – berechnet eine gewichtete Summe seiner Eingaben und transformiert diese durch eine nichtlineare Aktivierungsfunktion.

$$y = f(W_x + b)$$

Die Matrix W enthält dabei die gelernten Gewichte, b ist ein Bias, und f ist eine nichtlineare Aktivierungsfunktion. Für die Vorhersage von Wörtern in einer Übersetzung ist häufig die softmax-Funktion entscheidend, da sie eine Wahrscheinlichkeitsverteilung über alle möglichen Wörter erzeugt. Die softmax-Funktion stellt einen zentralen Bestandteil neuronaler Übersetzungsmodelle dar, da sie die Rohwerte (sog. Logits), die das Netz für jedes mögliche Ausgabewort berechnet, in Wahrscheinlichkeiten transformiert. Formal ergibt sich für ein Wort z_i.

$$\mathrm{softmax}(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

Durch die Exponentialfunktion werden größere Werte überproportional verstärkt, während die Normierung durch den Nenner sicherstellt, dass die Summe aller Wahrscheinlichkeiten gleich 1 ist. Auf diese Weise lässt sich eine klare Rangordnung der möglichen Zielwörter erzeugen. So kann das Modell bei der Übersetzung des deutschen Satzes „Das Haus ist groß“ mit hoher Wahrscheinlichkeit das englische Wort „house“ auswählen, während weniger passende Alternativen wie „building“ oder „cat“ deutlich geringere Wahrscheinlichkeiten erhalten. Die softmax-Funktion fungiert damit als Schnittstelle zwischen den internen Berechnungen des Netzes und der tatsächlichen Wortausgabe und ermöglicht eine probabilistische, kontextabhängige Generierung der Zielsprache. Zur Veranschaulichung sei angenommen, das Modell habe für das nächste Ausgabewort folgende Rohwerte berechnet.

Tabelle 2-2 *Beispielrechnung softmax-Aktivierungsfunktion*

	House	**Building**	**Cat**
$\boldsymbol{z_i}$	$z_{house} = 2.0$	$z_{building} = 1.0$	$z_{cat} = 0.1$
$\boldsymbol{\exp(z_i)}$	≈ 7.39	≈ 2.72	≈ 1.11
$\boldsymbol{\sum_j \exp(z_j)}$	$= 11.22$		
$\boldsymbol{P(z_i)}$	≈ 0.66	≈ 0.24	≈ 0.10

Das Modell würde somit mit größter Wahrscheinlichkeit „house“ als Übersetzung auswählen. Dieses Beispiel verdeutlicht, dass die softmax-Funktion nicht nur eine Normalisierung vornimmt, sondern zugleich die relativen Präferenzen des Netzes in eine klar interpretierbare Wahrscheinlichkeitsverteilung überführt.

Das Übersetzungsmodell entscheidet sich nicht deterministisch für ein einzelnes Wort, sondern wählt jenes mit der höchsten Wahrscheinlichkeit aus. Die Qualität dieser Vorhersage wird während des Trainings mit der Kreuzentropie-Verlustfunktion bewertet, die den Abstand zwischen der vorhergesagten Verteilung $\hat{y}$ und der wahren Zielverteilung y misst.

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

In der Verlustfunktion bezeichnen die Terme y_i die wahre Zielverteilung, die im Training als One-Hot-Vektor vorliegt, sodass lediglich das korrekte Wort den Wert 1 annimmt, während alle anderen 0 sind. Die Terme $\hat{y}_i$ stehen für die vom Modell mittels $\mathrm{softmax}$ berechneten Wahrscheinlichkeiten (siehe dazu Ergebnisse Tabelle *2-2*), mit denen es jedes mögliche Ausgabewort vorhersagt. Der Logarithmus $\log(\hat{y}_i)$ transformiert die Wahrscheinlichkeiten so, dass hohe Werte nur geringe Kosten verursachen, während niedrige Werte – insbesondere bei falschen, aber selbstbewussten Vorhersagen – stark bestraft werden. Das vorangestellte Minuszeichen kehrt das Vorzeichen der Summation um, sodass die Funktion minimiert werden kann. Insgesamt misst die Kreuzentropie damit den Abstand zwischen der vom Modell erzeugten Verteilung und der wahren Zielverteilung und stellt ein differenzierbares Optimierungskriterium dar, das die Grundlage für die Anpassung der Modellparameter im Training bildet.

Tabelle 2-3 *Beispielrechnung der Verlustfunktion*

	House	**Building**	**Cat**
$\hat{\boldsymbol{y}}_{\boldsymbol{i}}$	$= 0.7$	$= 0.2$	$= 0.1$
$\boldsymbol{y}$	$(1, 0, 0)$		
$\log(\hat{\boldsymbol{y}}_{\boldsymbol{i}})$	≈ -0.36	≈ -1.61	≈ -2.3
$\boldsymbol{L}$	$-[1 \cdot \log(0.7) + 0 \cdot \log(0.2) + 0 \cdot \log(0.1)] \approx 0.36$		

Der in Tabelle *2-3* dargestellte Rechnungsverlauf verdeutlicht, dass in die Berechnung des Kreuzentropie-Verlustes ausschließlich die Wahrscheinlichkeit des korrekten Zielwortes eingeht. Da im One-Hot-Vektor nur „house“ den Wert 1 trägt, reduziert sich der Verlust auf ≈ 0.36. Damit wird sichtbar, dass das Modell zwar das richtige Wort mit der höchsten Wahrscheinlichkeit gewählt hat, die Vorhersage aber noch nicht optimal ist, da eine perfekte Übersetzung erst bei einer Wahrscheinlichkeit nahe 1 erreicht wäre. Durch die Kombination von $\mathrm{softmax}$ und Kreuzentropie entsteht somit ein konsistentes Trainingsverfahren: $\mathrm{softmax}$ wandelt die internen Rohwerte in Wahrscheinlichkeiten über das Vokabular um. Gleichzeitig gibt die Kreuzentropie als Verlustfunktion an, wie weit diese Wahrscheinlichkeiten von den wahren Zielwerten entfernt sind. Über Backpropagation wird dieser Fehler auf die Parameter des Netzes zurückgeführt und dient als Signal für die Anpassung der Gewichte. Auf diese Weise verknüpfen neuronale Übersetzungsmodelle die Wahrscheinlichkeiten einzelner Logits mit einem klaren Optimierungskriterium, sodass sie über viele Iterationen hinweg lernen, die richtige Wortwahl immer präziser zu treffen.

Das von Sutskever et al. (2014) eingeführte **Encoder-Decoder-Modell** stellte den ersten Durchbruch für die Anwendung neuronaler Netze in der Übersetzung dar. Die Grundidee besteht darin, die gesamte Eingabesequenz Schritt für Schritt in einem Encoder einzulesen, der den Satz in eine komprimierte Vektorrepräsentation überführt. Diese Repräsentation wird anschließend an einen Decoder übergeben, der daraus die Zielsprache generiert. Die Verarbeitung im Encoder mit einem rekurrenten neuronalen Netz (RNN) lässt sich folgendermaßen darstellen.

$$h_t = \mathrm{sigm}(W^{hx} x_t + W^{hh} h_{t-1})$$

Im Encoder eines Sequenz-zu-Sequenz-Modells wird die Eingabesequenz Wort für Wort verarbeitet und dabei in eine interne Repräsentation überführt. Jedes Wort x_t wird zunächst als Vektor dargestellt, der semantische und syntaktische Eigenschaften des Wortes kodiert. Dieser Vektor dient als Eingabe für das rekurrente Netz, das zugleich den verborgenen Zustand h_{t-1} aus dem vorherigen Schritt übernimmt. In diesem Zustand ist die Information aller bislang gelesenen Wörter gespeichert, sodass nicht nur das aktuelle Wort, sondern auch der bisherige Kontext berücksichtigt wird. W^{hx} und W^{hh} sind Gewichtsmatrizen, wie stark die Eingabe x_t den aktuellen Zustand beeinflusst bzw. wie stark der vorherige Zustand h_{t-1} in den neuen Zustand eingeht. Mit Hilfe einer nichtlinearen Aktivierungsfunktion – hier $\mathrm{sigmoid}$, auch $\tanh$ wäre möglich – wird dafür gesorgt, dass das Modell komplexe Abhängigkeiten abbilden kann und nicht nur lineare Verknüpfungen berechnet. h_t ist der neue, verborgene Zustand zum Zeitpunkt t. Er enthält sowohl die Informationen des Wortes x_t als auch die des bisherigen Kontexts h_{t-1}. Die nachfolgende Tabelle zeigt die schrittweise Berechnung des Encoders mit vereinfachten Zufallswerten für den Eingabevektor. Zusätzlich wird $W^{hx} = \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & 0.7 \end{bmatrix}$ und $W^{hh} = \begin{bmatrix} 0.6 & 0.1 \\ 0.4 & 0.5 \end{bmatrix}$ gesetzt und $\mathrm{sigmoid}\ \sigma(z) = \frac{1}{1+e^{-1}}$ gewählt.

Tabelle 2-4 *Beispielrechnung Encoder-Funktion*

t	x_t (Wort)	x_t (Vektor)	h_{t-1}	**Berechnung**	h_t
0	–	–	–	*Initialisierung*	(0, 0)
1	„Das“	(0.1, 0.3)	(0, 0)	$\sigma[W^{hx} \cdot (0.1, 0.3)]$	(0.53, 0.56)
2	„Haus“	(0.4, 0.7)	(0.53, 0.56)	$\sigma[W^{hx} \cdot (0.4, 0.7) + W^{hh} \cdot (0.53, 0.56)]$	(0.67, 0.74)
3	„ist“	(0.3, 0.2)	(0.67, 0.74)	$\sigma[W^{hx} \cdot (0.3, 0.2) + W^{hh} \cdot (0.67, 0.74)]$	(0.66, 0.69)
4	„groß“	(0.6, 0.5)	(0.66, 0.69)	$\sigma[W^{hx} \cdot (0.6, 0.5) + W^{hh} \cdot (0.66, 0.69)]$	(0.70, 0.74)

Der vom Encoder erzeugte Satzvektor h_T dient im Encoder–Decoder-Modell als Ausgangspunkt für den Decoder, der schrittweise die Zielsprache generiert. Auch der Decoder arbeitet mit einem rekurrenten Netz, dessen Zustand s_t sich aus drei Informationsquellen speist: dem zuletzt ausgegebenen Wort, dem vorherigen Decoder-Zustand sowie dem Encoder-Vektor.

$$s_t = \mathrm{sigm}\left(W^{ys} y_{t-1} + W^{ss} s_{t-1} + W^{hs} h_T\right)$$

Hierbei steht y_{t-1} für das zuletzt erzeugte Wort der Zielsprache, dargestellt als Vektor. Der Term $W^{ys}y_{t-1}$ beschreibt, wie stark dieses Wort den neuen Decoder-Zustand beeinflusst. s_{t-1} ist der bisherige Zustand des Decoders; die Matrix W^{ss} gewichtet den Beitrag der bereits aufgebauten Zielsequenz. Schließlich trägt $W^{hs}h_T$ die Information aus dem Encoder bei, die den vollständigen Kontext des Quellsatzes enthält. Wie im Encoder wird auch hier eine nichtlineare Aktivierungsfunktion verwendet, um komplexe Abhängigkeiten abzubilden.

Aus dem resultierenden Zustand s_t werden im nächsten Schritt die sogenannten Logits $o_t = W^o s_t$ berechnet. Die Ausgabematrix W^o projiziert damit den Decoder-Zustand $s_t$in den Ausgaberaum. Der Ausgaberaum ist ein Vektor, dessen Dimensionen gleich der Größe des Zielvokabulars ist. Wenn das Zielvokabular z. B. 10.000 Wörter umfasst, dann ist W^o eine Matrix der Größe $10.000 \times d \times d$, wobei d die Dimension des Decoder-Zustands ist (z. B. 512 oder 1024). Diese werden über eine softmax-Funktion in Wahrscheinlichkeiten für jedes mögliche Ausgabewort $\hat{y}_t$ transformiert. Das Modell wählt das Wort mit der höchsten Wahrscheinlichkeit und gibt es als nächste Zielspracheinheit aus. Dieser Vorgang wiederholt sich so lange, bis ein Endsymbol <EOS> generiert wird.

Der im Encoder erzeugte Satzvektor $h_T = (0.70, 0.74)$ wird in den Decoder übernommen. In gleicher Weise zum Encoder sind die Werte der Matrizen vereinfacht und dienen der besseren Darstellung. In realen Übersetzungssystemen werden sie nicht manuell festgelegt, sondern während des Trainingsprozesses anhand großer Mengen paralleler Sprachdaten iterativ optimiert. Zusätzlich wird sigmoid $\sigma(z) = \frac{1}{1+e^{-1}}$ wieder als Aktivierungsfunktion gewählt, dazu die Gewichtsmatrizen $W^{ys} = \begin{bmatrix} 0.3 & 0.1 \\ 0.2 & 0.4 \end{bmatrix}$, $W^{ss} = \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & 0.6 \end{bmatrix}$ und $W^{hs} = \begin{bmatrix} 0.4 & 0.3 \\ 0.2 & 0.5 \end{bmatrix}$.

Tabelle 2-5 *Beispielrechnung Decoder-Funktion*

t	y_{t-1}	s_t	W^o	$W^o \cdot s_t \rightarrow$ softmax	Mini-Vokabular	$\hat{y}_t$
0	–	–	–	*Initialisierung*	–	–
1	(0, 0)	(0.731, 0.736)	$\begin{bmatrix} 3.0 & 0.5 \\ 0.2 & 0.6 \\ 0.1 & 0.3 \end{bmatrix}$	$\approx (2.56, 0.59, 0.29)$ $\rightarrow (0.81, 0.11, 0.08)$	{the, a, this}	„The“
2	(0.3, 0.4)	(0.792, 0.767)	$\begin{bmatrix} 2.5 & 0.8 \\ 0.7 & 0.3 \\ 0.1 & 0.2 \end{bmatrix}$	$\approx (2.52, 0.76, 0.23)$ $\rightarrow (0.78, 0.14, 0.08)$	{house, building, cat}	„house“
3	(0.4, 0.6)	(0.790, 0.778)	$\begin{bmatrix} 2.2 & 0.9 \\ 0.8 & 0.2 \\ 0.1 & 0.1 \end{bmatrix}$	$\approx (2.42, 0.78, 0.16)$ $\rightarrow (0.77, 0.15, 0.08)$	{is, was be,}	„is“
4	(0.2, 0.1)	(0.776, 0.782)	$\begin{bmatrix} 2.3 & 0.7 \\ 1.1 & 0.2 \\ 0.2 & 0.1 \end{bmatrix}$	$\approx (2.26, 0.98, 0.23)$ $\rightarrow (0.71, 0.20, 0.09)$	{big, large, fat}	„big“
5	(0.5, 0.5)	(0.806, 0.783)	–	–	{<EOS>}	<EOS>

Der finale Zustand h_T dient als komprimierte Repräsentation des gesamten Satzes und wird dem Decoder übergeben, der daraus die Ausgabesequenz generiert. Dieses Verfahren war zwar prinzipiell leistungsfähig, stieß jedoch bei langen Sätzen schnell an Grenzen: Da alle Informationen in einem einzigen Vektor gespeichert werden mussten, gingen semantische Details verloren, wodurch die Qualität der Übersetzung deutlich abnahm.

Als weiterer entscheidender Fortschritt ist die Einführung des **Aufmerksamkeits-Mechanismus** (engl. mechanism of attention) durch Bahdanau et al. (2014) zu betrachten. Anstatt den gesamten Eingabesatz in einem einzigen Vektor zu verdichten, kann das Modell für jedes Ausgabewort dynamisch gewichten, auf welchen Teil der Eingabe es besonders achten soll. Zunächst wird ein Alignment-Score e_{ij} berechnet, der die Relevanz des Eingabewortes j für das aktuelle Ausgabewort i beschreibt. Dazu werden der Decoder-Zustand s_{i-1} aus dem vorherigen Schritt und der Encoder-Zustand h_j des jeweiligen Eingabewortes miteinander verglichen. Eine spezielle Funktion $a(\cdot)$ berechnet daraus einen Rohwert.

$$e_{ij} = a(s_{i-1}, h_j)$$

Dieser Wert gibt an, wie stark das Eingabewort x_j für die Erzeugung des nächstes Ausgabewortes y_i relevant ist. Ein hoher Wert bedeutet eine starke inhaltliche Nähe, ein niedriger Wert eine geringere Relevanz. Da die Werte aus e_{ij} noch keine Wahrscheinlichkeiten abbilden, müssen sie mit der $\mathrm{softmax}$-Funktion normalisiert werden. Dadurch entstehen gewichtete attention values a_{ij}, die zwischen 0 und 1 liegen und sich über alle Eingabewörter hinweg zu 1 summieren.

$$a_{ij} = \frac{\exp(e_{ji})}{\sum_k \exp(e_{ik})}$$

Die $\mathrm{softmax}$-Normierung stellt sicher, dass die relative Bedeutung der einzelnen Eingabewörter klar gewichtet wird. Ein Eingabewort mit besonders hohem Score erhält nach dieser Transformation eine entsprechend große Wahrscheinlichkeit, während weniger relevante Wörter in den Hintergrund treten. In einem dritten Schritt wird aus den attention values ein Kontextvektor berechnet. Dazu werden alle Encoder-Zustände h_j mit den jeweiligen Gewichten a_{ij} multipliziert und anschließend aufsummiert.

$$c_i = \sum_j a_{ij} \cdot h_j$$

Der Kontextvektor c_i fasst somit die relevanten Informationen aus dem Quellsatz für die Erzeugung des aktuellen Ausgabewortes y_i zusammen. Er betont die wichtigen Teile der Eingabe und blendet irrelevante Bestandteile weitgehend aus.

Ein Beispiel verdeutlich diesen Ablauf: Soll der deutsche Satz „Das Haus ist groß“ ins Englische übersetzt werden und der Decoder möchte das Wort „house“ erzeugen, dann berechnet das Modell zunächst Scores für alle Eingabewörter. Während „Haus“ einen hohen Wert erhält, sind „Das“, „ist“ und „groß“ weniger relevant. Nach der $\mathrm{softmax}$-Normierung liegt der größte Anteil der Gewich-

tung auf „Haus“, sodass der Kontextvektor c_i im Wesentlichen von dessen Repräsentation geprägt ist. Dies führt dazu, dass das Modell eine gewichtete Kontextualisierung der Übersetzung zulässt.

Den bislang größten Fortschritt markierte die Entwicklung der **Transformer-Architektur** durch Vaswani et al. (2017). Der Transformer ersetzt rekurrente Strukturen vollständig durch Attention-Mechanismen. Der zentrale Baustein ist die Self-Attention, die es jedem Wort erlaubt, Abhängigkeiten zu allen anderen Wörtern im Satz herzustellen.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V$$

Dabei sind Q die Queries, also Anfragen, die vom aktuellen Wort ausgehen. Wenn der Decoder beispielsweise das Wort „house“ erzeugen will, stellt dessen Query die Anfrage, welche Wörter aus dem Quellsatz für diese Entscheidung relevant sind. Die Keys K sind die dazugehören Schlüssel aller Wörter in der Eingabesequenz. Jedes Wort im Quellsatz trägt einen eigenen Key-Vektor, der beschreibt, in welchem semantischen oder syntaktischen Kontext es steht. Sie bestimmen, wie gut eine Query zu einem bestimmten Eingabewort passt. In den Values V ist die eigentliche Information der Wörter enthalten, die später im Kontextvektor übernommen wird. Durch die Normierung mit $\sqrt{d_k}$, der Dimension der Key-Vektoren, wird numerische Stabilität gewährleistet.
Da der Transformer im Gegensatz zu rekurrenten Netzen keine zeitliche Abfolge mehr Schritt für Schritt abarbeitet, fehlt ihm zunächst eine Information, die für Sprache elementar ist: die Reihenfolge der Wörter. Während ein RNN implizit durch seine rekursive Struktur weiß, dass „Das“ vor „Haus“ kommt, betrachtet der Transformer alle Wörter gleichzeitig. Damit das Modell dennoch zwischen „Haus ist groß“ und „Groß ist Haus“ unterscheiden kann, müssen die Positionen der Wörter explizit in die Repräsentation eingebaut werden. Dies geschieht über Positional Encodings. Die Grundidee besteht darin, zu jedem Eingabevektor eine Positionsinformation hinzuzufügen. Diese wird nicht durch einfache Zähler (1, 2, 3, …) dargestellt, sondern durch eine Kombination aus Sinus- und Cosinus-Funktionen, die unterschiedliche Wellenlängen abbilden.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \qquad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

Hierbei ist d_{model} die Dimension des Embeddings. Durch die Potenzen von $10000^{2i/d_{model}}$ entstehen Funktionen mit unterschiedlichen Frequenzen: In den niedrigen Dimensionen sind die Wellenlängen lang und verändern sich nur langsam, in den höheren Dimensionen sind sie kurz und verändern sich sehr schnell. Jede Dimension schwingt also unterschiedlich, sodass die Kombination der Dimensionen eine eindeutige Signatur für jede Position im Satz ergibt.

Folgendes Beispiel verdeutlich diesen Zusammenhang: Für die Position $pos = 1$ (das erste Wort im Satz) erhält man Werte, die anders aussehen als für $pos = 2$ (zweites Wort). Je weiter man im Satz fortschreitet, desto mehr verschieben sich die Sinus- und Cosinuswerte. Dadurch entsteht ein Muster, das dem Modell nicht nur die absolute Position jedes Wortes mitteilt, sondern auch die Relationen zwischen Positionen kodiert.

2.2.3 Herausforderung bei der Verarbeitung historischer Sprachen

Die maschinelle Übersetzung historischer Sprachen wie des Biblisch-Hebräischen oder des Altgriechischen steht vor besonderen Schwierigkeiten, die weit über die Herausforderungen moderner Sprachpaare hinausgehen. Ein grundlegendes Problem besteht darin, dass es sich bei beiden Sprachen um tote Sprachen handelt, die über keine lebendige Sprachgemeinschaft mehr verfügen. Damit fehlt eine natürliche Referenzinstanz, an der Übersetzungen überprüft oder semantische Nuancen validiert werden können. Stattdessen basieren alle Rekonstruktionen auf überlieferten Handschriften, die zudem selbst erhebliche Varianz aufweisen (Rösel, 2006). Infolgedessen existiert kein eindeutig verbindlicher Urtext, sondern eine Vielzahl unterschiedlicher Textzeugen, die voneinander abweichen und deren Vergleich wiederum interpretative Entscheidungen erfordert. Das Altgriechische wird an dieser Stelle durch einige weitere, nicht-biblische Textzeugen gestützt, die Übersetzungen der LXX an bestimmten stellen unterstützen können. Dagegen steht das Althebräisch, von dem es nur wenig originale Überlieferungen gibt. Mehrdeutigkeit einzelner Worte und semantische Nuancen müssen mit ebenfalls zu übersetzenden Texten, z. B. aus dem Altgriechischen mit Hilfe der LXX oder dem Lateinischen durch die Biblia Sacra Vulgata, validiert werden. So entsteht ein hermeneutisches Validierungsdreieck.

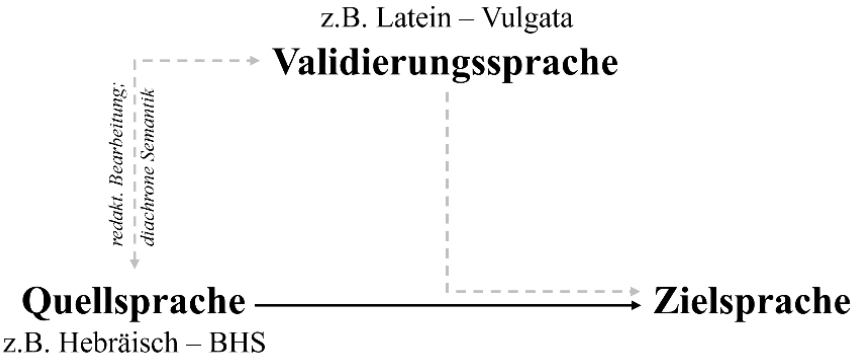


Abbildung 2-1 *Hermeneutischen Validierungsdreiecks (eigene Darstellung)*

Dennoch sei gesagt, dass redaktionelle Unterschiede, diachrone Semantik sowie sprachgeschichtliche Morphologie die Semantik der alten Sprachen verändert hat und Wörter in ihrer Bedeutungsvielfalt und Ambiguität gewonnen haben.

Hinzu kommen spezifische sprachliche Eigenschaften. Das Biblische Hebräisch zeichnet sich durch ein stark wurzelbasiertes Lexikon aus, bei dem die meisten Lexeme von trilateralen Wurzeln, bestehend aus drei Radikalen, abgeleitet werden (Jenni, 1981). Dadurch entsteht zwar eine hohe interne Kohärenz, zugleich jedoch auch eine Vielzahl morphologischer Varianten, die maschinelle Systeme korrekt segmentieren, analysieren und interpretieren müssen. Komplexität birgt auch das Tempussystem: Anders als in indogermanischen Sprachen werden nicht chronologische Zeiten, sondern Aspekte (z. B. abgeschlossene und unabgeschlossene Handlungen) markiert, was in der Übersetzung in moderne Sprachen kontextabhängig Entscheidungen erforderlich macht (Neef, 2018). Das Altgriechische wiederum ist eine hochflektierende Sprache, deren grammatischer For-

menreichtum mit Kasus, Aspekten, Modi und Partizipialkonstruktionen eine Vielzahl syntaktischer Möglichkeiten eröffnet (Decker, 2014). Beide Sprachen verfügen zudem über eine stark kontextgebundene Semantik: Ein und dasselbe Wort kann je nach literarischem Umfeld und historischer Schicht unterschiedliche Bedeutungen haben. Diese semantische Breite erschwert es neuronalen Modellen, eindeutige Entsprechungen zu erlernen, da Korpusdaten in der Regel nicht groß genug sind, um alle möglichen Kontexte abzudecken.

Ein weiteres Problemfeld betrifft die orthografische und textliche Uneinheitlichkeit. Im Hebräischen fehlen in den älteren Textzeugen häufig die Vokalzeichen, sodass Wörter ohne explizite Markierung mehrdeutig bleiben. Erst die masoretische Tradition fügte Vokalisation hinzu, die jedoch selbst interpretative Elemente enthält (Rösel, 2006). Zur Verdeutlichung: Der älteste, vollständig vokalisierte Quelltext wird als Codex Leningradensis bezeichnet, diente als Vorlage für die BHS und stammt aus den Jahren 1008/09 (Fischer, 1998). Die Qumranfunde belegen zwar die Genauigkeit der Überlieferungsarbeit der Masoreten, können die Ambiguität vieler Lexeme jedoch nicht aufheben. Im Griechischen der Septuaginta zeigt sich wiederum, dass Übersetzungen nicht nur sprachliche, sondern auch inhaltliche Abweichungen enthalten können. Diese betreffen sowohl den Umfang als auch die Anordnung und den semantischen Gehalt einzelner Bücher. So ist etwa das Buch Jeremia in der griechischen Fassung rund ein Siebtel kürzer als der masoretische Text. Anders im Buch Ester: dort enthält die Septuaginta ganze Passagen, die im Hebräischen fehlen. Für maschinelle Übersetzungssysteme bedeutet dies, dass sie nicht nur mit unterschiedlichen Sprachsystemen umgehen müssen, sondern auch mit textkritischen Divergenzen, die in den Trainingsdaten als vermeintlich gleiche Textstellen enthalten sind. Modelle können dadurch lernen, dass ein Vers im Hebräischen und Griechischen nicht immer deckungsgleich sein muss, sondern unterschiedliche semantische Inhalt transportieren kann. Die Eingabedaten weisen daher selbst eine hohe Varianz auf, die weder eindeutig normiert noch systematisch aufgelöst sind.

Eng damit verbunden ist die Frage nach der Datenlage. Während moderne Sprachen von großen, parallel annotierten Korpora profitieren, stehen für das Biblische-Hebräisch oder die Koine nur sehr kleine Textbestände zur Verfügung. Zwar lässt sich der Textbestand des Altgriechischen auf andere Sprachstufen, wie etwa die klassische oder archaische Phase, ausweiten und weitere literarische Varietäten, etwa Prosa und Lyrik, miteinbringen. Dennoch kann auch damit nicht erforderliche Datenfülle für neuronale Modelle bereitgestellt werden. Selbst die Kombination aus Bibeltexten, Qumran-Schriften und spätere Übersetzungen (u.a. LXX und Vulgata) ergeben keinen ausreichenden Korpus. Dies führt zu Problemen der Data Sparity: Modelle laufen Gefahr die Übersetzung zu stark zu konstruieren und zu sehr anzupassen – Überanpassung (*engl. overfitting)*, weil die Trainingsbasis zu klein, einseitig oder unvollständig ist.

Neben Fragen der Datenlage steht die maschinelle Übersetzung alttestamentlicher Texte auch vor technischen Fragen. Beispielweise müssen Tokenizer und Lemmatizer mit Sprachen umgehen, die z. B. keine Leerzeichen zwischen Wörtern markieren oder deren Flexionssystem sehr viele For-

men generiert. Während für moderne Sprachen standardisierte Tools existieren, müssen für historische Sprachstufen individuell angepasste Werkzeuge entwickelt werden. Ansätze wie Transfer Learning und Multilingual Training können zwar genutzt werden, bergen aber einige und bedeutend mehr Risiken im Vergleich zu modernen Alltagssprachen. Standardmetriken messen die Übereinstimmung mit einer Referenzübersetzung, doch gerade bei historischen und vor allem biblischen Texten existieren mehrere Übersetzungsmöglichkeiten. Damit werden nicht nur quantitative (technische) sondern auch qualitative (theologische) Maßstäbe notwendig, die semantische Äquivalenz und theologische Nuancen berücksichtigen. Nur in Verbindung beider kann die Funktionsweise und Qualität modernen NMT-Systeme und generierender LLMs überprüft und validiert werden.

2.3 Literaturübersicht und Forschungslücke

Stand in der Forschung

Die Forschung zur maschinellen Übersetzung hat sich in den letzten Jahren stark in Richtung neuronaler, Transformer-basierter Verfahren entwickelt. Neben allgemeinen Qualitätsgewinnen in datenreichen Szenarien zeigt die Literatur für spezielle Kontexte (z. B. Low-Resource-Settings oder domänenspezifische Daten) weiterhin deutliche Herausforderungen, da Datenknappheit, Domänenbias und fehlende Referenzstandards die Übertragbarkeit von Methoden und Ergebnissen begrenzen (Ranathunga et al., 2023). Für historische Sprachen kommt hinzu, dass Textüberlieferung, Editionen und Übersetzungstraditionen die Datenbasis sowie die Definition von Referenzqualität zusätzlich verkomplizieren (Sommerschield et al., 2023).

Ein wiederkehrender Schwerpunkt im Forschungsstand ist die Frage nach Kontexttreue. Viele MT-Systeme arbeiten satzweise, wodurch diskursive Bezüge, Referenzen, Terminologie-Kohärenz oder Argumentationszusammenhänge nur eingeschränkt erfasst werden. Entsprechend betont die Forschung seit Jahren die Grenzen satzbasierter Übersetzung und diskutiert dokument- bzw. kontextbasierte Verfahren, um Kohärenz und Konsistenz über Satzgrenzen hinweg besser abzubilden (Maruf et al., 2021). Auch aktuelle Übersichtsarbeiten heben hervor, dass kontextuell angemessene Übersetzungen gerade bei semantisch dichten Domänen weiterhin schwierig bleiben (Naveen & Trojovský, 2024). Auch im Bereich der Evaluation zeigt der Forschungsstand ein konsistentes Bild: Automatische Metriken sind nützlich für Standardisierung und Vergleichbarkeit, erfassen jedoch zentrale Fehlerarten wie Auslassungen, Bedeutungsverschiebungen, Paraphrasen oder Zusätze nur unvollständig. Daher wird zunehmend eine phänomen- und fehlerorientierte Betrachtung, unter anderem über Challenge Sets, sowie mehr Transparenz bzw. Erklärbarkeit von Metriken gefordert (Moghe et al., 2025). Mit dem Aufkommen von LLMs rücken zusätzlich Fehlerphänomene in den Fokus, die als Halluzinationen beschrieben werden. Hierzu wurden jüngst eigene Taxonomien und Benchmarks vorgeschlagen, um solche Fehler bei LLM-basierter Übersetzung systematisch zu diagnostizieren (X. Wu et al., 2025).

In Bezug auf antike/biblische Sprachen ist die direkte Studienlage im Vergleich zu modernen

Sprachpaaren deutlich geringer. Es existieren zwar Arbeiten in angrenzenden Bereichen, etwa Fallstudien zu biblisch-nahen historischen Sprachpaaren, z. B. Aramäisch–Hebräisch (Liebeskind et al., 2021), oder Untersuchungen zu interlinearer Übersetzung antiker griechischer Texte (Rapacz & Smywiński-Pohl, 2025). Diese Arbeiten verfolgen jedoch meist andere Zielsetzungen als eine domänenspezifische Übersetzungsbewertung unter hermeneutischer Perspektive (Sommerschield et al., 2023).

Forschungslücke und kontextualisierte Zielsetzung

Aus dem beschriebenen Stand der Forschung ergeben sich für diese Arbeit zwei zentrale Lücken. Erstens fehlt häufig eine domänenspezifische, nachvollziehbare Evaluation maschineller Übersetzungen biblischer Texte, die nicht ausschließlich auf Metriken basiert, sondern kontextbezogene Kriterien (z. B. Kohärenz, konsistente Schlüsselbegriffe, interpretative Stabilität) systematisch einbezieht—obwohl die Literatur die Grenzen klassischer Evaluationsmetriken und die Notwendigkeit transparenter bzw. phänomenorientierter Evaluation deutlich herausstellt.
Zweitens gibt es nur wenige Arbeiten, die den Unterschied zwischen NMT (Transformer Encoder–Decoder) und LLM-basierten Übersetzungsansätzen (Decoder-only Transformer) in diesem Domänenkontext systematisch entlang typischer Fehlerphänomene (z. B. Auslassung, Zusatzinformation/Halluzination, Bedeutungsverschiebung) vergleichen. Zwar liegen methodische Beiträge zur Diagnose von Halluzinationen bei LLM-Übersetzung vor und es existieren Einzelarbeiten zu antiken/biblisch-nahen Sprachen. Ein geschlossenes Forschungsdesign, das biblische Ursprachen, einen systematischen NMT-vs.-LLM-Vergleich und kontext- sowie hermeneutik-sensitive qualitative Kriterien kombiniert, ist in Literatur jedoch nicht als etablierter Standardansatz sichtbar.

Das Ziel ist daher, die Leistungsfähigkeit ausgewählter Systeme für die Übersetzung ausgewählter biblischer Textstellen zu untersuchen und dabei Kontexttreue und interpretative Zurückhaltung als zentrale Qualitätsdimensionen herauszustellen. Die Arbeit trägt damit zur Einordnung aktueller MT- und LLM-Ansätze in einem Bereich bei, in dem direkte empirische Evidenz bislang begrenzt ist, und schafft eine Grundlage für weiterführende Forschung zu Evaluation und Domänenanpassung in der Übersetzung historischer und religiöser Texte.

3 Methodik

Die vorliegende Arbeit verfolgt einen methodischen Ansatz, der sowohl technische als auch theologisch-linguistische Dimensionen berücksichtigt. Ausgangspunkt ist die Hypothese, dass Qualität maschineller Übersetzung alttestamentlicher Texte nicht allein an technischen Metriken gemessen werden kann, sondern auch eine hermeneutische Bewertung erfordert. Methodisch wird daher ein mehrstufiges Vorgehen gewählt, das auf einer triangulären Analyse basiert.
Statt Übersetzung lediglich als Relation von Quellsprache und Zielsprache zu begreifen, wird ein drittes Bezugssystem einbezogen: die Evaluierungssprache. Sie wird durch die Septuaginta und die Vulgata repräsentiert und folgt dem Prinzip des bereits bekannten hermeneutischen Validierungsdreiecks. Dieses geht davon aus, dass sich der Gehalt eines Textes nur im Abgleich mehrerer sprachlicher und kultureller Kontexte angemessen erschließen lässt. Entsprechend wird jede untersuchte Perikope in drei Vergleichsebenen parallelisiert:

[1] der hebräische Quelltext

[2] die maschinell generierte deutsche und englische Übersetzung

[3] die historischen Übersetzungszeugnisse in Griechisch und Latein

Die Evaluierungssprache übernimmt in dieser Analyse eine doppelte Funktion. Einerseits dient sie der diachronen Kontextualisierung, indem sie dokumentiert, wie in der Septuaginta und der Vulgata zentrale Übersetzungsentscheidungen bereits in der Antike bzw. Spätantike getroffen wurden und wie diese Tradition bis in die Gegenwart einwirken. Andererseits fungiert sie als hermeneutische Prüfgröße, da ihre Einbindung ermöglicht, die Ergebnisse maschineller Systeme nicht nur mit dem hebräischen Quelltext, sondern zugleich mit historischen und modernen Zielsprachen in Beziehung zu setzen. Konkret erfolgt dies in drei Schritten:

[1] **Hebräisch → Deutsch/Englisch** *Maschinen-Output*
Der Quelltext wird mit NMTs und LLMs ins Deutsche übertragen.

[2] **Hebräisch → Griechisch/Latein** *Maschinen-Output*
Die gleichen Systeme erzeugen Übersetzungen in die antiken Zielsprachen.

[3] **Vergleiche mit überlieferten Fassungen**
Die maschinellen Ausgaben werden anschließend sowohl mit den überlieferten Fassungen als auch mit modernen Referenzübersetzungen kontrastiert.

Auf diese Weise entsteht eine Triangulation, in der die Übersetzungsleistung systematisch zwischen drei Bezugspunkten eingeordnet wird.
In der dargestellten Methodik lässt sich ein leichtes Ungleichgewicht finden. Während der Korpus der Validierungssprache mit Latein und Griechisch aus zwei Sprachen besteht, beinhaltet der Zielsprachenvergleich nur das Deutsche. Für eine ausgeglichene Analyse kann dem Zielsprachenkorpus das Englische und entsprechende Referenzübersetzungen hinzufügt werden.

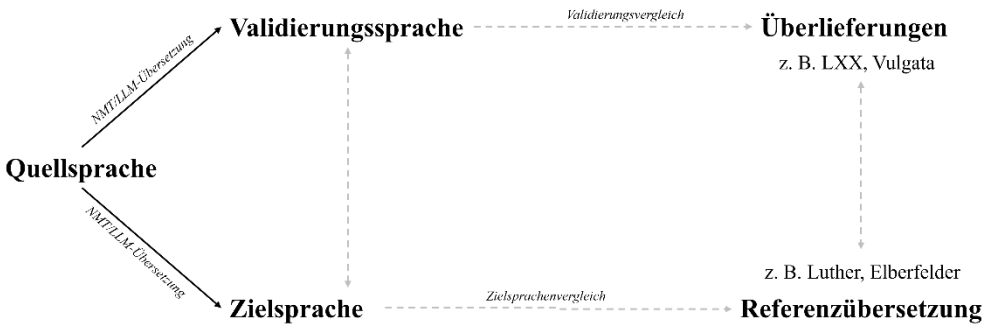


Abbildung 3-1 *eigene Darstellung der dreiseitigen Vergleichsmatrix (Triangulation)*

Diese Vergleichsmatrix – siehe Abbildung 3-1 – ist nicht bloß eine schematische Gegenüberstellung, sondern bildet die methodische Grundalge für eine vielschichtige Analyse, die sich auf drei Ebenen vollzieht:

[1] **Lexikalische Ebene**

Hier wird geprüft, ob die maschinellen Systeme bei der Wahl einzelner Begriffe der Tradition etablierter Übersetzungen folgen und diese übernehmen oder alternative Lösungen wählen, die stärker am semantischen Spektrum der Quellsprache orientiert sind.

[2] **Syntaktische Ebene**

Auf dieser Ebene geht es um die Frage, ob die Modelle Strukturen (z. B. Nominalsätze im Hebräischen) der Quellsprache beibehalten oder eigenständig anpassen. Der Abgleich mit LXX und Vulgata erlaubt es, solche Entscheidungen einzuordnen und zu bewerten, ob maschinelle Übersetzungssysteme eher die Tendenz antiker Übersetzung zur Glättung und Vereinfachung wiederholen oder neue Strukturen generieren.

[3] **Semantisch-theologische Ebene**

Gerade an Passagen mit inhärenter Mehrdeutigkeit zeigt sich die Stärke der Triangulation. Verschiedene NLP-Systeme legen je nach Trainingshintergrund den Schwerpunkt auf bestimmte Lesarten. So kann der Vergleich mit einen stark theologisch geprägten LXX bzw. Vulgata dies offenlegen. Daher eröffnet dieser Vergleich damit nicht nur Einblicke in die Übersetzungslogik der Systeme, sondern auch in der Frage, welche hermeneutischen Horizonte sie implizit reproduziert.

Durch diese dreifache Anlage wird die Analyse über eine rein sprachliche Evaluierung hinausgeführt. Die trianguläre Methodik verbindet quantitative Verfahren mit einer hermeneutischen Validierung, die die Ergebnisse in den historischen und theologischen Deutungshorizont einordnet.
So entsteht ein methodisches Fundament, in dem maschinelle Systeme nicht lediglich als technische Instrumente, sondern als Akteure im Kontinuum der Übersetzungsgeschichte sichtbar werden. Ihre Ausgaben können entweder in Kontinuität mit überlieferten Lösungen stehen, eigene Lesarten hervorbringen oder Bedeutungsnuancen nivellieren, die für eine Auslegung zentral sind.

3.1 Auswahl und Aufarbeitung der Bibeltexte

Zwei zentrale Elemente stellen die sorgfältige Auswahl und systematische Aufarbeitungen der zu analysierenden Bibeltexte dar. Da keine Einzelstellen untersucht werden sollen, sondern die Gesamtheit des alttestamentlichen Textkorpus in die Analyse einfließt, unterliegt die Auswahl einer gezielten Eingrenzung und Konkretisierung. Die Textauswahl bestimmt maßgeblich, welche sprachlichen Phänomene und theologischen Problemfelder sichtbar werden und stellt die methodische Grundlage für die gesamte Analyse dar.

Die Auswahl der Perikopen orientiert sich an zwei komplementären Zielen: Zum einen sollen sie sprachliche Vielfalt abbilden, indem unterschiedliche Gattungen (narrative, poetische und prophetische Texte) berücksichtigt werden. Zum anderen sollen sie theologische Relevanz aufzeigen, indem zentrale Begriffe, semantische Mehrdeutigkeiten oder historisch kontroverse Übersetzungsentscheidungen in den Blick genommen werden. Damit wird gewährleistet, dass die Analyse sowohl linguistisch als auch hermeneutisch aussagekräftig ist.
Die methodische Triangulation bedingt für die Aufarbeitung der einzelnen Perikopen weitere Schritte. Zur benötigen Quellsprache addieren sich die Textzeugen der Validierungssprache(n). So müssen ebendiese aufbereitet werden, indem sie aus wissenschaftlichen Standardeditionen übernommen und in ein einheitliches, digitales Format übertragen werden. Entscheidend ist dabei, dass die Texte in einer Form vorliegen, die sowohl die Eingabe in maschinelle Übersetzungssysteme ermöglicht als auch eine spätere systematische Vergleichbarkeit sicherstellt. Dazu gehören die Vereinheitlichung von Orthografie, die Segmentierung nach Versen sowie die Bereinigung von editorischen Zusätzen.

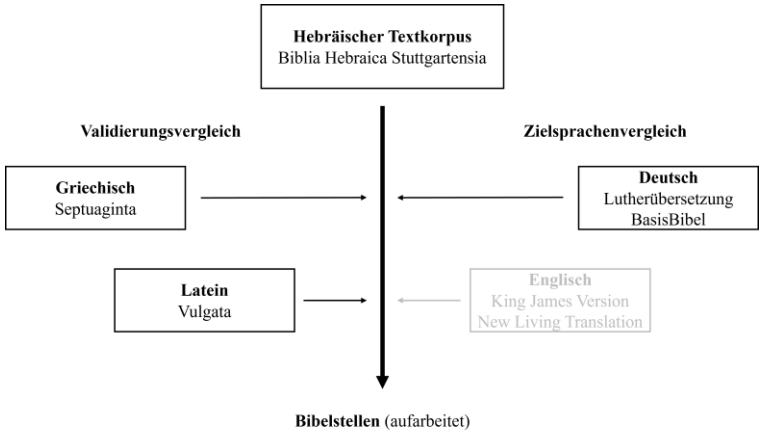


Abbildung 3-2 *Zusammenhang der aufzuarbeitenden Textkorpora (eigene Darstellung)*

Die Auswahl und Aufarbeitung der Bibeltexte bildet somit eine methodische Doppelbewegung: einerseits die inhaltlich-theologische Begründung der Perikopenwahl, andererseits die technische Vorbereitung für die computergestützte Übersetzung und Analyse. Beide Dimensionen sind eng miteinander verknüpft, da nur eine methodisch kontrollierte Textbasis eine valide Evaluierung maschineller Übersetzungen erlaubt.

3.1.1 Kriterien für die Textauswahl

Die Auswahl geeigneter Textstellen stellt einen methodisch entscheidenden Schritt dar, da sie bestimmt, welche sprachlichen Phänomene und theologischen Problemstellungen sichtbar werden. Eine zufällige oder rein pragmatische Auswahl wäre in diesem Kontext nicht tragfähig. Vielmehr ist es notwendig die Auswahl durch klare, nachvollziehbare Kriterien abzusichern, die sowohl die exegetische Methodik als auch die Übersetzungswissenschaft berücksichtigt. In Anlehnung an Nida & Taber (1969) sowie Nord (1995) wird die Textauswahl nicht primär als technischer Schritt verstanden, sondern als integraler Teil des Übersetzungsprozesses, in dem sprachliche, theologische und praktische Gesichtspunkte ineinandergreifen. Zur besseren Übersicht lassen sich die Kriterien in vier Kategorien gliedern.

[1] **Allgemeine Kriterien**
meinen den Umfang und die Repräsentativität der Gattung.

[2] **Sprachlich-linguistische Kriterien**
summieren vor allem morphologische Vielfalt, semantische Ambiguität, syntaktische Eigenheiten und Stilistik.

[3] **Theologische Kriterien**
fassen Schlüsselbegriffe, Rezeptionsgeschichte und hermeneutische Relevanz zusammen.

[4] **Praktisch-methodische Kriterien**
bestehen aus Machbarkeit, Vergleichbarkeit und Möglichkeiten zur Triangulation.

Allgemeine Kriterien

Ein grundlegendes Kriterium für die Auswahl betrifft den Umfang der Textstelle. Einzelverse sind für eine Untersuchung dieser Art nicht hinreichend, da sie zu wenig semantischen und syntaktischen Kontext liefern. Zugleich wäre ein ganzes Kapitel zu umfangreich, da so eine detaillierte Analyse verwischt werden und den Fokuspunkt verschieben würde. Daher wird eine Länge von fünf bis zehn Versen angesetzt, die sowohl literarisch geschlossene Einheiten darstellen als auch eine detaillierte Analyse erlauben. Für exegetische Untersuchung kann es ebenso von Vorteil sein, wenn Textabschnitte so gewählt sind, dass sie inhaltlich abgeschlossen, aber noch überschaubar sind. Neben dem Umfang ist auch die Repräsentativität entscheidend. Die ausgewählten Texte sollten nicht zufällig oder thematisch willkürlich gewählt sein, sondern exemplarisch die sprachliche und theologische Breite des Alten Testaments widerspiegeln. Aus diesem Grund werden narrative, prophetische und poetische Texte in die Analyse mit einbezogen. Diese Dreigliederung ist nicht nur literarisch sinnvoll, sondern erlaubt auch, unterschiedliche sprachliche und stilistische Merkmale in den Blick zu nehmen, die von den Gattungen in unterschiedlicher Weise bedingt werden. Gemeint sind z. B. die lineare Handlungsstruktur narrativer Texte, rhetorische Dichte prophetischer Rede sowie die die verdichtete Bildsprache poetischer Texte.

Sprachlich-linguistische Kriterien

Ein zentrales Ziel der Untersuchung besteht darin, die Leistungsfähigkeit maschineller Übersetzungssysteme im Umgang mit den spezifischen sprachlichen Strukturen des Hebräischen, Griechischen und Lateinischen zu prüfen. Dafür ist es notwendig, Texte auszuwählen, die ein möglichst breites Spektrum linguistischer Phänomene abdecken.

Besonders wichtig ist die morphologische Vielfalt der Texte. Das biblische Hebräisch kennt ein aspektorientiertes Verbsystem, dessen Bedeutung stark kontextabhängig ist. Jenni (1981) und Neef (2018) zeigen, dass die Interpretation von Perfekt- und Imperfektformen zu den größten Herausforderungen der Hebraistik zählt. Das dagegen deutlich komplexere Tempussystem des Griechischen bzw. die Koine der LXX transportiert semantische Feinheiten jenseits reiner Zeitstufen (Siebenthal, 2008). Dazukommend stellt die semantische Ambiguität ein weiteres wichtiges Kriterium dar. Dass viele biblische Schlüsselbegriffe ein breits Bedeutungsspektrum besitzen und deshalb nicht deckungsgleich ins Deutsche oder Englische übertragen werden können, wurde bereits erwähnt. Texte, die solche Termini enthalten, sind besonders geeignet, um die Kontextsensibilität von Übersetzungssystemen zu prüfen.

Auch syntaktische Einheiten sind relevant. Hebräische Nominalsätze ohne Kopula oder griechische Partizipialkonstruktionen stellen besondere Herausforderungen dar, da sie nicht ohne Weiteres in die Strukturen der deutschen oder englischen Sprache übertragbar sind. Schließlich gilt es, stilistische Besonderheiten zur berücksichtigen. Poetische Texte mit parallelismus membrorum und dichter Bildsprache erfordern andere Übersetzungsstrategien als narrative Texte mit linearer Handlung oder prophetische Texte mit rhetorischer Verdichtung.

Theologische Kriterien

Da es sich bei den untersuchten Texten um biblische Schriften handelt, sind theologische Aspekte für die Auswahl unverzichtbar. Von besonderem Interesse sind Schlüsselbegriffe, deren Übersetzungen das theologische Verständnis direkt prägen. Trotz der Ähnlichkeit allgemeingültiger Ambiguität ist der theologische Bedeutungskontext eine wichtige Validierungsgröße moderner, maschineller Übersetzungssysteme. So können neben der reinen Übersetzungsleistung interpretative Spielräume nicht nur linguistisch, sondern auch theologisch bewertet werden.

Darüber hinaus spielt die Rezeptionsgeschichte eine zentrale Rolle. Die LXX und Vulgata sind selbst Zeugnisse interpretierender Übersetzungen. Ein Vergleich der maschinellen Übersetzung mit verschiedene Zielsprachen kann so die Bedeutung der Rezeptionsgeschichte und anderen exegetischen Methoden hervorheben. So sind die LXX und Vulgata nicht nur Übersetzungen, sondern auch Versionen erster Bibelkommentare. Dieser interpretative Charakter macht sie für den Vergleich mit modernen, maschinellen Übersetzungssystemen besonders aufschlussreich.

Abschließend erweitert sich die Metrik zur Textauswahl der theologischen Kriterien um die hermeneutische Relevanz. Viele Texte sind intendiert mehrdeutig formuliert. Mit der korrekten Auswahl geeigneter Perikopen soll die Intention der Ambiguität durch NLP aufgelöst werden können.

Praktisch-methodische Kriterien

Neben sprachlichen und theologischen Gesichtspunkten spielen auch praktische Erwägungen eine Rolle. Da der Auswertungsrahmen begrenzt ist, muss die Analyse auf eine bewusste Anzahl an Perikopen beschränkt werden. Diese sollen exemplarisch die ersten drei Kriterienkategorien berücksichtigen und so eine Vielzahl genannter Phänomene abdecken.
Als weiteres praktisches Kriterium gilt die Vergleichbarkeit. Die ausgewählten Texte müssen in kritischen Editionen zuverlässig vorliegen und mit etablierten Referenzübersetzungen kontrastierbar sein. Nord (1995) betont, dass die Funktion von Übersetzung stets im Kontext ihrer Rezipienten und Zielkulturen verstanden werden muss. Ob moderne NMT-Systeme eine solche Kontextualisierung erreichen, soll die Analyse zwischen den Ziel- und der Validierungssprache mit Hilfe der methodischen Triangulation evaluiert werden.

Damit entsteht eine Textbasis, die sowohl sprachliche Vielfalt als auch theologische Tiefenschärfe repräsentiert und zugleich methodisch handhabbar bleibt. Sie bildet das Fundament für die weitere Analyse der Übersetzungsqualität maschineller Systeme und erlaubt eine Bewertung, die technische, linguistische und hermeneutische Dimensionen miteinander verbindet.

Die Anwendung der dargestellten Kriterien auf den alttestamentlichen Textkorpus führt zwangsläufig zu einer Eingrenzung der in Frage kommenden Textstellen. Nida & Taber (1969) betonen, dass Übersetzungsanalyse stets an einem „manageable corpus“ durchgeführt werden sollten, um die Ergebnisse differenziert und zugleich übersichtlich darstellen zu können. Die Entscheidung für drei Perikopen stellt daher einen methodischen Kompromiss dar: Sie sind ausreichend zahlreich, um verschiedene literarische Gattungen und unterschiedliche sprachliche Phänomene abzudecken, aber zugleich zu stark begrenzt, sodass jede Stelle intensiv analysiert werden kann. Während eine einzelne Perikope lediglich exemplarische Einblicke erlaubt und eine zweite den Vergleich unterschiedlicher Ergebnisse ermöglicht, schafft erst die dritte Perikope die notwendige Validität, indem sie den zuvor gewonnen Befund überprüft und absichert. Eine Konzentration auf zwei oder gar eine einzige Perikope birgt die Gefahr, dass die Ergebnisse einseitig ausfallen und nicht die gesamte Bandbreite der Übersetzungsproblematik abbilden. Die Zahl drei erweist sich damit methodisch als Mindestvoraussetzung für eine solche vergleichende als auch verifizierende Analyse.

Aus der Menge potenzieller Kandidaten ergeben sich schließlich drei Passagen, die alle genannten Kriterien erfüllen. Die nachfolgende Tabelle 3-1 gibt einen begründeten Überblick der Auswahl.

Tabelle 3-1 *Auswahl der Perikopen anhand methodischer Kriterien*

Perikope	**Gattung und Struktur**	**Linguistik**	**Theologie**
Ex 3,11–15	*Narrativ* - Erzählung in Dialogform - einheitlich und in sich abgeschlossen	- Gottesname: אֶהְיֶה אֲשֶׁר אֶהְיֶה → mehrdeut. Imperfekt - morphologische Vielfalt - Vergleich: LXX: ἐγώ εἰμι ὁ ὤν, Vulgata: ego sum qui sum - semantische Dichte und theologische Komplexität	- Offenbarung des Gottesnamen - Spannungsfeld von göttlicher Existenz - zentrale Frage nach Gottes Identität
Jes 7,10–16	*Prophetie* - eingebettet in narrative Rahmung - einheitlich und in sich abgeschlossen	- Schlüsselbegriff: עַלְמָה → semantisch ambig - prophetische Bildsprache - rhetorische Intensität - Übersetzungsproblem: LXX: παρθένος, Vulgata: virgo	- Verheißung des Immanuel-Zeichens - zentral für christologische Deutung Frage nach Gottes Treue
Ps 23,1–6	*Poesie* - parallelismus membrorum - kleine, geschlossene Einheit	- metaphor. Sprache: יְהוָה רֹעִי - parallelismus membrorum - poetische Wiederholungsstrukturen - einfacher Satzbau, aber komplexe Bildsprache - kulturspezifische Idiome	- theologische Deutung Gottes - Vertrauen und Schutz im Zentrum

3.1.2 Formatierung und Digitalisierung

Die Grundlage für die Analyse neuronaler Übersetzungssysteme bildet ein konsistenter, maschinenlesbarer Textkorpus. Dazu werden zunächst die als geeignet ausgewählten Textstellen digitalisiert, anschließend vereinheitlicht und schließlich in ein standardisiertes Datenformat überführt. Dieser mehrstufige Prozess gewährleistet, dass alle untersuchten Perikopen in den verschiedenen Sprachversionen direkt miteinander vergleichbar sind und in gleicher Form in die Übersetzungssysteme eingespeist werden können.

Auswahl der Quellen

Für die Untersuchung wurden bereits drei zentrale Perikopen des Alten Testaments genannt, die unterschiedliche Textgattungen repräsentieren: Exodus 3,11–16 (narrative Prosa), Jesaja 7,10–16 (prophetische Verkündigung) sowie Psalm 23,1–6 (poetische Dichtung). Diese Textstellen werden in mehreren maßgeblichen Sprachfassungen herangezogen: der Biblia Hebraica Stuttgartensia, der Septuaginta, der Vulgata sowie ausgewählten deutschen und englischen Übersetzungen (Lutherbibel 2017, BasisBibel sowie King James Version, New Living Translation). Letztere stellen jeweils eine traditionelle und eine freie, an die Moderne angepasste, Übersetzung dar. Damit wird einerseits die historische Mehrsprachigkeit der alttestamentlichen Überlieferung berücksichtigt, andererseits ein Vergleich mit etablierten zeitgenössischen Bibelübersetzungen ermöglicht.

Digitalisierung

Soweit möglich werden bestehende, digitale Textausgaben verwendet (z. B. frei verfügbare Digitalisate der BHS und der LXX). Für das Hebräische wird der Text in zwei Fassungen aufbereitet: eine vokalisierte Variante (mit masoretischen Vokalzeichen) sowie eine unvokalisierte Fassung (als reinen Konsonantentext). Diese Unterscheidung dient dazu, die Relevanz der Vokalisierung für neuronale Übersetzungssysteme zu prüfen. Alle Texte werden zusätzlich mit gedruckten Referenzausgaben abgeglichen, um Fehler und Abweichungen in der Transkription auszuschließen.

Formatierung

Die Texte wurden so strukturiert, dass sie für maschinelle Verarbeitung geeignet sind. Jede Perikope wurde nach dem Schema Buch – Kapitel – Vers segmentiert, wobei jeder Vers eine eigenständige Einheit bildet. Jede Sprachfassung erhält eine eindeutige ID, die das Kürzel der Textquelle, die biblische Referenz und gegebenenfalls Zusatzangaben (z. B. ob sie unvokalisiert ist) enthält. Die Dateien sind in UFT-8 kodiert, sodass auch hebräische und griechische Schriftzeichen einheitlich verarbeitet werden können. Zusätze wie Fußnoten oder textkritische Apparate wurden entfernt, um eine reine Textgrundlage sicherzustellen.
Ein kurzer Abschnitt verdeutlicht den Aufbau der gespeicherten .txt-Dateien am Beispiel der poetischen Perikope Psalm 23,1.

```
#-------------------------------------------
# Perikope: Psalm 23, Verse 1-6 | Buch: Psalmen
# UTF-8 codiert, ID-Format: <Version>_<Abkürzung>_<Kapitel>_<Vers>
#-------------------------------------------

#------- Vers 1 -------
ID: Heb_vok_Ps_23_1
Text: מִזְמוֹר לְדָוִד יְהוָה רֹעִי לֹא אֶחְסָר

ID: Heb_unvok_Ps_23_1
Text: מזמור לדוד יהוה רעי לא אחסר

ID: LXX_Ps_23_1
Text: Ψαλμὸς τῷ Δαυιδ. Κύριος ποιμαίνει με, καὶ οὐδέν με ὑστερήσει.

ID: VUL_Ps_23_1
Text: canticum David Dominus pascit me nihil mihi deerit

ID: LUTH_Ps_23_1
Text: Ein Psalm Davids. Der Herr ist mein Hirte, mir wird nichts mangeln.

ID: BASIS_Ps_23_1
Text: EIN PSALM, VON DAVID. Der Herr ist mein Hirte. Mir fehlt es an nichts.

ID: KJV_Ps_23_1
Text: A Psalm of David. The LORD is my shepherd; I shall not want.

ID: NLT_Ps_23_1
Text: The LORD is my shepherd; I have all that I need.
...
```

Abbildung 3-3 *Auszug aus einer digitalisierten Perikope (Ps23_1-6.txt)*

Dieser Aufbau erlaubt sowohl eine menschliche Lesbarkeit als auch eine einfache Weiterverarbeitung in Skripten und Übersetzungssystemen.

Standardisierung

Zur Vereinheitlichung werden alle Texte in strukturierte Formate überführt. Neben der .txt-Darstellung wird insbesondere ein .json-Format genutzt, das pro Vers die verschiedenen Sprachfassungen in einem Objekt zusammenfasst. Diese Struktur erleichtert den Import in Programmiersprachen wie Python und ermöglicht eine systematische Weiterverarbeitung z. B. Tokenisierung, Alignment oder statistische Analysen.

Die Entscheidung für eine standardisierte Datenstruktur erfolgt aus mehreren Überlegungen. Zuerst wird so gewährleistet, dass alle Sprachfassungen exakt auf Vers-Ebene miteinander korrespondieren. In dieser Weise lassen sich Übersetzungsvergleiche oder automatische Auswertungen ohne zusätzliche Abgleiche durchführen. Weiter wird durch die Vergabe von eindeutigen IDs (z. B. LXX_Jes_7_14) Transparenz und Nachvollziehbarkeit geschaffen: Jede Textstelle kann eindeutig referenziert und problemlos wiedergefunden werden. Abschließend erlaubt die konsequente Nutzung von Unicode (UTF-8) eine verlustfreie Darstellung hebräischer, griechischer und lateinischer Schriftzeichen in einem gemeinsamen Korpus.

Darüber hinaus ist die Standardisierung auch im Hinblick auf die maschinelle Übersetzung unverzichtbar. Neuronale Modelle sind in der Regel hochsensibel gegenüber Inkonsistenzen in der Eingabe. Abweichungen in der Segmentierung, unterschiedliche Zeichencodierungen oder zusätzliche

editorische Markierungen könnten die Auswertung erheblich verfälschen. Durch die Reduktion auf den reinen Bibeltext und die einheitliche Strukturierung wird sichergestellt, dass Unterschiede in den Ergebnissen tatsächlich auf sprachliche Eigenheiten und das Verhalten der NMTs und LLMs zurückzuführen sind – und nicht auf technische Inkonsistenzen der Datenbasis.
Ein kurzer Abschnitt verdeutlicht den Aufbau der gespeicherten .json-Dateien erneut am Beispiel der poetischen Perikope Ps 23,1.


```
[
  {
    "Book": "Psalms",
    "Chapter": 23,
    "Verse": 1,
    "Versions": {
      "Heb_vok": " מִזְמוֹר לְדָוִד יְהוָה רֹעִי לֹא אֶחְסָר",
      "Heb_unvok": "מזמור לדוד יהוה רעי לא אחסר",
      "LXX": "Ψαλμὸς τῷ Δαυιδ. Κύριος ποιμαίνει με, καὶ οὐδέν με ὑστερήσει.",
      "VUL": "canticum David Dominus pascit me nihil mihi deerit",
      "LUTH": "Ein Psalm Davids. Der Herr ist mein Hirte, mir wird nichts mangeln.",
      "BASIS": "EIN PSALM, VON DAVID. Der Herr ist mein Hirte. Mir fehlt es an nichts.",
      "KJV": "A Psalm of David. The LORD is my shepherd; I shall not want.",
      "NLT": "The LORD is my shepherd; I have all that I need."
    }
  },
…
]
```


Abbildung 3-4 *Auszug aus einer strukturierten Perikope (Ps23_1-6.json)*

Technische Vorbereitung

Die vorbereiteten Texte werden in einer einheitlichen, versweisen Struktur aufbereitet und ohne zusätzliche manuelle Tokenisierung oder Lemmatisierung in die Modelle eingespeist. Zwar spielen diese Verfahren grundsätzlich eine Rolle in der Sprachverarbeitung – etwa durch die Zerlegung in kleinere Einheiten (Tokenisierung) oder die Rückführung auf Grundformen (Lemmatisierung) – moderne neuronale Übersetzungssysteme und Large Language Models verfügen jedoch meist über integrierte Tokenizer, die diese Schritte automatisch durchführen.
Durch den Verzicht auf externe Vorverarbeitung bleibt das Vorgehen nah der gängigen Praxis: Die Texte werden so verwendet, wie es auch typische Anwender:innen tun würden. Dadurch ist einerseits die Vergleichbarkeit zwischen verschiedenen Modellen gewährleistet, andererseits wird sichtbar, wie die Systeme selbst mit historischen Sprachen umgehen. Auf dieser Grundlage erfolgt die Auswahl der Übersetzungsmodelle.

3.2 Auswahl der Übersetzungsmodelle

Die Auswahl geeigneter und einzusetzender Übersetzungssysteme folgt dem Ziel, die Leistungsfähigkeit spezialisierter NMTs mit generativen LLMs systematisch zu vergleichen. Anstelle eines breiten Querschnitts vieler Systeme wird der Fokus gezielt auf wenige, aber konzeptionell unterschiedliche Modelle gelegt. Diese Vorgehensweise ermöglicht es, qualitative Unterschiede zwischen domänenspezifischen und generischen Übersetzungsansätzen herauszuarbeiten – insbesondere im Hinblick auf Kontexttreue, stilistische Kohärenz und semantische Präzision bei der Übersetzung historischer Texte.

Die Entscheidung, sowohl klassische NMTs als auch LLMs einzubeziehen, beruht auf aktuellen Forschungsergebnissen, die auf eine zunehmende konzeptionelle Annäherung beider Paradigmen hinweisen. Während neuronale Übersetzungssysteme ursprünglich auf klar abgegrenzte Sprachpaare und Parallelkorpora trainiert wurden, integrieren moderne LLMs Übersetzungsprozesse als Teil eines allgemeinen Sprachverständnisses. Stahlberg (2020) beschreibt diesen Übergang als „paradigm shift“ in der maschinellen Übersetzung, bei dem die Grenze zwischen Übersetzung und Sprachmodellierung zunehmend verschwimmt. Neuere Studien bestätigen diese Entwicklung: Pang et al. (2024) sowie Enis und Hopkins (2024) zeigen, dass große Sprachmodelle in standardisierten Übersetzungstests teils mit spezialisierten NMTs konkurrieren, insbesondere wenn längere Kontexte oder semantische Ambiguitäten eine Rolle spielen.

Für die empirische Untersuchung wurden daher vier Modelle ausgewählt, die unterschiedliche Klassen und Anwendungsszenarien repräsentieren. Genauer werden zwei NMTs mit zwei LLMs verglichen und differenzierte Dimensionen in Bezug zueinander gesetzt. Zuerst soll die *Trainingsstrategie* der Analysemittelpunkt sein. So kann ein spezialisiertes NMT mit einem generischen LLM verglichen und die Übersetzungsergebnisse bewertet werden, inwiefern die Spezialisierung eines konkreten NMTs die Übersetzungsqualität eines generischen LLM überbieten kann. In einer zweiten Betrachtungsweise soll die *Implementierungsstrategie* beobachtet und bewertet werden. Mit Hilfe kleiner, lokal installierter Übersetzungssysteme, in Form eines einfachen NMT und eines einfachen LLM, sollen die Fähigkeiten dieser denen der großen, cloudbasierten Implementierungen gegenübergestellt werden. Dies soll die Bedeutung der Cloudbasiertheit bewerten und mit lokalen Systemen in einen entsprechenden Kontext gerückt werden.

Als kommerzielles NMT dient *DeepL*, das durch seinen proprietären Transformer, Framework und kontinuierliche Domänenanpassung eine hohe Übersetzungsqualität in gegenwartssprachlichen Kontexten erzielt (Koehn & Knowles, 2017). Ergänzend wird mit *NLLB* ein offenes Forschungsmodell eingesetzt. Dieses Modell steht exemplarisch für die jüngere Entwicklung wissenschaftlicher NMTs, die auf mehrsprachige Daten trainiert und zur Erschließung ressourcenschwacher Sprachen konzipiert wurden. Als Vertreter der LLM-basierten Systeme werden *Gemini 2.5 Pro* und *LLaMA-3* einbezogen. Beide Modelle basieren auf der Transformer-Architektur, unterscheiden sich aber in ihrer Ausrichtung: Gemini 2.5 Pro ist ein großskaliges, multimodales Sprachmodell mit starkem Fokus auf Übersetzung, Textverständnis und Kontextwahrung über lange Sequenzen hinweg.

Es repräsentiert damit den derzeitigen Stand kommerziell geschlossener LLM-Systeme. LLaMA-3 hingegen ist ein offenes Basismodel, das lokal betrieben und gezielt auf spezifische Datensätze oder Aufgaben feinjustiert werden kann. Dadurch bildet es die forschungsnahe Variante eines LLMs und erlaubt eine transparentere Kontrolle der Modellarchitektur und der Ausgabengenerierung.

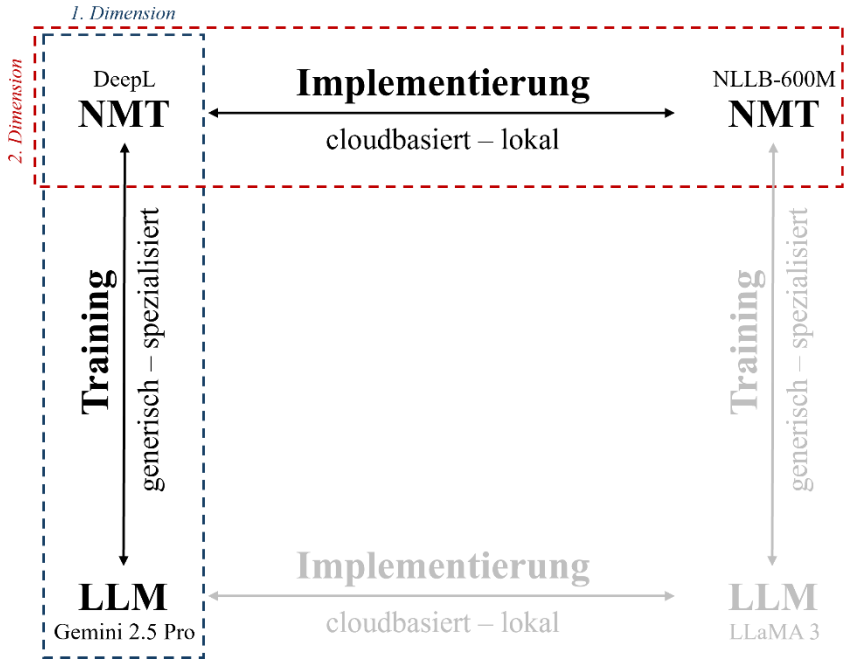


Abbildung 3-5 *Gegenüberstellung der Vergleichsdimensionen*

Die Kombination aus kommerziellen und wissenschaftlichen Systemen ermöglich es, Leistungsunterschiede entlang zweier Dimensionen zu untersuchen: **Trainingsstrategie** (spezialisiert vs. generisch) sowie **Implementierungsstrategie** (cloudbasiert vs. lokal). Ähnliche Vergleichsstudien, etwa von Sizov et al. (2024) und Wu et al. (2025), zeigen, dass diese Faktoren entscheidend für die Übersetzungsqualität, Kontexttreue und Interpretierbarkeit der Ergebnisse sind. Durch den gezielten Vergleich von DeepL und NLLB-200 einerseits und Gemini 2.5 Pro und LLaMA-3 andererseits lassen sich die jeweiligen Stärken und Grenzen der beiden Modelltypen empirisch erfassen und in Hinblick auf ihre Eignung zur Übersetzung kontextsensitiver, historischer Texte evaluieren. Diese Konstellation erlaubt es, die gesamte Bandbreite moderner NLP-Methoden – von datengetriebener Satz-zu-Satz-Übersetzung bis hin zu kontextsensitiver generativer Sprachmodellierung – in einem gemeinsamen Untersuchungsrahmen zu betrachten. Die vorangegangene Abbildung 3-5 zeigt den Zusammenhang der eingesetzten Systeme und verdeutlicht den Vergleichsbereich der beiden Dimensionen. Sie verdeutlicht zugleich den theoretischen Spannungsbogen zwischen spezialisierter, deterministischer Übersetzung und generativer, kontextadaptiver Sprachverarbeitung.
Im Nachfolgenden werden die einzelnen Systeme genauer beschrieben und hinsichtlich ihrer theoretischen Eignung kurz vorab bewertet, wobei der Fokus zunächst auf ihren technischen Grundlagen und architektonischen Besonderheiten liegt.

3.2.1 Beschreibung der eingesetzten Systeme

Der folgende Abschnitt bietet eine technische und methodische Charakterisierung der eingesetzten Übersetzungsmodelle. Im Zentrum steht die Analyse ihrer architektonischen Grundlagen, Trainingsparadigmen und funktionalen Unterschiede im Hinblick auf ihre Eignung für kontexttreue Übersetzung historischer Texte. Dabei werden die Systeme nicht primär deskriptiv, sondern im theoretischen Zusammenhang der neuronalen Sprachverarbeitung betrachtet – insbesondere mit Bezug auf die Prinzipien der Transformer-Architektur und des Aufmerksamkeits-Mechanismus.
Zur strukturellen Kohärenz werden zunächst die kommerziellen Cloud Systeme (DeepL und Gemini 2.5 Pro) beschreiben, gefolgt von offenen, lokalen Modellen (NLLB-200 und LLaMA-3).

NMT – **DeepL**

Das kommerzielle System DeepL stellt einen der technologisch ausgereiftesten Vertreter der neuronalen maschinellen Übersetzung dar. Es basiert auf der Transformer-Architektur, deren zentrales Merkmal die Self-Attention-Mechanik ist. Diese ermöglicht, dass jedes Wort (Token) einer Sequenz seine Beziehung zu allen anderen Elementen des Satzes modelliert, wodurch syntaktische und semantische Abhängigkeiten über die gesamte Eingabe hinweg erfasst werden. Damit überwindet der Transformer die lineare Beschränkung rekurrenter Netze (RNN, LSTM) (Bahdanau et al., 2014; Hochreiter & Schmidhuber, 1997). DeepL implementiert dieses Prinzip in einer mehrschichtigen Encoder-Decoder-Struktur mit proprietären Erweiterungen, deren Details zwar nicht veröffentlich sind, sich aber anhand empirischer Analysen rekonstruieren lassen (Koehn & Knowles, 2017; Läubli et al., 2018). Das System scheint ein Ensemble-Training auf mehrsprachigen Parallelkorpora und eine dynamische Domänenadaption zu nutzen, bei der spezifische Textdomänen nachtrainiert werden, um lexikalische Präzision und stilistische Kohärenz zu erhöhen. Ein typisches Merkmal moderner NMT-Implementierungen, wie sie auch DeepL verwendet, ist der Einsatz adaptiver $\mathrm{softmax}$-Schichten zur effizienteren Gewichtung seltener Vokabeln (Akhbardeh et al., 2021). Durch die Kombination von aufmerksamkeitsbasiertem Satzverständnis und kontinuierlicher Modelloptimierung erreicht DeepL eine hohe idiomatische Natürlichkeit und fluide Syntax – insbesondere in modernen Sprachpaaren. Allerdings bleibt das System weitgehend auf zeitgenössische Sprachen beschränkt. In Domänen mit nichtstandardisierter Grammatik, poetischer Struktur oder archaischen Wortformen – wie bei altgriechischen oder althebräischen Texten – wird es aufgrund mangelnder domänenspezifischer Trainingsdaten an seine Grenzen stoßen.
Für die anstehende Übersetzungsanalyse fungiert DeepL somit als Referenz für ein hochoptimiertes, aber datenabhängiges NMT, das den Stand kommerzieller Übersetzungstechnologien repräsentiert.

LLM – **Gemini 2.5 Pro**

Das von Google DeepMind entwickelte Gemini 2.5 Pro markiert den derzeitigen Stand großskaliger, generativer Sprachmodellierung. Es basiert auf einer Decoder-only-Transformer-Architektur,

bei der jedes Token autoregressiv auf Basis der vorhergehenden Token generiert wird. Im Gegensatz zu NMTs, die Übersetzung explizit als $P(t|s)$ modellieren (Zieltext | Quelltext), erzeugen LLMs und Gemini 2.5 Pro Text auf Grundlage des bedingten Wahrscheinlichkeitsmodells $P(w_t|w_1, \dots, w_{t-1})$, wodurch Übersetzung als Spezialfall generativer Sprachproduktion entsteht.
Gemini kombiniert diese Architektur mit einem Mixture-of-Expoerts-Ansatz (MoE), bei dem spezialisierte Teilnetzwerke abhängig vom Eingabekontext aktiviert werden (Fedus et al., 2021). Anders als herkömmliche dichte Transformatoren, bei denen alle Neuronen in jeder Schicht gleichzeitig aktiv sind, verteilt MoE die Berechnung auf einen Vielzahl von Expert-Subnetzen, die jeweils auf bestimmte sprachliche oder semantische Muster spezialisiert sind. Ein sogenannter Gating-Mechanismus entscheidet für jedes Eingabetoken, welches dieser Expertennetze aktiviert werden, sodass nur ein kleiner Teil des gesamten Netzwerks pro Berechnungsschritt tatsächlich genutzt wird (Lepikhin et al., 2020). Dieses Verfahren erhöht die Modellkapazität erheblich, ohne den Rechenaufwand proportional zu steigern, da nicht alle Parameter bei jeder Inferenz beteiligt sind. Gleichzeitig erlaubt die Spezialisierung einzelner Expertennetze eine feinere, semantische Differenzierung, etwa zwischen syntaktischer Struktur, lexikalischer Bedeutung und pragmatischer Kontextinformationen. Dadurch kann Gemini komplexe, kontextübergreifende Abhängigkeiten effizienter modellieren und Übersetzungen erzeugen, die nicht nur grammatikalisch, sondern auch semantisch konsistent sind (Fedus et al., 2021; Shazeer et al., 2017). Deshalb kann das Modell sowohl Rechenaufwand als auch Kontexttiefe adaptiv optimieren. Hinzu kommt eine multimodale Trainingspipeline, die Text-, Bild- und Audiodateien integriert, um semantische Relationen auf mehreren Ebenen zu modellieren. Empirische Vergleiche in WMT-24 und FLORES-200 zeigen, dass Gemini bei mehrsprachigen Korpora teilweise bessere Werte in BLUE COMET und chrF erreicht als ChatGPT-5 (Pang et al., 2024; Wu et al., 2025). Aus methodischer Sicht ist Gemini insbesondere deshalb relevant, weil es semantische und stilistische Kohärenz über sehr lange Textfenster aufrechterhalten kann. Damit steht es exemplarisch für den Paradigmenwechsel von spezialisierter Übersetzung zu kontextsensitiver Sprachmodellierung, bei der Bedeutung nicht mehr explizit übertragen, sondern implizit generiert wird.

NMT – **NLLB-200**

Das Forschungsmodell NLLB-200 (No Language Left Behind) in der Variante distilled-600M sowie 1.3B wurde von Meta AI mit dem Ziel entwickelt, die neuronale Übersetzung auf Sprachen mit geringer Datenverfügbarkeit auszuweiten. Im Gegensatz zu monolingual trainierten Systemen nutzt NLLB ein mehrsprachiges Multitask-Training auf einem gemeinsamen Encoder-Decoder-Netz, sodass es über Sprachgrenzen hinweg semantische Strukturen generalisieren kann. Die Architektur basiert ebenfalls – wie DeepL als kommerzielles System – auf dem Transformer-Prinzip, erweitert jedoch das Training um Verfahren wie Back-Translation (Sennrich et al., 2016), Knowledge Distillation (Hinton et al., 2015) und Adaptive Fine-Tuning (Howard & Ruder, 2018). Bei der Back-Translation werden Trainingsdaten durch Rückübersetzung erzeugt, um den Mangel an parallelen Texten in ressourcenschwachen Sprachen auszugleichen. Die Knowledge Distillation ermöglicht

es, Wissen eines größeren „Teacher“-Modells auf eine kompaktere Variante zu übertragen, wodurch das Modell auch bei reduzierter Parameterzahl semantische Zusammenhänge präzise abbilden kann. So lernt das kompaktere „Student“-Modell (…-600M) aus Referenzübersetzungen, die vom leistungsstärkeren „Teacher“-Modell (NLLB-200) vorab erzeugt wurden, ohne signifikant an Qualität zu verlieren. Durch Adaptive Fine-Tuning werden die Gewichte einzelner Modellschichten gezielt an spezifische Sprachdomänen angepasst, was die Generalisierungsfähigkeit erhöht und zugleich eine feinere Kontrolle über domänenspezifische Ausdrucksformen erlaubt. Damit vereint NLLB hohe Übersetzungsleistung mit geringerem Ressourcenbedarf.
Besonders hervorzuheben ist das Ziel der domänenübergreifenden Robustheit: Durch die simultane Modellierung vieler Sprachfamilien (darunter semantische und indogermanische) bildet NLLB tieferliegende syntaktische Regularitäten ab, die auch für historische Sprachformen relevant sind. Das Modell kann lokal ausgeführt werden und erlaubt so eine transparente Kontrolle.

LLM – **LLaMA-3**

Das LLaMA-3-Modell ist ein offenes LLM, das in seinen Varianten von acht bis 70 Milliarden Parametern verfügbar ist. Es basiert – genau wie Gemini – auf einer Decoder-only-Transformer-Architektur, verwendet jedoch vollständig offene Trainingsdaten und bietet damit mehr Transparenz und Reproduzierbarkeit. Die zum Einsatz kommende 8B-Variante nutzt Rotary Position Embeddings (RoPE) zur Positionskodierung sowie optimierte Flash-Attention-2-Mechanismen zur effizienteren Berechnung von Abhängigkeiten über lange Sequenzen. RoPE ersetzt klassische, feste Positionsvektoren durch rotationsbasierte Transformationen im komplexen Vektorraum, wodurch relative Wortpositionen präziser und kontextabhängiger erfasst werden können. Flash-Attention 2 optimiert die Speicher- und Rechenprozesse der Attention-Berechnung, indem relevante Token-Paare in den GPU-Speicher geladen und Zwischenergebnisse direkt zusammengefasst werden (Dao, 2023). LLaMA-3 verfolgt einen auf Effizienz und Anpassbarkeit ausgerichtetes Trainingsparadigma: Es kann über Instruction-Tuning und Low-Rank Adaption (LoRA)-Fine-Tuning domänenspezifisch nachtrainiert werden, ohne die Hauptarchitektur zu verändern. Beim Instruction-Tuning wird das Modell gezielt mit Anweisungs- und Aufgabenformaten weitertrainiert, um seine Fähigkeiten zu verbessern, natürlichsprachliche Eingaben zu interpretieren und drauf angemessen zu reagieren (Wei et al., 2021). LoRA hingegen verändert nicht die Hauptgewichte des Modells, sondern fügt zusätzliche, komplexe Matrizen hinzu, die während des Trainings aktualisiert werden (Hu et al., 2021). Dadurch lassen sich spezifische Aufgaben oder Domänen anpassen, ohne den gesamten Parameterraum neu zu optimieren.
Für die Untersuchung historischer Texte ist LLaMA-3 besonders relevant, weil es – im Gegensatz zu geschlossenen Modellen – eine präzise Steuerung der Hyperparameter (z. B. Temperature, Top-p-Sampling, Beam Width) und eine detaillierte Analyse der Attention-Gewichte erlaubt. Damit kann die semantische Strukturierung der Übersetzung direkt nachvollzogen werden.
Die lokale Ausführbarkeit auf handelsüblichen GPUs (z. B. einer Nvidia GeForce RTX 4080 Super) ermöglicht darüber hinaus eine vollständige Kontrolle der Reproduzierbarkeit.

3.2.2 Durchführung der Übersetzungen

Die praktische Durchführung der Übersetzung stellt einen wesentlichen Kern der vergleichenden Analyse dar, da erst durch die systematische Erzeugung paralleler Übersetzungen die spätere Modellbewertung, der qualitative Vergleich und die linguistische Analyse möglich werden. Die Übersetzung der ausgewählten hebräischen Perikopen erfolgt auf Basis einer einheitlichen, skriptgestützten Pipeline, die eine kontrollierte, reproduzierbare und modellübergreifende, vergleichbare Verarbeitung gewährleistet. Im Folgenden wird detailliert beschrieben, wie die Übersetzungen mit DeepL, Gemini und den lokal ausgeführten Modellen, NLLB sowie LLaMA-3, technisch umgesetzt wurden, welche Schwierigkeiten sich dabei ergaben und welche methodischen Entscheidungen aus diesen Erfahrungen abgeleitet wurden.

Gemeinsame Übersetzungspipeline

Die Grundlage aller Übersetzungen sind die zuvor erstellten .json-Dateien, in denen jede untersuchte Perikope versweise segmentiert wurde. Diese Struktur erlaubt es, die gesamte Übersetzungspipeline konsequent auf Vers-Ebene zu betrieben und damit eine hohe Genauigkeit in der späteren Auswertung zu gewährleisten. Jedes Skript liest die .json-Dateien automatisch ein und erweitert das bestehende Sprachobjekt um die entsprechenden Modellübersetzungen, ohne die bestehende Struktur zu verändern. Bereits vorliegende Übersetzungen werden erkannt und übersprungen, sodass der Prozess inkrementell fortgesetzt werden kann.

Zur Sicherstellung einer effizienten Verarbeitung wird ein Batching-Verfahren eingesetzt. Die Verse werden in Gruppen von acht bis zwölf Segmenten zusammengefasst und in einem einzelnen Anfragevorgang an das jeweilige Modell übermittelt. Dadurch wird die Anzahl der API-Aufrufe deutlich reduziert, was insbesondere bei Gemini aufgrund strenger Rate-Limits notwendig ist. Gleichzeitig bleibt die Zuordnung der Übersetzungsergebnisse zu den einzelnen Eingabeversen eindeutig nachvollziehbar.

Darüber hinaus schreiben sämtliche Skripte nach jeder erfolgreich erzeugten Übersetzung die aktualisierte Datei sofort zurück, sodass der Abbruch der Pipeline – etwa infolge von Netzfehlern, API-Rate-Limits oder GPU-bezogenen Problemen – keine bereits erzeugten Zwischenergebnisse zerstört. Diese redundante Speicherung stellt einen wichtigen Bestandteil der technischen Reproduzierbarkeit dar.

Durchführung mit DeepL

Die Integration von DeepL erfolgte über die offizielle SDK, ergänzt durch eine separate HTTP-Schnittstelle für die Bearbeitung der lateinischen Zielsprache. Dies ist notwendig, da DeepL Latein ausschließlich als Beta-Sprache bereitstellt und diese in der API nur über explizite HTTP-Aufrufe nutzbar ist. Die DeepL-Pipeline extrahiert zunächst die hebräischen Quelltexte – einmal vokalisiert und einmal unvokalisiert – und übermittelt diese abschließend in batched Requests an die API, wobei Deutsch und Englisch regulär über das SDK verarbeitet werden, während Latein über die HTTP-Schnittstelle angefragt wird.

Die Übersetzungseinstellungen wurden so gewählt, dass der Satzbau der hebräischen Versstruktur so weit wie möglich erhalten bleibt. Zu diesem Zweck wurde `preserve_formatting` aktiviert und `split_sentences` auf „`nonewlines`“ gesetzt. DeepL zeigt im Bereich moderner Gegenwartssprachen eine hohe Robustheit und Stabilität, weist jedoch bei den beta-basierten Sprachen deutliche Einschränkungen auf. Da für Latein weder Stilregister noch Glossare unterstützt werden und Altgriechisch über die API gänzlich nicht verfügbar ist, kann DeepL nur einen begrenzten Anteil der angestrebten Übersetzungen der benötigen Zielsprachen abdecken.
Diese Beschränkungen machen erstmals deutlich, dass DeepL als spezialisiertes NMT-System zwar sehr leistungsfähig im Bereich moderner Sprachen ist, jedoch im Bereich historischer Sprachen keine ausreichende Flexibilität besitzt. Insbesondere die fehlende Möglichkeit, stilistische oder philologische Instruktionen in irgendeiner Form zu übergeben, zeigt ein erstes Defizit

Durchführung mit Gemini 2.5 Pro

Die Gemini-Pipeline stellt den komplexesten Teil der Übersetzungsdurchführung dar, da dieses Modell eine generative Architektur nutzt und somit sowohl linguistische Feinsteuerung durch Prompting als auch die simultane Erzeugung mehrerer Zielsprachen ermöglicht. Die technische Umsetzung erfolgt über die „`google-genai`“-Bibliothek, die ein einheitliches Interface zur Kommunikation mit dem Modell bereitstellt. Ein zentraler Aspekt ist, dass Gemini in einem einzigen API-Aufruf vier unterschiedliche Zielsprachen liefern soll: Deutsch, Englisch, Latein und Altgriechisch. Diese Multi-Target-Strategie ist notwendig, um die Rate-Limits der API einzuhalten, da Gemini ansonsten nur zwei Anfragen pro Minute akzeptieren würde.
Die Prompt-Konfiguration ist präzise kalibriert. Sie verlangt für Deutsch einen sachlich-hochsprachlichen Stil, wie er in bibelwissenschaftlichen Kontexten üblich ist; Englisch einen nüchternen, akademischen Duktus; für Latein eine an die Vulgata orientierte Terminologie und Form; und für Altgriechisch eine Form, die sich an den Stil und die Morphologie der Septuaginta – dem Koine-Griechisch – annähert. Die klare Definition der gewünschten Register ist notwendig, da generative Modelle ansonsten dazu neigen, paraphrasierende und stilistisch inkonsistente Formulierungen zu erzeugen.
Ein wesentlicher Bestandteil der Pipeline ist die umfassende, automatische Fehlererkennung und -behandlung während der Kommunikation mit cloudbasierten Modellen, insbesondere bei der Nutzung des Gemini 2.5 Pro Modells. Da generative cloudbasierten LLMs häufig mit strengen Nutzungsbegrenzungen, zeitlichen Aufrufbeschränkungen und variablen Antwortformanten arbeiten, müssen Skripte eine Reihe spezialisierter Mechanismen implementieren, die den Ablauf stabilisieren und eine vollständige reproduzierbare Verarbeitung gewährleisten. Die häufigste Fehlerquelle bei Gemini sind Rate-Limit-Überschreitungen. Die API signalisiert solche Ereignisse über den gRPC-Fehlercode `ResourceExhausted`, der im Python-Client typischerweise als Exception mit dem HTTP-Statuscode 429 erscheint. Dieser Fehler entsteht, wenn das Modell innerhalb eines bestimmten Zeitfensters mehr Anfragen erhält, als das jeweilige Nutzungskontigent zulässt. Da

Übersetzungspipeline mehrere Zielsprachen gleichzeitig erzeugt und für jeden Vers einen vollständigen Prompt überträgt, ist die Wahrscheinlichkeit einer Limitüberschreitung signifikant erhöht. Um einen Abbruch der Verarbeitung zu vermeiden, implementiert das Skript daher eine robust gestaltete Retry-Logik: Tritt ein Fehler der Klasse `ResourceExhausted` auf, wird dieser abgefangen, das Skript legt eine vordefinierte Wartezeit von 70 Sekunden ein und wiederholt anschließend denselben API-Aufruf. Die Wahl dieser Wartezeit basiert auf empirischen Beobachtungen; sie liegt oberhalb der typischen Abkühlzeit des Modells und garantiert damit eine erfolgreiche Abfrage beim erneuten Versuch. Neben der expliziten Fehlerbehandlung ist auch die Steuerung der Aufrufintervalle ein kritischer Bestandteil des technischen Designs. Gemini erlaubt im Standard nur etwa zwei Anfragen pro Minute. Da ein einzelner Request vier Zielsprachen gleichzeitig erzeugt, muss die Pipeline sicherstellen, dass die Effektivlast der API dauerhaft unter der erlaubten Schwelle bleibt. Zu diesem Zweck implementiert das Skript eine Zwangspause von 35 Sekunden nach jedem erfolgreich verarbeiten Vers, unabhängig davon, ob zuvor ein Fehler aufgetreten ist oder nicht. Diese künstliche Verzögerung verhindert eine schleichende Erhöhung der Anfragerate über längere Batch-Prozesse hinweg und stellt sicher, dass die Pipeline auch bei mehreren aufeinanderfolgenden Versen stabil bleibt. Das Zusammenspiel aus expliziten Pausen und automatischer Fehlerkorrektur macht die Ausführung deterministisch und verhindert, dass Langläuferprozesse unvollständig bleiben oder in nicht reproduzierbare Zustände geraten. Ein weiterer technischer Schwerpunkt besteht in der Validierung des .json-Ausgabeformats. Da generative Antworten nicht deterministisch, sondern probabilistisch erzeugt werden, können selbst bei präziser Prompt-Definition Formatierungsabweichungen auftreten, etwa durch eingeschlossene Markdown-Blöcke, zusätzliche Textfragmente oder unvollständiger .json-Strukturen. Das Skript bereinigt zunächst alle bekannte Artefakte – typischerweise die ungewollte Einbettung in .json-Blöcke – und versucht anschließend, die Antworten in ein Python-Objekt zu parsen. Schlägt dieser Parsing-Vorgang fehl, wird der Prompt erneut an das Modell übergeben. Da nur korrekt formatiertes .json in die strukturierte Versdatei übernommen wird, kann die Pipeline garantieren, dass die nachfolgenden Auswertungsschritte auf formal korrekten Daten basieren.

Durch diese Kombination aus Fehlerklassifikation, Retry-Mechanismus, zeitlicher Drosselung und formaler Validierung entsteht ein hochgradig robustes und fehlertolerantes Übersetzungssystem, das auch unter den technischen Randbedingungen eines starken Cloud-Modells zuverlässig vollständig Daten liefert. Die lokale Pipeline der NMT-Systeme benötigt diese Mechanismen nicht in gleichem Umfang, da dort weder API-Limits noch probabilistische Antwortformate bestehen. Der besondere technische Aufwand für Gemini ist damit nicht zufällig, sondern eine direkte Folge seiner Modellklasse und Architektur, die eine höhere Flexibilität bietet, aber dessen Nutzung zugleich die Notwendigkeit präziser Kontrollmechanismen erforderlich macht.

So zeigen die ersten Ergebnisse, dass Gemini als cloudbasiertes LLM in der Lage ist, stilistisch fein gesteuerte Übersetzungen zu erzeugen, die für die ausgewählten Zielsprachen verfügbar sind.

Durchführung mit NLLB-200

Die Nutzung der NLLB-Modelle erfordert eine differenzierte Betrachtung des zugrunde liegenden Seq2Seq-Ansatzes, der Sprachkodierung, der Tokenisierung sowie des Decoding-Verhaltens. NLLB folgt einer strikt strukturierten Architektur, die das Übersetzungsverhalten deterministisch gestaltet und daher im methodischen Vergleich einen klar abgegrenzten Systemtypen repräsentiert.

Die folgende Beschreibung erläutert die internen Abläufe im Modell und zeigt, weshalb NLLB für historische Sprachen sowohl funktionale Stärken als auch inhärente Einschränkungen aufweist.
NLLB-200 – sowohl -1.3B als auch -600M – basiert auf einem klassischen Encoder-Decoder-Modell, das im Kern der Transformer-Architektur folgt, sich jedoch in zwei entscheidenden Punkten von generativen Decoder-only-Systemen unterscheidet:

[1] Der Encoder verarbeitet die Quellsprache in ihrer Gesamtheit und erzeugt eine hochdimensionale Repräsentation aller Token.

[2] Der Decoder generiert die Zielsprache sequenziell, wobei jede Tokenvorhersage strikt an die vom Modell festgestellte Sprachumgebung gebunden ist.
(forced language conditioning)

Dieser Aufbauschritt führt dazu, dass das Modell die Übersetzung nicht in einem freien Fluss erzeugt, sondern eine explizite strukturierte linguistische Abbildung vollzieht. Die interne Aufmerksamkeit (Self- und Cross-Attention) unterscheidet sich insofern von LLMs, als dass der Decoderzugriff auf den Encoderausgang systematisch reguliert wird und kein globaler Kontext über sämtliche Übersetzungsaufgaben hinweg entsteht. NLLB kann daher keine stilistischen Muster über längere Textfenster erlernen, sondern behandelt jeden Vers als insolierte Eingabeeinheit.
Die technische Grundlage jeder Übersetzung bildet bei NLLB die Wahl der Sprachcodes über den `NllbTokenizerFast`. Dieser Tokenizer stellt sicher, dass:

[1] die Quellsprache korrekt normalisiert wird,

[2] die Zielsprachen eindeutig identifiziert werden,

[3] sämtliche Eingaben in das vom Modell erwartete Schema überführt werden.

Die Sprachcodes wie `heb_Hebr`, `deu_Latn` und `eng_Latn` sind nicht bloße Metadaten, sondern funktionale Steuerbefehle an das Modell. Sie sorgen dafür, dass das gesamte Vokabular des Decoders, die tokeninterne Normalisierung und die Vektorraumbereiche korrekt initialisiert werden. NLLB unterscheidet streng zwischen Skript (z. B. *Hebr* für Hebräisch Alphabet) und Sprachfamilien. Gerade bei historischen Sprachen ist diese Unterscheidung relevant, da beispielweise Altgriechisch kein dediziertes Skriptmodell erhält und somit nicht adressierbar ist. Mit dem Setzen von z. B. `tokenizer.src_lang = "heb_Hebr"` wird der Tokenizer angewiesen, die Eingabe nach dem hebräischen Unicode-Normalisierungsschema zu segmentieren, das vollständig auf Konsonanten-

ebene arbeitet und diakritische Zeichen entweder verwirft oder als separate Tokens behandelt. Die Tokenisierung ist für NMT optimiert, nicht jedoch für philologische Feinheiten – ein methodisch bedeutender Unterschied zu LLMs.
Ein weiterer zentraler Mechanismus des Modells ist die Nutzung eines forced BOS (Beginning-of-Sentence) Token. Dieser Token entscheidet technisch darüber:

[1] in welchem Sprachraum der Decoder startet,

[2] welches Zielvokabular aktiviert wird,

[3] welche Weight-Matrizen im Decoder zur Anwendung kommen.

Da NLLB ein mehrsprachiges Modell mit über 200 Sprachen ist, würde der Decoder ohne diesen Zwangtoken theoretisch frei beginnen – was bei mehrsprachigen Modellen regelmäßig zu Sprachmischungen (*engl. language bleed*) führen würde. Der Skript nutzt deshalb die Anweisung:

```
forced_bos_token_id = tokenizer.lang_code_to_id[tgt_lang_code]
```

So kann sichergestellt werden, dass der erste generierte Token in der gewünschten Zielsprache verortet ist. Dieser Schritt fixiert das Sprachmodell auf ein bestimmtes Vokabularset und eine spezifische Wahrscheinlichkeitstopologie, sodass die darauffolgende Sequenz deterministisch bleibt. Für historische Sprachen wird hier ein methodischer Zusammenhang sichtbar. Da Altgriechisch nicht im Vokabular enthalten ist, existiert kein gültiger BOS-Token für GRC. Auch für das Lateinische ist NLLB nicht trainiert und listet keinen validen Token für LA auf. Das Modell kann somit keinen regulären Decoder-Prozess für lateinische und koine-griechische Ausgaben initiieren. Die Leerstellen „NLLB_LA“ und „NLLB_GRC“ sind daher nicht nur ein semantisches Ergebnis, sondern ein architekturbedingter Grenzpunkt des Modells.
Das Decoding erfolgt bei NLLB typischerweise über beam search, wodurch das Modell mehrere mögliche Sequenzen gleichzeitig evaluiert und jene auswählt, die nach dem internen Sprachmodell die Wahrscheinlichste darstellt. Im Unterschied zu generativen LLMs wird kein Sampling verwendet, sodass die Ergebnisse vollständig deterministisch sind. Identischer Input führt damit zu identischem Output, unabhängig vom Kontext oder vorherigen Übersetzungsvorgängen. Dieser Prozess hat damit zwei relevante Konsequenzen:

[1] Stilistische Variationen sind nicht möglich, da der Decoder stets dieselbe maximale Wahrscheinlichkeitssequenz erzeugt.

[2] Semantische Feinheiten, Ambivalenzen oder poetische Strukturen werden häufig in die durchschnittlichste, glatteste Lösung überführt.

Gerade die hebräische Poesie ist davon betroffen: Komplex-parallel strukturierte Passagen werden durch das probabilistische Ranking zu linearen, inhaltlich eindeutiger formulierten Aussagen reduziert. Der mögliche Verlust kontextueller Ambiguität ist an dieser Stelle eine erste Bewertung der systemischen Eigenschaften des Modells.

Durchführung mit LLaMA-3

Das lokal ausgeführte Modell Meta LLaMA-3 stellt im Rahmen der Untersuchung einen Vertreter der modernen Decoder-only-Transformator-Architekturen dar und unterscheidet sich damit grundsätzlich von den Encoder-Decoder-basierten NMT-Systemen sowie den cloudbasierten, hochgradig abstrahierten Modellen. Die technische Ausgestaltung des LLaMA-3-Workflows beruht auf einer Kombination aus lokalem GPU-Inference-Betrieb, einem systematisch aufgebauten Chat-Prompting-Schema sowie restriktiven Sampling-Parametern, die eine möglichst literal orientierte Übersetzung erzeugen sollen. Die folgende Darstellung beschreibt die zentralen Funktionsweisen und architektonischen Besonderheiten im Detail und erläutert die Implikationen dieser Struktur für die Übersetzung historischer Texte.

Die LLaMA-3-Pipeline lädt das Modell über die `transformers`-Bibliothek mittels `AutoModelForCausalLM` und nutzt `device_map="auto"` sowie `torch_dtype=torch.bfloat16`, wodurch das Modell optimal an die Ressourcen der eingesetzten GPU – eine NVIDIA RTX 4080 Super – angepasst wird. Diese Konfiguration minimiert den Verbrauch von VRAM der Grafikkarte und erlaubt dennoch schnelle Inferenz mit gleichzeitiger Erhaltung der numerischen Stabilität. Die lokale Ausführung stellt sicher, dass das Modell nicht durch Netzwerklatenzen, API-Limits und Serverlast variabel beeinflusst wird, sodass der gesamte Inferenzprozess vollständig deterministisch bezüglich Hardware und Parametrierung ist. Die Modellinitialisierung umfasst das Laden des Tokenizers (`AutoTokenizer.from_pretrained`), der die Tokenisierung nach dem Byte-Pair-Encoding-ähnlichen Schema der LLaMA-Reihe durchführt. Der Tokenizer definiert dabei zugleich das Chat-Template, das zur formalen Strukturierung der System- und Nutzerinstruktionen genutzt wird. Diese klar bestimmte Input-Struktur ist notwendig, um das Modell zuverlässig in einem instruktionalen Modus zu versetzen, da Instruct-Modelle ihre Verhaltenssteuerung ausschließlich über Prompt-Kontexte erhalten.

Ein zentrales Unterscheidungsmerkmal zu NMT-Systemen ist die Nutzung eines Chat-Templates, das explizit zwischen Systemrolle und Nutzerrolle differenziert. Das Skript erzeugt die Eingabe, indem es ein Array aus zwei strukturierten Chat-Elementen aufbaut und anschließend das vollständige Prompt-Objekt über das Chat-Template des Tokenizers zu einer modellkompatiblen Eingabesequenz zusammenführt. Dabei übernimmt der System-Prompt die Funktion eines sprachlichen und stilistischen Rahmens, der Folgendes näher beschreibt:

[1] die gewünschte Zielvarietät (z. B. LXX-ähnlich bei Altgriechisch),

[2] die Literalität (keine Paraphrasen),

[3] die Registeranforderung (hochsprachlich, wissenschaftlich),

[4] die Strukturvorgaben (keine Meta-Kommentare, keine Erklärungen).

Der User-Prompt besteht ausschließlich aus dem hebräischen Vers. Die Trennung der beiden Rollen ist technisch besonders relevant, da der Decoder-only-Mechanismus alle Tokens – einschließ-

lich der Instruktionen – sequenziell verarbeitet und die Rolle nur über explizite Markierungen unterscheiden kann. Das Modell generiert die Ausgabe anschließend als Fortsetzung dieses Prompt-Kontexts. Diese Architektur unterscheidet sich fundamental von NMT-Systemen: Während z. B. NLLB vorab definierte Sprachräume nutzt, entscheidet LLaMA aufgrund der Promptsteuerung dynamisch, welchen Stil und welche Varietät es erzeugt. Das Prompting ist somit keine Ergänzung, sondern ein essenzieller Bestandteil des Modus der Übersetzung.
Die Tokenisierung in LLaMA-3 folgt einem BPE-ähnlichen (Byte-Pair-Encoding) Schema im Unicode Raum und verarbeitet auch historische Alphabete – wie etwa Hebräisch und Griechisch – über universelle Subword-Einheiten. Dies führt zu technisch relevanten Konsequenzen:

[1] Die Modellarchitektur unterscheidet nicht zwischen Schriftsystemen, sondern bildet semantische Zusammenhänge ausschließlich über statisches Auftreten von Subword-Fragmenten im Trainingskorpus ab.

[2] Semantische Nähe entsteht nicht über linguistische Merkmale, sondern über Verteilungsmuster der Tokenfragmente im Pretraining.

Dies bedeutet, dass das Modell Hebräisch, Latein und Altgriechisch nicht als kohärente Sprachsysteme erfasst, sondern als Cluster von Subword-Einheiten, die es über Kontextwahrscheinlichkeiten interpretiert. Die Qualität der Übersetzung hängt daher entscheidend davon ab, wie häufig und in welchen Kontexten die entsprechende Tokenstrukturen im Trainingsmaterial vorkamen.
Da LLaMA ein generatives Modell ist, basiert seine Ausgabe nicht auf maximaler Wahrscheinlichkeit einer festen Sequenz, sondern auf sampling-basierten Verfahren. Das Skript nutzt stark restriktive Samplingparameter: `temperature = 0.1`; `top_p = 0.9`; `do_sample = True`. Eine niedrige Temperatur reduziert die Varianz der Aktivierungsverteilung, sodass das Modell die wahrscheinlichste Fortsetzung bevorzugt. Der Parameter `top_p` begrenzt die Auswahl auf den kleinsten Wortschatzbereich, der mindestens 90 % der kombinierten Wahrscheinlichkeitsmasse enthält. Diese Parameterwahl ist wichtig, da höhere Temperaturen zu semantisch abweichenden, stilistisch inkonsistenten oder interpretativen paraphrasierenden Texten führen würden – ein Risiko, das bei historischen Texten besonders stark ausgeprägt ist. Durch diese restriktive Parametrisierung nähert sich LLaMA funktional einem NMT-System an, bleibt jedoch im Kern ein generatives Modell. Die Aussagen sind somit weniger deterministisch stabil als NLLB, jedoch deutlich steuerbarer als DeepL.
LLaMA nutzt zwei wichtige End-Token-Mechanismen. Zum einen das Standard-EOS-Token (`eos_token_id`) und zum zweiten das Instruct-Template-spezifische Abschluss-Token (`<|eot_id|>`). Die Pipeline definiert beide als mögliche Terminatoren. Diese Token sind notwendig, da generative Modelle autoregressiv arbeiten. Sie erzeugen Wort für Wort immer weiter – so lange, bis ein Stoppkriterium erreicht wird. Ohne ein solches Signal könnten Modelle den Text endlos fortsetzen, oder unter anderem zusätzliche Kommentare anhängen. Dieser doppelte Abschlussmechanismus verhindert, dass das Modell Metadiskurse anhängt, weitere Chat-Rollen

übernimmt, Erklärungen generiert oder alternative Übersetzungsvorschläge produziert. Gerade bei generativen Modellen ist diese Form der Kontrolle notwendig, um eine konsistente, formatierte und verwertbare Ausgabe zu gewährleisten. Ohne solche Terminatoren würden viele Übersetzungen in Absätze übergehen, die den Übersetzungsprozess kommentieren oder reflektieren.
Das Skript implementiert System-Prompts (Deutsch, Englisch, Latein und Altgriechisch). Diese unterscheiden sich nicht nur im Zielregister, sondern auch in der Modellinstruktion:

[1]	Deutsch	wissenschaftlicher Duktus, semitisch-syntaktische Nähe
[2]	Englisch	standardisierter und akademischer Sprachgebrauch
[3]	Latein	Fokus auf kirchlichem Latein und Vulgata-Kompatibilität
[4]	Altgriechisch	expliziter Ausschluss moderner griechischer Formen

Die voneinander abgegrenzten Instruktionen sind notwendig, da LLaMA keine eingebauten Mechanismen besitzt, die Varietäten eindeutig trennen. Ohne klare Instruktion würde das Modell Latein mit Neulatein (Verwendung im Renaissance-Humanismus bis in die Gegenwart) oder Altgriechisch in modernen Strukturen formulieren. Die Präzision dieser Prompts ist damit nicht nur stilistisch, sondern technisch funktional erforderlich.
Generative Modelle neigen systematisch dazu, sogenannte *role-prefix artifacts*, bei den Ausgaben mit rollenbasierten Präfixen zu beginnen – z. B. „German: …“. Das Skript erkennt diese Artefakte und entfernt sie über heuristische Regeln. Diese Bereinigung ist notwendig, da LLaMA als generatives Modell keine inhärenten Formatgarantien gibt. Für die spätere Vergleichbarkeit der Übersetzungen in der Triangulation ist die standardisierte Normalisierung daher ein kritischer Schritt.

Erste technische Evaluation der vier Übersetzungsmodelle

Die praktische Durchführung der Übersetzungen offenbart deutliche strukturelle Unterschiede zwischen den vier eingesetzten Systemen. Diese erste Differenzen sind dabei nicht das Resultat der beobachteten Übersetzungen, sondern entstehen vielmehr aus ihrer Modellarchitektur, den technischen Rahmenbedingungen, der Steuerbarkeit, der Stabilität sowie der Operationalisierbarkeit innerhalb einer skriptgestützten Pipeline, wobei auf Qualitätsmerkmale der Übersetzungen zu einem späteren Zeitpunkt eingegangen wird. In der Summe zeichnen sich nicht zwei, sondern viel mehr vier Modelltypen ab, die sich in ihrer Funktionsweise grundlegend voneinander unterscheiden und deren Verhalten während der Durchführung charakteristische Muster erkennen lässt.
DeepL zeigt sich im praktischen Betrieb als das formal zuverlässigste und stabilste System. API-Anfragen werden deterministisch verarbeitet, liefern reproduzierbare Ergebnisse und erfordern kaum Eingriffe in den Ablauf. Insbesondere bei Deutsch und Englisch ist die API vollständig normiert, reagiert vorhersehbar und erzeugt konsistente Ausgaben. Die Pipeline kann Textmengen in klar definierten Batches verarbeiten, ohne in zeitliche oder technische Limitierungen zu geraten. Die zentralen Probleme betrifft jedoch die Funktionsbereite des Systems. Historische Sprachen wie Altgriechisch sind über die API nicht verfügbar, während Latein vorerst nur in einer einge-

schränkten Beta-Version angesprochen werden kann, die aber wesentliche Parameter wie `formality` oder `glossaries` nicht unterstützt. Dabei besitzt DeepL die zwei geringste Sprachabdeckung unter allen gezeigten Modellen. Zudem besteht keine Möglichkeit zur stilistischen Feinsteuerung oder Promptkontrolle, da DeepL keine Anweisungen in Textform auswertet. Das Modell arbeitet strikt innerhalb eines deterministischen Rahmens, der zwar zuverlässig ist, aber keinen philologischen oder historischen Stileinfluss erlaubt. In einer Gesamtbetrachtung erweist sich DeepL als robustes, einfaches Basissystem mit geringer Flexibilität.

Gemini stellt unter den vier Systemen die technisch anspruchsvollste Komponente dar. Das Modell erlaubt die gleichzeitige Erzeugung mehrerer Zielsprachen in einem einzigen Request, was die Pipeline funktional vereinfacht und zugleich eine präzise stilistische Steuerung ermöglicht. Die Fähigkeit, stil- und registerbezogene oder philologische Instruktionen direkt in einem Prompt zu übergeben, unterscheidet Gemini als generatives LLM grundlegend von den NMT-Systemen. Dadurch kann das Modell streng instruiert werden, beispielweise Vulgata-nahes Latein oder LXX-ähnliches Altgriechisch zu erzeugen. Die technische Durchführung ist jedoch mit Herausforderungen verbunden. Durch die inhärente Komplexität generativer LLMs unterliegen APIs – auch die von Gemini – häufig strikter Rate-Limits, was zu regelmäßigen Fehlern und damit verbunden zu Wartezeiten führt. Die notwendige Stabilität muss durch den Einsatz von Throttling-Mechanismen, einer Retry-Logik und einer automatisierten Abkühlphase gewährleistet werden. Zusätzlich müssen alle Antworten durch .json-Parsing validiert werden, um auch bei generativen Modellen ein garantiertes Format zu liefern. Diese formalen Anforderungen verlangsamen den Gesamtprozess und machen die Pipeline anfälliger für Unterbrechungen und anwendungsbezogene Wartungen. Summiert lässt sich evaluieren, dass diese technischen Überprüfungen Folgen der gebotenen Flexibilität innerhalb des Modells sind. In der praktischen Verwendung zeigt es eine hohe Ausdrucksfähigkeit und Steuerbarkeit, die allerdings das Benutzen und den Einsatz im Entwicklungsprozess komplexiert.

Die NLLB-Pipeline zeichnet sich durch eine sehr hohe technische Stabilität und eine vollständig deterministische Arbeitsweise aus. Als klassisches Seq2Seq-Modell benötigt NLLB keine Prompting-Mechanismen. Sämtliche Steuerung erfolgt über Sprachcodes und forced BOS-Token, wodurch jede Zielsprachsequenz sauber initialisiert wird. Die lokale Ausführung vermeidet externe Abhängigkeiten, API-Fehler oder Rate-Limits, sodass die Verarbeitung auch größerer Textmengen problemlos ohne Unterbrechung erfolgen kann. Die Grenzen dieser Stabilität liegen in der Struktur des Modells selbst. Da NLLB ausschließlich moderne Sprachen abdeckt, ist Altgriechisch nicht adressierbar, und auch Latein wird nur als moderne Varietät ausgegeben. Die fehlende Möglichkeit zur stilistischen Steuerung führt dazu, dass poetische Strukturen, historische Varietäten oder philologische Besonderheiten nicht reproduzierbar steuerbar sind. Durch die deterministische Decoding-Architektur entstehen linearisierte, geglättete Zielstrukturen. Gleichwohl ist die Konsistenz der Pipeline ein Vorteil für syntaktische Vergleiche und für technische Reproduzierbarkeit. NLLB bietet somit die höchste Betriebssicherheit, jedoch die geringste stilistische Flexibilität.

LLaMA-3 repräsentiert die flexibelste lokal ausführbare Modellklasse im Vergleich. Das Modell akzeptiert komplexe Systemprompts, wodurch Registerwahl, Literalität oder historische Stilnähe präzise gesteuert werden kann. Die lokale Ausführung auf einer performanten GPU ermöglicht vollständige Kontrolle über Sampling-Parameter, Chat-Templates und Terminierungsmechanismen, wodurch das Verhalten des Modells weitgehend regulierbar bleibt. Zudem unterstützt LLaMA alle vier benötigten Zielsprachen, einschließlich eines Koine-ähnlichen Griechisch – eine Fähigkeit, die beide einbezogenen NMT-Systeme, DeepL und NLLB, nicht besitzen. Gleichzeitig zeigt sich im praktischen Einsatz eine deutliche Sensitivität gegenüber kleinsten Promptänderungen. Da der gesamte Kontext – System- und Userprompt – in die Generierung einfließt, reagiert das Modell stärker variabel als NMT-Systeme. Die Notwendigkeit restriktiver Samplingparameter (`temperature = 0.1`) zeigt, dass Varianz ansonsten schnell zu stilistischen oder semantischen Drifterscheinungen führt. Zudem muss die Pipeline, wie auch schon bei Gemini, typische LLM-Artefakte entfernen, da generative Modelle keine inhärente Formatdisziplin besitzen. LLaMA bildet damit einen Mittelweg zwischen Flexibilität und technischer Kontrolle, erfordert jedoch eine sorgfältige Gestaltung der Prompts, um konstante Ergebnisse zu erzielen.

Die nachfolgende Tabelle zeigt die technischen Eigenschaften im Modellvergleich.

Tabelle 3-2 *Modelleigenschaften im Überblick (eigene Darstellung)*

Modell	**Gattung**	**Stabilität**	**Flexibilität**	**Stilsteuerung**	**Sprachabdeckung**
DeepL	NMT	Sehr hoch	Niedrig	Keine	Begrenzt *(ohne GRC)*
NLLB-200		Sehr hoch	Niedrig	Keine	Nur moderne Sprachen
Gemini 2.5 Pro	LLM	Mittel	Sehr hoch	Umfassend	Voll *(inkl. GRC)*
LLaMA-3		Mittel	Hoch	Präzise	Voll *(inkl. GRC)*

Insgesamt zeigt der technische Vergleich der vier Systeme, dass jedes Modell eine deutlich unterscheidbare operative Charakteristik besitzt, die bereits auf der Ebene der Durchführung maßgeblich prägt, wie Übersetzungen erzeugt, kontrolliert und weiterverarbeitet werden können. Die Kombination dieser Modelle liefert ein breits methodisches Spektrum, dessen technische Diversität der späteren Analyse eine differenzierte Grundlage bietet. Der Vergleich der Systemarchitekturen zeigt zugleich, dass die Übersetzungsleistung nicht isoliert betrachtet werden kann, sondern im engen Zusammenhang mit den jeweiligen technischen Rahmenbedingungen, Modellgrenzen und Stärken der eingesetzten Verfahren steht.

3.3 Evaluationsmethoden

Die im Rahmen der Übersetzungen gewonnenen Texte bilden die Grundlage für eine detaillierte Auswertung, die über eine rein technische Betrachtung hinausgeht. Da die Modelle nicht nur formale Strukturen erzeugen, sondern zugleich Bedeutungen setzen und Interpretationsspielräume unterschiedlich nutzen, ist eine systematische Evaluation notwendig, die sowohl sprachliche als auch inhaltliche Aspekte berücksichtigt. Übersetzungen biblischer Texte bewegen sich grundsätzlich im Bereich zwischen sprachlicher Form, historischer Einbettung und theologischer Auslegung. Entsprechend muss auch die Analyse verschiedene Ebene miteinander verbinden, um ein differenziertes Bild der Modelle zu erhalten.

Die Evaluationsmethoden sind daher so angelegt, dass sie den Übersetzungen in ihrer Mehrschichtigkeit gerecht werden. Einerseits wird berücksichtigt, wie die Modelle mit Strukturen historischer Sprachen umgehen, welche formalen Muster sich erkennen lassen und wo typische Eigenschaften neuronaler Übersetzungssysteme sichtbar werden. Andererseits wird untersucht, welche Entscheidungen die Modelle auf der Bedeutungsebene treffen, insbesondere an den Stellen, die für das theologische Verständnis des Textes eine zentrale Rolle spielen. Zusätzlich eröffnet ein Abgleich mit etablierten Übersetzungen einen traditionsgeschichtlichen Rahmen, innerhalb dessen die maschinellen Ausgaben verortet und besser eingeschätzt werden können.

Auf diese Weise entsteht ein kohärenter methodischer Zugang, der technische, semantische und historische Beobachtungen miteinander verbindet, ohne die Bereiche künstlich voneinander zu trennen. Die folgenden Unterkapitel entfalten diese Perspektiven in ihrer jeweiligen Zielrichtung, greifen aber aufeinander zurück, wo dies für das Gesamtverständnis notwendig ist. Die Evaluationsmethoden dienen damit nicht nur der Überprüfung der Übersetzungsqualität, sondern auch der Einordnung der Modelle im Spannungsfeld aus formaler Präzision, interpretativer Offenheit und der Geschichte bereits etablierter Lesarten.

3.3.1 Quantitative, technische Analyse

Die quantitative, technische Analyse dient der systematischen Untersuchung der maschinellen Übersetzung anhand formaler und strukturbezogener Eigenschaften, die unabhängig von theologischen oder interpretativen Fragestellungen erfasst werden können. Sie ergänzt die in Kapitel 3.3.2 dargestellte qualitative und hermeneutische Auswertung, indem sie die Verhaltensweisen der Modelle auf einer rein technischen Ebene beschreibt. Die Auswahl der Methoden orientiert sich insbesondere an den strukturellen Merkmalen historischer Sprachen wie dem Biblisch-Hebräischen und dem Koine-Griechischen, deren morphologische und syntaktische Eigenschaften moderne Übersetzungssysteme vor besondere Herausforderungen stellen. Dazu gehören das trilaterale Wurzelsystem, die häufig fehlende Vokalisierung, die ausgedehnte Flexionsmorphologie des Griechischen sowie die freie Wortstellung beider Sprachen. Die technische Analyse ist darauf ausgerichtet, diese Strukturen sichtbar zu machen und daraus Rückschlüsse auf die Funktionsweise und Grenzen der Modelle zu erhalten.

Ein erster Schwerpunkt liegt auf der Untersuchung der Tokenisierung. Neuronale Übersetzungssysteme segmentieren ihre Eingaben nicht in vollständigen Wörtern, sondern in Subword-Einheiten, die durch Verfahren wie BPE oder SentencePiece erzeugt werden (Kudo & Richardson, 2018; Sennrich et al., 2016). Diese Algorithmen lernen wiederkehrende Zeichenfolgen, indem sie die Häufigkeiten von Zeichenpaaren xy maximieren, formal ausgedrückt durch

$$\arg\max_{x,y \in V} f(xy)$$

wobei V den jeweils aktuellen Tokenvorrat und $f(xy)$ die Frequenz des Zeichenpaares bezeichnet. Für historische Sprachen kann diese Subword-Segmentierung erhebliche Auswirkungen auf die Übersetzungsqualität haben. Hebräische Wörter, die auf Konsonantenwurzeln basieren und deren vokalische Muster im unvokalisierten Text nicht sichtbar sind, können so häufig übermäßig fragmentiert werden; altgriechische Flexionsformen neigen dazu, aufgrund ihrer Vielfalt eine Vielzahl seltener Subtoken-Sequenzen zu erzeugen. Um diese Phänomene zu erfassen, werden die Ausgaben der Modelle tokenisiert und die Token Fragmentation Rate berechnet, definiert als

$$TFR = \frac{Anzahl\ der\ Subtoken}{Anzahl\ der\ Wörter}$$

Eine erhöhte TFR weist auf Unsicherheiten in der internen Repräsentation der Modelle hin und ist insbesondere bei hebräischen Konsonantenclustern und griechischen Partizipialformen zu erwarten. Zusätzlich werden die modellinternen Tokenisierungen mit externen linguistischen Tokenizern verglichen, um zu beurteilen, ob die Modelle Wurzeln, Stämme und Flexionsmuster kohärent erfassen oder systematisch aufbrechen.

Ein zweites zentrales Verfahren bildet die Alignment-Analyse. Sie untersucht, inwieweit die Modelle Quell- und Zielwörter einander zuordnen und wie sie syntaktische Relationen rekonstruieren. Dies ist bei historischen Sprachen besonders relevant, da das Hebräische häufig kopulalose Nominalsätze bildet, die im Deutschen normalerweise durch ein explizites Kopulaverb wiedergegeben wer-

den müssen und altgriechische Partizipialkonstruktionen häufig die im Deutschen verwendeten Nebensätze ersetzen. Zur Analyse wird sowohl ein statisches Verfahren wie `fast_align` verwendet (Dyer et al., 2013), das auf dem IBM Model 2 basiert und positionsbasierte Wahrscheinlichkeiten schätzt, als auch ein embeddingbasiertes Verfahren wie `SimAlign`, das kontextuelle Vektorrepräsentation nutzt. Die statistische Zuordnung orientiert sich an einer Wahrscheinlichkeitsfunktion $P(a_j = i)$, die die Distanz der Tokenpositionen modelliert, während die embeddingbasierte Variante auf der Kosinusähnlichkeit

$$\cos(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$$

zwischen Quell- und Zielvektoren beruht. Durch die Kombination beider Ansätze lassen sich sowohl formale als auch vektorraumbasierte Nahbeziehungen erfassen. Auf diese Weise können Einfügungen, Auslassungen oder strukturelle Transformationen sichtbar gemacht werden, ohne dass bereits eine theologische Bedeutungsebene berührt wird.

Ergänzend zur Tokenisierung und zum Alignment wird die Strukturtreue der Übersetzungen analysiert. Dazu gehören Satzlängenverhältnisse, Wiedererkennbarkeit syntaktischer Muster und der Umgang mit poetischen Strukturen wie dem parallelismus membrorum. Die Satzlänge wird formal als

$$Length\ Ratio = \frac{|T|}{|S|}$$

definiert und erlaubt Aussagen über systematische Einfügungen oder Verkürzungen. Besondere Aufmerksamkeit gilt modelltypischen Fehlerarten, die in der MT-Forschung umfassend dokumentiert sind. Dazu zählen Halluzinationen, also die Erzeugung nicht im Quelltext vorhandener Wörter (Raunak et al., 2021), sowie die Over-Normalization, die insbesondere bei komplexen und seltenen Formen auftritt (Müller et al., 2020). Halluzinationen werden technisch erfasst, indem Elemente im Zieltext identifiziert werden, die weder lexikalisch noch strukturell aus dem Quelltext motiviert sind. Over-Normalization beschreibt hingegen die Tendenz des Modells, seltene hebräische oder griechische Formen durch gebräuchliche deutsche Standardformen oder syntaktische Vereinfachungen zu ersetzen.

Schließlich wird die Stabilität der Modelle gegenüber Prompt-Variation untersucht. Diese wird anhand der Ähnlichkeit zweier vom selben Modell erzeugter Übersetzungen gemessen, wobei sich die strukturelle Nähe über die normalisierte Levenshtein-Distanz (Zhao et al., 2021) ergeben. Eine hohe Sensitivität weist darauf hin, dass das Modell keine deterministische Repräsentation der Quellsprache entwickelt, sondern den Übersetzungsstil stark am Nutzereingang orientiert.

In der Gesamtheit ermöglicht die quantitative, technische Analyse eine differenzierte Betrachtung der MT, die sich auf formale und technische Eigenschaften konzentriert und damit die Grundlage für weitere Bewertungen bildet. So werden besondere Eigenschaften historischer Sprachen systematisch abgebildet und die Unterschiede zwischen NMTs und LLMs präzise herausgearbeitet.

3.3.2 Qualitative, theologische Analyse

Die qualitative, theologische Analyse bildet die zweite Säule des Evaluationsverfahrens und ergänzt die zuvor dargestellte technische Untersuchung um eine hermeneutisch orientierte Betrachtungsebene. Während Kapitel 3.3.1 ausschließlich formale und strukturelle Merkmale der Modelloutputs beschreibt, widmet sich die qualitative Analyse der inhaltlichen und theologischen Bedeutung der maschinell erzeugten Übersetzungen. Diese Ebene ist für alttestamentliche Texte unverzichtbar, da Sprache hier nicht nur ein neutrales Kommunikationsmittel darstellt, sondern selbst Träger theologischer Konzepte, kultureller Konnotationen und historisch geprägter Interpretationsspielräume ist. Hebräisch und Griechisch verfügen über semantische Strukturen, deren inhaltliche Tiefenschärfe ohne exegetische Reflexion nicht erfasst werden kann. Die qualitative Analyse zielt daher darauf ab, die von den Modellen vorgenommenen Bedeutungsentscheidungen zu identifizieren und im Hinblick auf ihre theologischen Implikationen einzuordnen.

Grundlage ist hierbei die Erkenntnis, dass maschinelle Übersetzungssysteme – unabhängig von ihrer technischen Ausrichtung – immer auch implizite Interpretationen vornehmen. Dies gilt insbesondere für historisch mehrdeutige Begriffe, bei denen eine lexikalische Entscheidung zugleich eine theologische Gewichtung darstellt. So umfasst das hebräische Lexem עַלְמָה (*ālmah*) etwa ein Bedeutungsspektrum von ‚junge Frau' bis ‚unverheiratetes Mädchen', wobei die Wahl des deutschen Zielbegriffs erhebliche Relevanz für die christologische Auslegung von Jesaja 7,14 besitzt (Donner & Gesenius, 2007). Ähnlich verhält es sich mit dem Gottesoffenbarung אֶהְיֶה אֲשֶׁר אֶהְיֶה *(e-hyeh ăšer ehyeh)* in Exodus 3,14, dessen grammatische Form eine große semantische Bandbreite erlaubt, die vom performativen Selbstprädikat *(„Ich werden sein, welcher ich sein werde")* bis zu einer ontologischen Selbstdefinition reicht. Die qualitative Analyse untersucht daher, wie die Modelle diese lexikalischen und grammatischen Ambiguitäten auflösen, welche Bedeutungsfacetten übernommen oder getilgt werden und ob die Modelle historische Interpretation – etwa der Septuaginta oder der Vulgata – reproduzieren.

Methodisch erfolgt die qualitative Analyse auf mehreren, ineinandergreifenden Ebenen. Die erste Ebene umfasst die lexikalische Bedeutungserschließung. Hier werden semantisch zentrale Begriffe der jeweiligen Perikopen identifiziert und mit den maschinellen Ausgaben vergleichen. Die Analyse orientiert sich an klassischen exegetischen Verfahren, insbesondere der Wortfeldanalyse und der diachronen Semantik. Ihr Ziel ist es, zu bestimmen, ob die Modelle die semantische Breite altsprachlicher Lexeme angemessen wiedergeben oder ob sie zu monosemischer Vereinfachung – gemeint ist die Reduzierung auf nur eine Bedeutung pro Wort – tendieren. Da neuronale Modelle häufig zu normalisierenden Übersetzungsstrategien greifen, bei denen mehrdeutige Begriffe in eine eindeutige Zielsprachform überführt werden, muss die Auswahl der automatischen Metriken angepasst werden. Popović (2015) weist darauf hin, dass verschiedene Metriken, wie z. B. BLEU solche semantischen Verschiebungen nicht sichtbar machen und schlägt mit chrF eine alternative Methode vor. Allerdings misst auch chrF ausschließlich formale Übereinstimmungen und kann daher keine Bedeutungsnuancen oder kontextuelle theologische Interpretationen abbilden. Um die

semantische Bedeutungsvielfalt alttestamtlicher Texte und die dazugehörigen Übersetzungsentscheidungen quantifizieren zu können, ergänzt der BERT Score, eine embeddingbasierte Metrik, die hermeneutische Analyse (Zhang et al., 2020).

Die zweite Ebene betrifft die syntaktisch-semantische Interpretation. Historische Sprachen verfügen über spezifische syntaktische Formen, die inhaltlich bedeutsam sind. Hebräische Nominalsätze ohne Kopulaverb erzeugen semantische Offenheit, da sie sowohl habituelle als auch ontologische Aussagen ermöglichen. Die qualitative Analyse untersucht daher, wie die Modelle u.a. diese Strukturen interpretieren und welche theologischen Verschiebungen dadurch entstehen. Wenn etwa ein hebräischer Nominalsatz als explizites Prädikat wiedergegeben wird, kann dies den Aussagegehalt verengen. So besteht die Gefahr spezifische hermeneutische Lesarten zu verlieren, die im Ausgangstext nicht eindeutig determiniert sind. Die Analyse erfolgt hier nicht normativ, sondern deskriptiv und prüft, ob die modellinterne Strukturinterpretation mit etablierten exegetischen Beobachtungen korrespondiert oder ob sie eigenständige, möglicherweise modernisierte Lesarten generiert.

Die dritte Ebene umfasst die literarisch-theologische Dimension. Biblische Texte operieren häufig mit metaphorischer Bildsprache, poetischen Parellelismen, oder kulturspezifischen Idiomen. Diese Elemente tragen wesentlich zur theologischen Aussage eines Textes bei und müssen daher in der Zielübersetzung angemessen berücksichtigt werden. Die qualitative Analyse untersucht, ob die Modelle metaphorische Strukturen beibehalten oder in erklärende Umschreibungen übertragen. Insbesondere bei poetischen Texten, wie der untersuchten Perikope Psalm 23, ist die Frage ob poetische Strukturen sowohl formal als auch semantisch erhalten bleiben oder ob diese von den Modellen nivelliert werden. Da derartige Strukturen nicht nur ästhetisch sind, sondern auch theologisch relevante Gottesbilder und anthropologische Aussagen beinhalten, sind sie unverzichtbarer Teil der qualitativen Analyse.

Die vierte Ebene betrifft die intertextuelle Kohärenz und die Anschlussfähigkeit an die Übersetzungstradition. In vielen Fällen lässt sich beobachten, dass Modelle – insbesondere solche, die mit großen allgemeinen Datensätzen trainiert wurden – entweder an moderne Zielsprachenkonventionen anschließen oder unbewusst historisch tradierte Übersetzungsmuster reproduzieren. Die qualitative Analyse untersucht daher in einer vierten Ebene, ob maschinelle Übersetzungen im Fall von modernen Zielsprachen, inhaltlich eher an der Tradition der traditionellen (Lutherbibel, King James Version) oder modernen Referenzübersetzungen (BasisBibel, New Living Translation) anschließen oder ob sie eine eigenständige Lesart erzeugen. Dieser Vergleich dient nicht der Feststellung von Richtigkeit, sondern der hermeneutischen Kontextualisierung, da die theologische Bedeutung eines Begriffs oder Satzes häufig aus seiner Wirkungsgeschichte erschlossen wird (Becker, 2021). Auch hier leistet die qualitative Analyse einen Beitrag, den automatische Metriken nicht erfassen können, wie Mathur et al. (2020) betonen, die gerade für die semantische und stilistische Qualitätssicherung die Grenzen metrischer Systeme hervorheben.

4 Analyse und Ergebnisse

Im folgenden Kapitel wird der Output der vier ausgewählten Systeme – DeepL, NLLB, Gemini und LLaMA-3 – systematisch den etablierten Referenzübersetzungen gegenübergestellt. Ausgangspunkt ist dabei jeweils der hebräische Quelltext, der zunächst in seiner Gattung verortet und anschließend mit den klassischen altsprachlichen Fassungen (LXX, Vulgata) sowie ausgewählten deutschen und englischen Bibelübersetzungen abgeglichen wird. Entlang der drei Perikopen Ex 3,11–15, Jes 7,10–16 und Ps 23,1–6 werden dabei nicht einzelne schöne oder auffällige Beispiele isoliert betrachtet, sondern wiederkehrende Muster der Modelle herausgearbeitet: zunächst im Bereich der Narration, wo sich eine vergleichsweise hohe Stabilität zeigt, anschließend in der Prophetie mit ihrer verdichteten Bildsprache und schließlich in der Poesie, die sich als deutlich störanfällige Gattung erweist.

Das Kapitel verfolgt dabei zwei Ziele: Zum einen werden typische Fehlertypen beschrieben – von orthografischen und segmentierungsbedingten Oberflächenphänomenen über semantische Drifts bis hin zu theologisch relevanten Bedeutungsverschiebungen bei Schlüsselbegriffen wie אֶהְיֶה אֲשֶׁר אֶהְיֶה (*ehyeh ăšer ehyeh*), הָעַלְמָה (hāalmāh) oder dem Tetragramm יְהוָה. Zum anderen wird sichtbar gemacht, in welchen Konstellationen die Modelle den Referenzübersetzungen näherkommen und wo sie sich deutlich von ihnen entfernen. Die Gegenüberstellung versteht sich damit als qualitative Vorarbeit für die anschließende technische Auswertung in Kapitel 4.2 sowie die theologisch-hermeneutische Vertiefung in 4.3: Sie markiert jene Stellen, an denen maschinelle Übersetzungen bereits heute als heuristische Unterstützung genutzt werden können, und jene Bereiche, in denen sie aus philologischer und theologischer Perspektive deutlich hinter den etablierten Übersetzungen zurückbleiben.

4.1 Vergleich mit etablierten Übersetzungen

Im Vergleich mit etablierten Übersetzungen zeigt sich zunächst, wie stark die maschinellen Modelle einerseits an klassische Übersetzungsmuster anknüpfen, andererseits aber auch regelmäßig von ihnen abweichen. Diese Abweichungen sind nicht zufällig, vielmehr verweisen sie auf charakteristische Muster, die sich entlang der drei untersuchten Gattungen – Narration, Prophetie und Poesie – deutlich abzeichnen. Bereits bei der narrativen Perikope Exodus 3,11–15 (Anhang 1-1 bis Anhang 1-10), in der Mose vor Gott zurücktritt und den Gottesnamen offenbart erhält, wird sicherbar, dass alle Modelle auf den ersten Blick stabil wirken. Die Satzstrukturen im Hebräischen – etwa in Folge von Verbalsätzen, die Wiederholung von אֶהְיֶה אֲשֶׁר אֶהְיֶה (*ehyeh ăšer ehyeh*) oder die Funktion der Nominalsätze – sind von vergleichsweise geringer syntaktischer Komplexität. Die Modelle scheinen offenkundig gerade bei narrativen Texten die höchste Zuverlässigkeit zu besitzen. Dennoch zeigen sich einerseits gravierende Übersetzungsfehler, auf der anderen Seite beim genauerem Hinsehen wiederkehrende Muster mit bedeutsamen theologischen Fehlern.

So übersetzt DeepL beispielweise mehrfach Satzanfänge ohne korrekte Großschreibung oder lässt Interpunktion weg – eine Form von Oberflächenfehlern, die philologisch leicht zu beheben wären, deren Auftreten aber drauf hinweist, dass das Modell intern keine explizite Satzschreibung erzwingt. NLLB liefert mit *„daß“* eine seit knapp 30 Jahren veraltete Rechtschreibung und setzt auch im Englischen mit *„shalt“* auf nicht mehr gebräuchliche Ausdrucksformen. Noch auffälliger sind die Fälle, in denen DeepL Leerzeichen falsch behandelt. In Ex 3,14 (Anhang 1-8) erscheint etwa „*zumose“* an Stelle von *zu Mose*, und im Englischen „*tomoses“*, obwohl die Modelle offensichtlich keinen semantischen Grund für die Zusammenschreibung besitzen. Einzig der hebräische Bindestrich *Maqqef* lässt bei אֶל־מֹשֶׁה *(’el-mosheh)* den Grund für das fehlende Leerzeichen vermuten. Dazu skizziert sich eine grundlegende Herausforderung moderner NMTs, die über kein lexikalisch-gestütztes Inventar heiliger Namen verfügen – anders als bei einer regelbasierten Übersetzung. Daher zeigen diese Fehler, dass selbst hochentwickelte Systeme die Segmentierung biblischer Eigennamen und dazugehöriger Präpositionen nicht zuverlässig beherrschen, selbst wenn sie in zeitgenössischen Sprachpaaren problemlos funktionieren. Noch deutlicher wird dies in der von DeepL erzeugen Übersetzung von Ex 3,12, in der das verdoppelte Wort „*dasdas“* erzeugt wird – eine Form maschineller Halluzination, wie sie im Englischen als „*thisthis“* ebenfalls vorkommt. Derartige Doppelungen sind charakteristisch für Fehltrigger im Tokenisierungsvorgang, bei denen das Modell denselben Token zweimal generiert, weil der Übersetzungswahrscheinlichkeitsraum kurzfristig instabil wird. Sie wirken trivial, zeigen aber, dass selbst in einfachen Kontexten keine garantierte formale Konsistenz besteht.

Hinzu kommen semantische Driftphänomene, die sich besonders bei der Wiedergabe von אֶל־בְּנֵי יִשְׂרָאֵל (*el-benê Jiśrāēl*) zeigen. Obwohl der hebräische Plural בְּנֵי (*benê* – „Söhne“*)* sich eindeutig übersetzen lässt, wechseln die Modelle unstet zwischen *„Söhne Israels“* und *„Kinder Israels“*. LLaMA-3 übersetzet im Englischen noch in V.11 *„sons of Isreal“*, nur um wenige Verse später auf

„children“ umzuschwenken. Gegensätzlich dazu bleibt die Übersetzung mit Gemini im Deutschen stabil. Dieser Wechsel ist nicht als bewusste inklusionsorientierte Anpassung zu verstehen, sondern als statistisches Echo der Trainingsdaten, in denen beide Varianten vorkommen – jedoch ohne hermeneutische Differenzierung. Die etablierten Übersetzungen zeigen hier eine klare Konstanz: LXX (υἱοὺς Ἰσραήλ) und Vulgata (filiis Israel) bevorzugen ausnahmslos die genealogisch präzise Wiedergabe, während Luther und die KJV sich ebenfalls terminologisch festgelegt haben. Die maschinellen Modelle hingegen geben teilweise ein unstetes Bild ab.

Eine weitere zentrale Beobachtung betrifft die Offenbarung des Gottesnamens: die hebräische Formel אֶהְיֶה אֲשֶׁר אֶהְיֶה (*ehyeh ăšer ehyeh*), ist als Imperfektform sowohl präsentisch als auch futurisch interpretierbar, wird aber von den Modellen unterschiedlich realisiert. Während einige Modelle eine futurische Lesart annähern, scheitert insbesondere NLLB – gerade im Englischen – daran, den verbalen Charakter wiederzugeben. Auch das Tetragramm als Gottesname יְהוָה wird von den Modellen inkonsistent behandelt. DeepL präferiert *„Herr“*, NLLB greift auf *„Jehova“* zurück, Gemini verwendet *„JHWH“* und LLaMA-3 setzt *„HERR“* mit Majuskeln. Diese vierfache Variabilität ist aus linguistischer Perspektive insofern problematisch, als der Gottesname traditionell nicht semantisch, sondern kultisch übersetzt bzw. paraphrasiert wird. Die klassischen Übersetzungen wählen durchgängig *„Herr“* oder Dominus. Dass maschinelle Modelle ohne erkennbare Regeln zwischen verschiedenen Formen alternieren, zeigt, dass sie aufgrund der Trainingsdaten lediglich statistische Wahrscheinlichkeiten ausgeben, aber keine hermeneutische Kohärenz besitzen.

Trotz dieser Probleme lässt sich festhalten, dass narrative Texte insgesamt stabil wiedergegeben werden, sodass die entstandenen Übersetzungen auch hermeneutische und theologische Analysen zulassen. Die Modelle arbeiten überwiegend wörtlich und bilden meist erkennbare Parallelitäten zwischen der deutschen und englischen Übersetzung, wodurch der Sinn größtenteils erhalten und erkennbar bleibt. Die Narration ist also über alle Modelle gut lesbar und zeigt eine hohe Stabilität. Schwierigkeiten entstehen erst dann, wenn ein Satz mehrere Aspekte hat, woraufhin die Modelle mitunter ungenaue und verschwimmende Strukturen bilden. Diese Phänomene entsprechen bekannter Schwächen neuronaler Modelle bei der Verarbeitung komplexer Satzgefüge, die mehrstufige Informationshierarchien erfordern.

Ein signifikant anderes Bild ergibt sich bei der prophetischen Perikope Jesaja 7,10–16. Bereits der Einstieg zeigt eine gravierende Instabilität – allerdings nur in der deutschen Übersetzung: Die Wiedergaben des Eigennamens אָחָז (*Āḥāz*) variieren erheblich von „Ahas“ (DeepL) über „Ohaz“ (NLLB) bis hin zur völlig unzutreffenden Form „Achab“ bei LLaMA-3, die sogar einen anderen König bezeichnet. Die Wiedergabe im Englischen dagegen bleibt stabil. Dieser Kontrast macht deutlich, dass insbesondere die deutschen Trainingsdaten an dieser Stelle weniger konsistent bezüglich altorientalischer Eigennamen sind, sodass Modelle phonetisch ähnliche Formen generalisieren. Der Effekt, dass im Englischen keine vergleichbare Variation auftritt, bestätigt die Vermutung: Die Modelle priorisieren bei uneindeutigen Signalmustern stabilere Korpusmuster.

Ein weiteres ausgeprägtes Fehlerbild betrifft die komplexen hebräischen Parallelstrukturen. Die Formulierung הַעְמֵק שְׁאָלָה אוֹ הַגְבֵּהַ לְמָעְלָה (*haămeq šeālāh ô hagbēah ləmalāh*, – „fordere tief unten oder hoch oben“) bildet im Hebräischen einen klaren antithetischen Parallelismus. Während LXX und Vulgata sowie die modernen Referenzübersetzungen diese Struktur präzise wiedergeben, scheitert NLLB als einziges Modell daran, sowohl im Deutschen als auch im Englischen, den Kontrast zu erfassen und erzeugt Formulierung, die weder syntaktisch klar noch semantisch korrekt sind. Dass gerade ein NMT-System den parallelen Aufbau nicht erkennt, zeigt, wie sehr semitische Stilfiguren von den Modellen übergangen werden, wenn sie nicht mit expliziten Attention-Mustern verknüpft sind. Hinzu kommt in Jes 7,12 die Fehlinterpretation des Verbs אֲנַסֶּה (*anasséh*, – „ich will nicht versuchen“). NLLB generiert mit *„vergessen“* eine kaum erkennbare Form, während LLaMA-3 im Deutschen das Verb fälschlich mit *„forschen“* wiedergibt, obwohl im Englischen dasselbe Modell mit *„try“* eine überraschend passende Entsprechung liefert. Diese divergente Performance innerhalb desselben Modells belegt, dass semantische Entscheidungsprozesse nicht sprachübergreifend stabil sind, sondern kontextuell fragmentiert ablaufen.

Besonders gravierende Abweichungen erscheinen jedoch in Jes 7,14, der bekannten Immanuel-Verheißung. Die Übersetzung von הָעַלְמָה (*hāʿalmāh*) variiert extrem: NLLB produziert im Englischen *„widow“*, was semantisch völlig verfehlt ist. Gemini interpretiert *„young woman“,* während DeepL und LLaMA-3 sowohl im Deutschen als auch im Englischen wiederholt *„Jungfrau“* bzw. *„virgin“* wählen. Hier zeigt sich, dass auch die Art des eingesetzten Modells keine Relevanz auf die Übersetzungsentscheidung hat. Während sich Gemini und LLaMA-3 als LLMs sowie DeepL und NLLB noch in der Narration typweise für gemeinsame Übersetzungen entscheiden haben - z. B. Söhne Israels und Kinder Israels – signalisiert die Prophetie ein anderes Bild. Die Divergenz spiegelt die bekannte Spannung zwischen der masoretischen Lesart („junge Frau“) und der christologisch geprägten LXX-Interpretation παρθένος („Jungfrau“) wider. Doch anders als etablierte Übersetzungen, die sich bewusst für eine der beiden Linien entscheiden, spiegeln die Modelle lediglich die statistische Uneinheitlichkeit ihrer Trainingsdaten wider. Auch im folgenden Vers treten deutliche Fehlleistungen auf. Die Phrase חֶמְאָה וּדְבַשׁ (*ḥemāh ûdəvaš*, – „Butter/Dickmilch und Honig“) wird von NLLB sowohl im Deutschen als auch im Englischen derart verzerrt wiedergegeben, dass der zugrundeliegende Sinn nicht mehr erkennbar ist. LLaMA-3 hingegen überträgt den Sinn im Deutschen zwar formal verständlicher, jedoch in einer unnatürlichen, übermäßig wörtlichen Struktur. Im Englischen produziert dasselbe Modell die Aussage *„he shall eat to know evil and to choose the bad“*, was eine vollständige Verdrehung des hebräischen Sinns darstellt, da aus der Fähigkeit des Kindes, moralisch zu unterscheiden, fälschlich eine Art absichtlichen Strebens nach dem Bösen konstruiert wird.

Diese prophetischen Beispiele zeigen in ihrer Gesamtheit, dass die Modelle in dieser Gattung deutlich instabiler arbeiten. Prophetische Sprache ist dichter, bildhafter und syntaktisch weniger linear als narrative Prosa. Prophetie stellt alle Modelle vor Probleme, und die Übersetzungen nähern sich zwar den Referenzübersetzungen an, erreichen jedoch keine durchgehende Kohärenz.

Während die prophetische Passage bereits erhebliche Schwierigkeiten für die Modelle sichtbar macht, zeigt sich in der poetischen Perikope Psalm 23 eine nochmals deutlich intensivere Instabilität, die sowohl syntaktische als auch semantische Ebenen betrifft. Die Übersetzungen (Anhang 3-1 bis Anhang 3-12) belegen klar, dass die Modelle in einem poetischen Kontext teils kaum in der Lage sind, die hebräische Ausgangsstruktur adäquat zu erfassen, noch weniger können sie eine inhaltlich belastbare oder stilistisch kohärente Zielsprachform erzeugen. Psalm 23 stellt aufgrund seiner dichten Bildsprache, der kulturspezifischen idiomatischen Wendungen und des parallelismus membrorum hohe Anforderungen an jede Übersetzung. Dass die maschinellen Modelle bei dieser Gattung besonders stark einbrechen, gliedert sich in die gesteigerte Komplexität zwischen den drei ausgewählten Perikopen.

Bereits der erste Vers lässt grundlegende Fehlentwicklungen erkennen. Während die traditionelle Lutherübersetzung „Der Herr ist mein Hirte, mir wird nichts mangeln“ eine ruhige, klare und semantisch weite Aussage formuliert, neigt LLaMA-3 dazu, die Bedeutung von אֶחְסָר (*eḥsār*, – „ich werde nicht fehlen“) stark zu verengen und mit *„hungern“* wiederzugeben. Damit verliert der Satz seine metaphorische und theologische Tiefe und reduziert sich auf eine körperlich-biologische Dimension, die im Hebräischen nicht intendiert ist. Die englischen Modellausgaben dagegen bleiben in allen Modellen stabil, was darauf hindeutet, dass gerade das deutsche Sprachmodell von LLaMA-3 weniger robuste Vorkommen des Psalms in seinen Trainingsdaten besitzt als die englischsprachigen Gegenstücke.

Der zweite Vers verstärkt dieses Bild. NLLB gelingt es im Deutschen nicht, eine klare, grammatisch kohärente Struktur zu bilden, stattdessen entsteht eine verdoppelte, semantisch unklar verschobene Struktur, die auf eine fehlerhafte interne Verarbeitung der poetischen Bildlogik hindeutet. Damit bestätigt sich ein Muster: NLLB produziert teilweise Satzfragmente oder Doppelungen, ohne erkennbare Grundlage im hebräischen Quelltext. Dabei ist zu beachten, dass die englischen Übersetzungen von Gemini bei diesem Vers deutlich stabiler sind, was die These eines stabileren Trainingstextkorpus im Englischen bestätigt.

Im dritten Vers zeigt sich ein besonders gravierendes Fehlerbild. LLaMA-3 scheitert im Deutschen daran, Personal- und Reflexionspronomen korrekt zuzuordnen. Die Formulierung *„Meine Seele wird mich umkehren …“* stellt eine klare Fehlinterpretation dar, da der Text zwar נַפְשִׁי (*nafšī*, – „meine Seele“) und יְשׁוֹבֵב (*ješôvēv*, – „er bringt zurück“) enthält, aber keine syntaktische Konstruktion, in der die Seele aktiv auf das Subjekt zurückwirkt. Diese Fehlleistung verdeutlicht, dass alle Modelle, besonders aber die lokal ausgeführten, die grammatischen Funktionsbeziehungen nicht sicher erkennen und insbesondere im poetischen Kontext Schwierigkeiten haben, hebräische Verbvalenzen korrekt aufzulösen. Noch entscheidender ist die Beobachtung, dass kaum eins der Modelle im Deutschen eine theologisch verwertbare Übersetzung für diesen Vers liefert. Im Englischen zeigt sich eine ähnliche Tendenz, wenn auch abgeschwächt. Nur Gemini erzeugt eine Version, die inhaltlich verständlich ist, während die übrigen Modelle zwar einzelne Schlüsselbegriffe wie נַפְשִׁי korrekt wiedergeben, aber darin scheitern, den יְשׁוֹבֵב angemessen zu interpretieren.

Der vierte Vers zeigt exemplarisch, wie stark maschinelle Modelle bei poetischen Texten kollabieren können, wenn metaphorische Ausdrücke hinzukommen. Besonders auffällig ist LLaMA-3 im Deutschen: Die Übersetzung von וּמִשְׁעַנְתֶּךָ (*ûmišantēkā*, – „und dein Stab“) wird in eine Form überführt, die nicht einmal mehr derselben Wortfamilie angehört. Der Begriff *„Steigerung“*, den das Modell generiert, steht in keiner Beziehung zum hebräischen Ausgangstext und spiegelt damit eine Fehlassoziation wider, die darauf hindeutet, dass die poetische Umgebung vom Modell falsch modelliert wird. Besonders beachtenswert ist der Kontrast zum Englischen: LLaMA-3 gelingt an derselben Stelle eine korrekte Übersetzung. Dies zeigt erneut, dass selbst innerhalb desselben Modells unterschiedliche Sprachroutinen unabhängig voneinander scheitern oder gelingen können. Ein weiterer Befund betrifft NLLB, das den Satz in diesem Vers im Deutschen deutlich verkürzt und im Englischen als einziges Modell keine Form hervorbringt, die erkennbar eine Verbindung zur Referenzübersetzung hat – weder die traditionelle KJV, noch die modernere NLT. Es wird deutlich, dass kürzere, pauschale Vorkommnisse poetischer Wendungen in Trainingsdaten nicht ausreichen, um die Modelle zu stabilisieren. Komplexere semantische Abschnitte erscheinen leichter übersetzbar, da sie mehr Kontext bieten, den die Modelle für ihre Wahrscheinlichkeitsverteilungen nutzen können.

Auch der fünfte Vers besteht als schwierige Stelle, vor allem für die lokal ausgeführten Modelle NLLB und LLaMA-3. Ersteres verfehlt hier bereits im Ansatz die Fähigkeit, überhaupt einen Satz zu bilden, der mit dem hebräischen Quelltext in Verbindung zu bringen wäre. LLaMA-3 erzeugt im Deutschen eine Formulierung, die zwar einzelne Motive andeutet, aber eine insgesamt unbrauchbare, fast karikierende Bildlichkeit hervorbringt: *„mein Haupt hat ein Ölbad“*. Dieses Beispiel demonstriert ein weiteres Muster: Nahezu alle Modelle greifen in einem Kontext, den sie nicht sicher interpretieren können, auf interne stereotypische Muster zurück, die zwar thematisch verwandt sind, aber nicht dem biblischen Text entsprechen. Im Englischen zeigt NLLB ein ähnliches Verhalten, indem es Teile des Textes ganz auslässt. Dieses Ergebnis wirkt modelluntypisch mehr wie eine generierten Halluzination als eine wortgetreue Übersetzung. NLLB findet im poetischen Kontext keine verlässlich semantische Struktur und generiert statt einer wörtlichen Übersetzung unverständliche Sequenzen, die den Quelltext bei weitem verfehlen. Demgegenüber gelingen DeepL und Gemini sowohl im Deutschen als auch im Englischen eine erkennbar verwendbare Übersetzung zu erzeugen, was darauf schließen lässt, dass diese Modelle stärkere Regularitäten in ihren Sprachrepräsentationen für poetische Bildlichkeiten besitzen und von größeren Datenmengen profitieren, die poetische Texte im modernen Sprachgebrauch umfassen.

Im abschließenden Vers des Psalms zeigt sich nochmals die deutliche Divergenz zwischen den Modellen. DeepL etwa interpretiert יִרְדְּפוּנִי (*yirdəfūnī*, – „sie werden mir folgen“) negativer und gibt es mit „*verfolgen*“ wieder. Diese Bedeutungsverschiebung ist besonders interessant, da das Verb רדף (*radaf* – „hinter jmd. her sein“) zwar je nach Kontext eine aggressive Konnotation tragen kann, aber im Psalm bewusst positiv intendiert ist. Die Übersetzung als *„verfolgen“* verzerrt daher den theologischen Gehalt des Verses. Auffällig ist auch, dass Gemini im Englischen als einziges Modell

die semantisch zutreffende Übersetzung von חֶסֶד (*ḥesed*, – „Gnade“/„Barmherzigkeit“) mit „*mercy*“ erreicht. Bei der Wiedergabe von בְּבֵית־יְהוָה לְאֹרֶךְ יָמִים (*bəvêt-YHWH ləōrek jāmīm*, – „im Haus des Herrn für die Länge der Tage“) neigen Gemini und LLaMA-3 dazu, die Wendung wörtlich zu reproduzieren, während DeepL und NLLB die Struktur stärker glätten und beispielsweise mit „*for ever*“ wiedergeben. Diese Glättung steht einerseits im Einklang mit moderner Bibelsprache – wie etwa die Referenzübersetzungen der BasisBibel und der NLT – zeigen aber auch deutlich, dass die Modelle bei poetischen Wendungen und Bildsprache divergieren.
Damit wird im gesamten Psalm deutlich, dass poetische Texte für neuronale Übersetzungssysteme die größte Herausforderung darstellen. Der Grund liegt dabei nicht nur in der Bildsprache oder der metaphorischen Dichte, sondern vor allem in der Kombination dreier Faktoren: zuerst den kurzen, kompakten Strukturen der einzelnen Verse, die wenig Kontext für die Wahrscheinlichkeitsverteilung der Modelle liefern. Zum zweiten, die semitische Parallelstruktur, die anders als narrative Syntax nicht linear fortschreitet. Und zuletzt der theologischen Tiefendimension, die oft von einzelnen Verben oder Begriffen getragen wird, deren Bedeutungsfeld die Modelle mitunter nicht zuverlässig erfassen können.

In den altsprachlichen Zielsprachen zeigt sich ein differenziertes Bild, das zwar dieselben Grundmuster wie in den modernen Sprachen erkennen lässt, diese jedoch in anderer Weise ausprägen. Während Griechisch und Latein aufgrund ihrer festen Übersetzungstradition eigentlich eine größere Stabilität erwarten lassen würden, gelingt es den Modellen nur teilweise, diese Tradition abzubilden. Die Systeme greifen häufig auf altsprachlich klingende Formen zurück, ohne jedoch konsequent der Logik der LXX oder der Vulgata zu folgen. Dadurch entsteht eine Oberfläche, die philologisch vertraut wirkt, semantisch aber nicht dieselbe Zuverlässigkeit aufweist wie bei etablierten Übersetzungen.
Besonders deutlich wird dies an Stellen, in denen LXX und Vulgata historisch eindeutige, fest tradierte Lösungen anbieten. So ist die Wiedergabe von אֶל־בְּנֵי יִשְׂרָאֵל (*el-benê Jiśrāēl* – wörtl. „Söhne Israels“) in beiden klassischen Übersetzungen vollständig stabil (υἱοὺς Ἰσραήλ bzw. *filiis Israel*). Die Modelle hingegen reproduzieren diese Stabilität nicht. Teilweise übernehmen sie die tradierten Formen, wechseln aber innerhalb derselben Passage zu alternativen Ausdrücken oder verlieren die genealogische Bedeutung des Begriffs בְּנֵי. Dieses Verhalten entspricht zwar dem im Deutschen und Englischen beobachteten Muster, wird aber im Griechischen und Lateinischen dadurch auffälliger, dass es eine seit Jahrhunderten feste Übersetzungsnorm unterläuft. Die Instabilität ist daher nicht nur eine lexikalische, sondern auch historisch-hermeneutische Abweichung.
Der Umgang mit אֶהְיֶה אֲשֶׁר אֶהְיֶה (*ehyeh ăšer ehyeh* – „Ich werde sein, der ich sein werde.“) zeigt ebenfalls eine geringere Bindung an tradierte Formen, als man es erwarten könnte. LXX und Vulgata reproduzieren hier nicht den hebräischen Verbmodus, sondern eine theologisch interpretierende Lesart. Die Modelle hingegen schwanken zwischen futurisch verstandenen und stärker normalisierten Fassungen und erzeugen sowohl im Griechischen als auch im Lateinischen Strukturen, die weder der hebräischen Vieldeutigkeit noch der Tradition entsprechen. Damit wiederholt sich

nicht einfach das Problem von Übersetzungen in moderne Sprachen, sondern es entsteht ein zusätzlicher Konflikt zwischen der statistischen Modelllogik und der überlieferten altsprachlichen Form.

Daneben setzt sich die bereits im Deutschen und Englischen beobachtete Inkonsistenz im Umgang mit dem Gottesnamen יְהוָה ebenso in den altsprachlichen Validierungssprachen fort. Obwohl LXX als auch Vulgata diesen fast ausnahmslos als κύριος bzw. Dominus wiedergeben, erzeugen die Modelle uneinheitliche Lösungen: mal transliterierte Formen, mal moderne Umschreibungen und vereinzelt auch direkte Bezeichnungen. Das zeigt, das neuronale Modelle die traditionelle kultische Funktion des Tetragramms nicht zuverlässig reproduzieren. Während die Fehlleistung im modernen Sprachraum oft stilistisch wirkt, unterläuft sie im griechisch-lateinischen Kontext sogar den Kern der historischen Übersetzungspraxis. In prophetischen Texten verstärkt sich dieser Befund. Die Fehlinterpretation von Eigennamen, die im Deutschen besonders deutlich werden, treten in ähnlicher Form auch in griechischen und lateinischen Modellübersetzungen auf. Das zeigt, dass die Fehlleistung nicht nur eine Frage der Zielsprache ist, sondern strukturell im Modell selbst angelegt ist. Auch hier greifen die Systeme gelegentlich auf phonetisch ähnliche, aber sachlich falsche Namen zurück – ein Umstand, der vor allem deshalb auffällt, weil die LXX an diesen Stellen sehr stabil ist und die Modelle diese Stabilität dennoch nicht übernehmen. Die semantische Divergenz beim Begriff הָעַלְמָה (*hāʿalmāh*) setzt sich in den altsprachlichen Ausgaben nahezu unverändert fort. Während die LXX eindeutig παρθένος und die Vulgata virgo verwendet, zeigen die Modelle teilweise dieselbe Übersetzungssteuerung wie in modernen Sprachen. Das bedeutet, dass selbst dort, wo eine eindeutige historische Norm existiert, können die Modelle keine robuste Orientierung daran entwickeln. Die Konsequenz ist, dass nicht nur die hebräische Vorlage, sondern auch die traditionsgeschichtliche Tiefenschicht der altsprachlichen Bibelübersetzungen verfehlt wird.

Die bekannten Probleme mit metaphorischer und poetischer Sprache treten auch in den altsprachlichen Zielsprachen hervor, wenn auch in anderer Form. Die Wendung חֶמְאָה וּדְבַשׁ (*ḥemāh ûdəvaš* – „Butter/ Dickmilch und Honig“) etwa, die in LXX und Vulgata ebenfalls klar kodifiziert ist, wird vor allem von den lokal ausgeführten Modellen durch Bedeutungsverschiebungen oder unvollständigen Wiedergaben verfremdet. Ebenso lassen sich Fehlinterpretationen theologischer Aussagen beobachten, etwa wenn LLaMA-3 die ethische Reifung des Kindes in Jes 7,15f, in eine Wahl des Bösen umkehrt – ein Fehler, der in den altsprachlichen Ergebnissen nicht immer wortgleich auftritt, aber dieselbe Verzerrung erkennen lässt.

Diese Kombination zeigt, dass die Übertragung ins Altgriechische und Lateinische trotz einer scheinbar höheren strukturellen Nähe zu den tradierten Formen nicht zuverlässiger ist. Die Modelle verfügen zwar über altsprachliche Morphologie und Syntax, aber diese Fähigkeiten wirken oft wie isolierte Bausteine, die nicht durch ein strukturelles Verständnis der biblischen Textlogik verbunden sind. Wo die LXX und Vulgata eine klare, über Jahrhunderte eingeübte philologische Linie verfolgen, unterlaufen die Modelle diese Linie regelmäßig zugunsten statistischer Muster.

In der Zusammenschau der Ergebnisse wird deutlich, dass die maschinellen Übersetzungen zwar in bestimmten Bereichen an etablierte Bibelübersetzungen anschließen, insgesamt jedoch eine deutliche Varianz und unterschiedlich ausgeprägte hermeneutische Verlässlichkeit aufweisen. Während narrative Strukturen noch eine hohe Stabilität unter allen Modellen zeigen, geraten prophetische und insbesondere poetische Texte schnell an die Grenzen der Systeme. Hier lösen sich semantische Präzision, metaphorische Kohärenz und syntaktische Rollenverteilung teilweise so weit auf, dass die erzeugten Fassungen mitunter nur eingeschränkt an eine philologische oder theologische Auswertung anschlussfähig sind. Die altsprachlichen Validierungssprachen verstärken diesen Eindruck zusätzlich: Trotz des Eindrucks formaler Korrektheit gelingt es den Modellen nicht, die traditionsgebundenen Übersetzungsentscheidungen der LXX und Vulgata zuverlässig abzubilden. Vielmehr entstehen auch dort statistische Approximationen, die nur oberflächlich an klassische Strukturen erinnern, jedoch ohne deren interpretative Teile auskommen. Insgesamt zeigt sich, dass die Modelle biblische Texte nicht als Teil einer gewachsenen Auslegungstradition verarbeiten, sondern als lineare sprachliche Sequenzen berechnen. Damit erfüllen sie zwar punktuell funktionale Anforderungen, erreichen aber nicht die Konsistenz und Präzision, die etablierte Übersetzungen kennzeichnet.

Im Hinblick auf die einzelnen Systeme lassen sich übergreifende Tendenzen erkennen. **DeepL** zeigt im narrativen Bereich häufig solide Ergebnisse, verliert aber bei längeren oder komplexeren Strukturen an Genauigkeit und produziert formale Oberflächenfehler, die auf inkonsistente Segmentierung hinweisen. **NLLB** fällt durch deutliche Instabilitäten auf, insbesondere bei poetischen Texten und bei der Zuordnung syntaktischer Rollen. Gleichzeitig weist es vereinzelt Stärken in der wörtlichen Annäherung an altsprachliche Formen auf, ohne diese konsequent zu erhalten. **Gemini** erweist sich insgesamt als das stabilste der getesteten Modelle, insbesondere im Englischen, und erreicht mehrfach Übersetzungen, die inhaltlich und strukturell deutlich näher an etablierten Fassungen liegen. Dennoch zeigt auch Gemini keine einheitliche Strategie für theologisch zentrale Begriffe. **LLaMA-3** ist das am stärksten schwankende System: Es liefert in einzelnen Passagen nachvollziehbare Ergebnisse, weist jedoch gleichzeitig deutliche Fehlübertragungen bis hin zu völligen Bedeutungsumkehrungen auf. Insgesamt bestätigt sich damit, dass kein Modell eine durchgehend zuverlässige Übersetzungsleistung für hebräische Bibeltexte erbringt, wenngleich alle Systeme in begrenztem Umfang nutzbare Teilaspekte erkennen lassen.

4.2 Quantitative, technische Analyse

Die in Kapitel 3.3.1 eingeführte quantitative, technische Analyse wird im Folgenden auf die erzeugten Übersetzungen der vier untersuchten Systeme angewendet. Ziel ist es, die formalen Eigenschaften der Modelloutputs systematisch zu beschreiben und die Unterschiede zwischen NMT-Systemen (DeepL, NLLB-200) und LLM-basierten Ansätzen (Gemini 2.5 Pro, LLaMA-3) sichtbar zu machen, ohne bereits eine theologische Deutungsebene zu betreten. Im Mittelpunkt stehen dabei die definierten Kennzahlen – die Token Fragmentation Rate als Maß für die Subwordsegmentierung, die Alignmet-Struktur zwischen Quell- und Zieltexten, das Längenverhältnis der Sätze, modelltypische Fehlerphänomene wie Halluzination und Over-Normalization. Diese Größen erlauben es, die Übersetzungsleistung zunächst auf der Ebene von Form, Struktur und Robustheit zu bewerten, bevor im darauffolgenden Abschnitt die semantischen und theologischen Implikationen in den Blick genommen werden.

4.2.1 Bewertung der Übersetzungsqualität nach Metriken

Length Ratio | Längenverhältnis

Die Length Ratio erlaubt eine Einschätzung der strukturellen Transformation vom Quelltext zum Zieltext. Werte über 1.0 markieren eine Expansion, also eine Verlängerung des Satzes basierend auf der Wortzahl. Gegenteilig wird bei Werten unter 1.0 von einer Kompression gesprochen, einer Verkürzung des Quellsatzes. In allen drei Perikopen zeigt sich, dass sämtliche Modelle signifikant erweitern, wobei die Höhe und Streuung dieser Erweiterung textgattungs- und modellabhängig ist. Die nachfolgenden Grafiken zeigen das Längenverhältnis der einzelnen Perikopen und Modelle. Zur Einordnung wird die Lutherübersetzung als Referenzlinie eingeführt. Alle Werte basieren auf den deutschen Übersetzungsausgaben der verschiedenen Modelle, die in Anhang 4-1 zu finden sind.

Exodus 3, 11–15 **Jesaja 7, 10–16** **Psalm 23, 1–6**

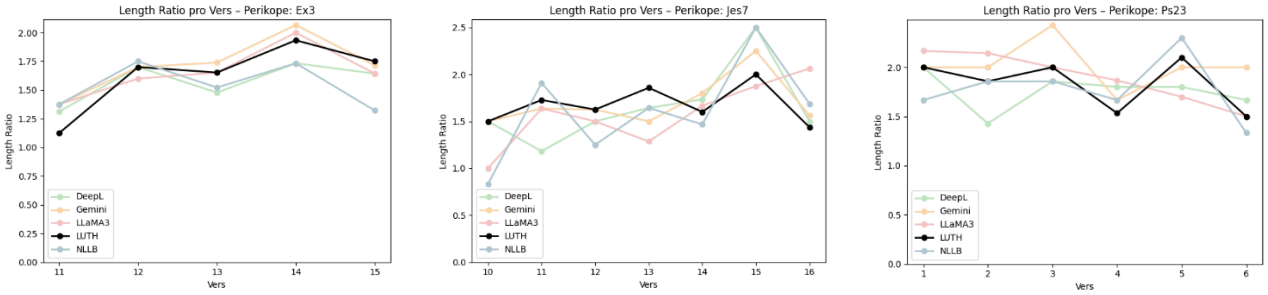


Abbildung 4-1 *Liniendiagramme – Length Ratio | Längenverhältnis*

Die narrativen Verse von Exodus 3 weisen deutlich geringere Varianz auf. Hier bewegen sich sowohl die Werte der Lutherübersetzung als auch die aller maschinellen Übersetzungen größtenteils im moderaten Bereich von etwa 1.3 bis 1.8 – nur LLaMA-3 und Gemini nähern sich im 14. Vers punktuell der Verdopplungsgrenze. Diese Abweichungen treten häufig in solchen Versen auf, in denen die Modelle zusätzliche Markierungen von Sprecherwechseln, expliziten Tempusangaben

oder syntaktische Klarstellungen einfügen. Zu beachten ist ebenfalls, dass in der narrativen Perikope eine Gruppierung zwischen spezialisierten NMT-Systemen und generativen LLMs sichtbar wird. Über annähernd die gesamten fünf betrachteten Verse liegen die Längenverhältniswerte der NMTs auf der einen Seite und LLMs auf der anderen Seite fast deckungsgleich. Dazu lässt sich erkennen, dass LLaMA-3 und Gemini als generative Modelle über der Referenzlinie der Lutherübersetzung liegen und damit eine weniger wörtliche Übersetzung darstellen. Auch für prophetische Texte lässt sich eine höhere Stabilität der cloudbasierten Systeme gegenüber den lokal ausgeführten Modellen wahrnehmen.

In der prophetischen Perikope zeigt sich ein intermediäres Bild: Die Werte der Lutherübersetzung liegen in weiten Teilen zwischen 1.2 und 1.8, während gerade NLLB stark ausgeprägte Fluktuationen zeigt. Besonders auffällig ist Vers 15, der von fast allen Modellen eine deutliche Erweiterung des Quelltextes beansprucht. Dies weist auf eine modelltypische Tendenz zur interpretativen Ausfaltung hin, die über das hinausgeht, was im Rahmen einer historisch-philologischen Übersetzung notwendig wäre.

Besondere Aufmerksamkeit verdienen Werte oberhalb von 2.0, die auf eine Verdoppelung der Zieltextlänge hindeuten. Solche Werte treten regelmäßig in Psalm 23 auf, besonders etwa bei LLaMA-3 in Vers 3, aber auch bei Gemini in mehreren Versen der Perikope. Diese hohen Expansionswerte lassen sich durch die poetische Struktur des Ausgangstextes erklären, die im Deutschen häufig syntaktisch entfaltet wird. Die Lutherreferenzübersetzung zeigt in denselben Versen ebenfalls erhöhte Length-Ratio-Werte, jedoch meist unterhalb der Verdopplungsgrenze, was den notwendigen Grad der Explikation markiert, ohne in paraphrasierende Ausdehnungen überzugehen. Aber gerade auch die der hebräischen Grammatik zugrundliegende Personalsuffixe, verringert die im Quelltext verwendete Wortzahl pro Satz im Allgemeinen – gerade im Vergleich zu indogermanischen Sprachen, wie Deutsch und Englisch. Im Ansatz lässt sich hier ein Unterschied in der Implementierungsebene erkennen, da DeepL und Gemini eher der Referenzlinie folgen als die lokal ausgeführten Modelle NLLB und LLaMA-3.

Damit ergeben sich drei grundlegende Beobachtungen:

[1] Modellbasierte Zusammenhänge sind in narrativen Prosatexten erkennbar.

[2] Prophetische und poetische Texte sind mit cloudbasierten Systemen stabiler an der Referenzübersetzung.

[3] Poetische Bildsprache benötigt eine ausgeprägte Expansion.

Halluzinationsrate

Die Halluzinationsrate gibt Auskunft darüber, wie viele Wörter des Zieltextes keine zuordenbare Entsprechung im hebräischen Quelltext besitzen. Eine hohe Rate zeigt, dass das Modell Begriffe oder Phrasen einfügt, die nicht im Quelltext verankert sind (potenzielle Halluzinationen oder interpretative Zusätze). Die Grundlage dafür bieten die von SimAlign gefundene Anzahl der Wortpaare als korrespondierende Übersetzungseinheiten. Dazu lassen sich die unalignierten Wörter

(`src_unaligned_count`) im hebräischen Quelltext zählen, die keine Entsprechung im Deutschen Zieltext haben. Ein hoher Wert weist an dieser Stelle auf Auslassungen oder starke strukturelle Umstellungen hin. Als Grenzwerte für die Halluzinationsrate lassen sich folgende Zonen unterscheiden:

[1] 0 – 30 %: moderate Explikation

[2] 30 – 60 %: deutliche, aber strukturell erklärbare Erweiterung

[3] > 60 %: starke interpretative Ausdehnung

[4] > 75 %: potenzielle überschießende Paraphrasen oder Strukturaufweichungen

Exodus 3, 11–15 **Jesaja 7, 10–16** **Psalm 23, 1–6**

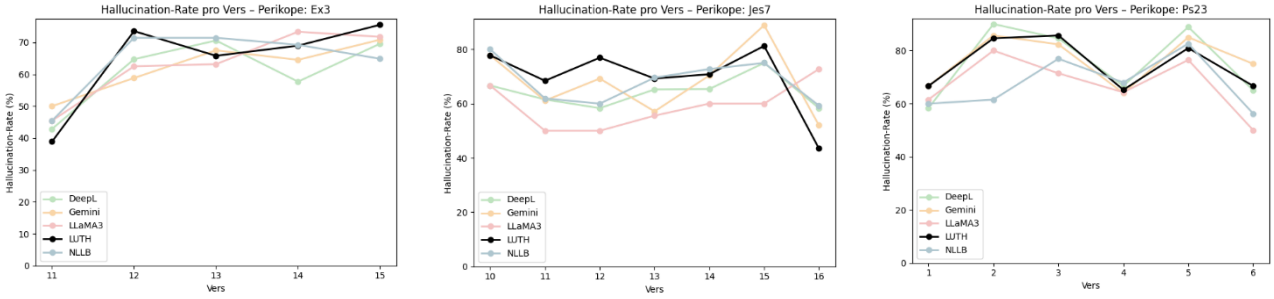


Abbildung 4-2 *Liniendiagramme – Halluzinationsrate der generierten Übersetzungen*

Bemerkenswert ist, dass die Lutherreferenzübersetzung in allen drei Perikopen meist über der dritten Zonenschwelle liegt und damit den hebräischen Quelltext stark interpretativ ausdehnt. So zeigt sich, dass eine Halluzination in Modellen, die zur Übersetzung von alttestamentlichen Texten verwendet werden, im technischen Sinn nicht zwingend eine Fehlerqualität markieren, sondern eine notwendige Folge interlingualer Strukturunterschiede. Besonders in poetischen und prophetischen Textformen sind Funktionswörter, syntaktische Verbindungen und kontextualisierende Ergänzungen unvermeidlich. Alle Modelle zeigen meist parallel zur Referenzlinie verlaufende Halluzinationswerte, die überwiegend unter diesem Wert liegen. Somit bieten die betrachteten Modelle hinsichtlich der errechneten Halluzination eine vermeintlich bessere Performance als die menschengemachte Referenzübersetzung. Erkennbar ist außerdem, dass einzelne Verse (z. B. Ps 23,4) eine nahezu deckungsgleiche Halluzinationsrate aufweisen, die auch die Lutherreferenzübersetzung einschließt. Werden für einen Erklärungsversuch die Übersetzungen (Anhang 3-9) betrachtet, zeigen sich klar divergierende Formulierungen und Übertragungen. Gerade die von NLLB erzeugte Übersetzung zeigt deutliche Schwächen. So wird klar, dass die Halluzinationsrate ohne entsprechende Kontextualisierung kein eindeutiges Qualitätsmerkmal für die Übersetzungen ist. Für eine Einordnung zeigt die nachfolgende Abbildung den Vergleich der Halluzinationsraten zwischen den englischen und deutschen Referenzübersetzungen – Lutherübersetzung und Basis-Bibel sowie King James Version und New Living Translation.

Exodus 3, 11–15 **Jesaja 7, 10–16** **Psalm 23, 1–6**

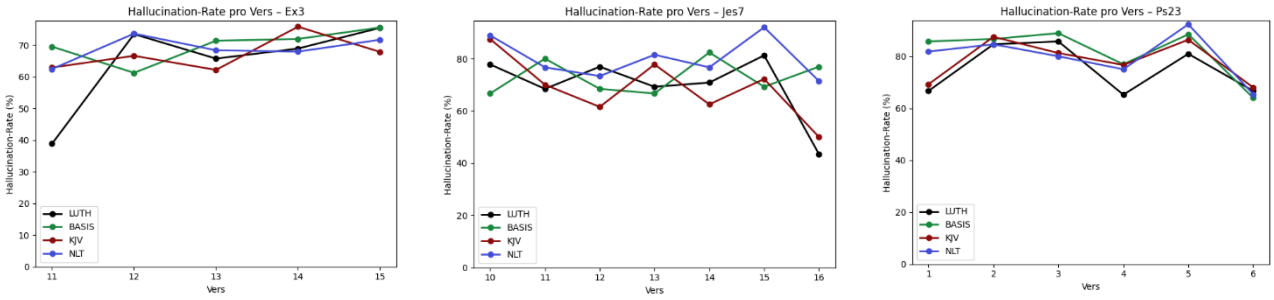


Abbildung 4-3 *Liniendiagramme – Halluzinationsrate der Referenzübersetzungen*

Dadurch wird klar, dass auch die menschengemachten Referenzübersetzungen variierende Halluzinationsraten aufweisen, wenngleich diese erkennbar weniger voneinander bzw. der Lutherübersetzung als Referenzpunkt divergieren. Dazu lässt sich erkennen, dass – anders als bei der Längenverhältnismetrik – nicht der Psalm die größte Varianz aufweist, sondern die prophetischen Verse aus Jesaja 7.

Drop-Rate

Die Drop-Rate zählt Wörter im deutschen Text, die keinem Wort des hebräischen Quelltext zugeordnet werden konnten (`dst_unaligned_count`). Diese deuten auf zusätzliche, im Ausgangstext nicht auffindbare Elemente hin.

Exodus 3, 11–15 **Jesaja 7, 10–16** **Psalm 23, 1–6**

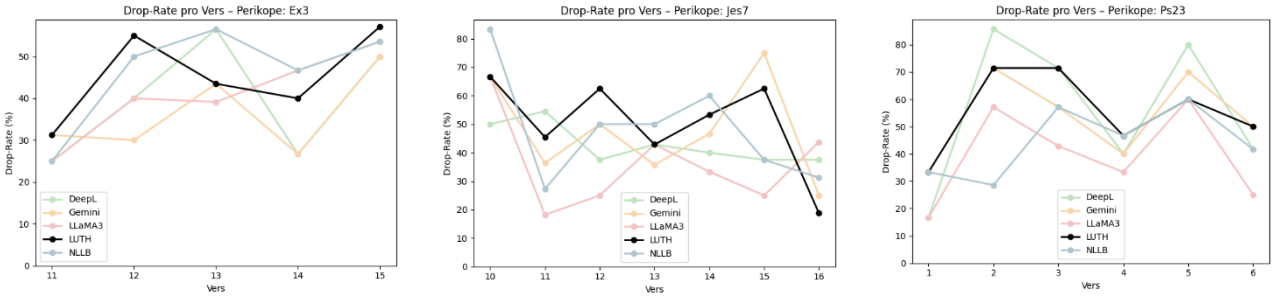


Abbildung 4-4 *Liniendiagramme – Drop-Rate-Metrik*

Auffällig ist, dass Versbereiche mit hoher Drop-Rate nicht zwingend jene mit hoher Halluzinationsrate sind. Während Halluzinationen oft interpretative Ergänzungen anzeigen, markieren erhöhte Fehlraten eher strukturelle Anpassungen und Verdichtungen. Die Diagramme helfen, diese Beziehung visuell zu differenzieren: In Psalm 23 sind beispielsweise Stellen erkennbar, die eine sehr niedrige Fehlrate darstellen, aber die Halluzination dagegen deutlich erhöht ist.

Ergänzend muss erwähnt werden, dass SimAlign, das für das Zählen der unalignierten Wörter eingesetzt wurde, mit modernem Hebräisch trainiert wurde. Auch wenn viele Wortwurzeln aus dem Althebräischen ins moderne Ivrith überführt wurden, kann es zu verfehlten Verbindungen durch SimAlign kommen. Das grundsätzliche Verhalten zur Referenzübersetzung sollte davon unberührt bleiben.

Token Fragmentation Rate (TFR)

Ergänzend zu den satz- und alignmentbasierten Metriken wurde für NLLB und LLaMA-3 die interne Segmentierung der Zieltexte untersucht. Grundlage ist die in Kapitel 3.3.1 beschriebene Subword-Segmentierung der Modelle. Die TFR misst dabei das Verhältnis von Subtoken- zu Wortanzahl und gibt an, in welchem Ausmaß ein Modell die Oberfläche der Zielsprache in kleinere Einheiten aufspaltet. Werte nahe bei 1.0 deuten auf eine weitgehend wortweise Segmentierung hin, während höhere Werte auf eine stärkere Fragmentierung und damit auf eine feinere, aber potenziell weniger stabile interne Repräsentation schließen lassen.

In den drei untersuchten Perikopen liegen die TFR-Werte insgesamt in einem moderaten Bereich, zeigen aber klare systematische Unterschiede zwischen Textgattungen und Modellen. Für die narrative Perikope Exodus 3 bewegen sich die Werte für NLLB überwiegend zwischen 1.32 und 1.54 (Mittelwert ca. 1.40), während LLaMA-3 hier leicht höhere Werte zwischen 1.38 und 1.61 (Mittelwert ca. 1.54) aufweist. Dies spricht dafür, dass beide Modelle den regulären narrativen Zieltext grob segmentieren und nur dort feiner auflösen, wo komplexere Wortformen oder seltene Ausdrücke auftreten. In der prophetischen Perikope Jesaja 7 steigen die TFR-Werte beider Modelle an und liegen typischerweise im Bereich zwischen 1.36 und 1.70. Hier nähern sich die Mittelwerte von NLLB (ca. 1.50) und LLaMA-3 (ca. 1.49) deutlich an. Die stärkere Fragmentierung spiegelt die höhere morphologische und syntaktische Komplexität der Zieltexte wider, insbesondere dort, wo hebräische Verdichtungen in deutsche Nebensätze, Modalstrukturen oder explizite Kausalverhältnisse aufgelöst werden. Am deutlichsten treten die Unterschiede in der poetischen Perikope Psalm 23 hervor. NLLB bewegt sich in einem Bereich von etwa 1.46 bis 1.87 (Mittelwert ca. 1.56), LLaMA-3 erreicht Werte zwischen 1.46 und knapp 1.93 (Mittelwert ca. 1.69). Die poetischen Zieltexte werden damit durchgängig stärker fragmentiert als die narrativen und prophetischen, was darauf hinweist, dass die Modelltokenizer in diesem Kontext häufiger auf seltene Lexeme, idiomatische Wendungen und syntaktisch ungewöhnliche Konstruktionen treffen.

Im Vergleich zur Lutherübersetzung ist hervorzuheben, dass die TFR keine direkte Entsprechung im Referenztext besitzt, sondern eine rein modellinterne Größe ist. Dennoch erlaubt sie eine indirekte Bezugnahme: Dort, wo die Lutherübersetzung im Zieltext eine einfache, morphologisch wenig komplexe deutsche Form verwendet, die Modelle aber eine hohe Fragmentierung aufweisen, ist von einem Spannungsfeld zwischen orthographischer Glätte und interner Verarbeitungsunsicherheit auszugehen. Umgekehrt deuten niedrige TFR-Werte in Kontexten, in denen die Referenzübersetzung bereits eine stark explizierte und syntaktisch ausformulierte Struktur bietet, darauf hin, dass die Modelle diese Zieltexte intern relativ stabil repräsentieren. Die TFR fungiert damit als ergänzender Indikator, der die in den anderen Metriken sichtbar werdenden Unterschiede – insbesondere zwischen narrativen, prophetischen und poetischen Texten – auf der Ebene der Modellarchitektur bestätigt.

4.2.2 Unterschiede in den Systemen

Die Analysen aus Kapitel 4.2.1 lassen sich durch die Betrachtung von Durchschnittswerten verdichten und damit als systematische Profile der Modelle lesbar machen. Für alle drei Perikopen wurden die Metriken Längenverhältnis, Halluzinationsrate, Drop-Rate sowie – für NLLB und LLaMA-3 – die TFR zu Mittelwerten pro System zusammengefasst. Diese Aggregation zeigt, dass die beobachteten Unterschiede zwischen den Modellen nicht nur punktuell in einzelnen Versen auftreten, sondern sich konsistent über den gesamten Korpus hinweg durchhalten.

Auf Ebene des **Längenverhältnisses** liegen die Modelle insgesamt in einem relativ engen Bereich, bilden aber dennoch deutlich unterscheidbare Profile aus. Über alle 18 Verse hinweg erreicht NLLB eine durchschnittliche Length Ratio von rund 1,65, DeepL liegt mit etwa 1,67 nur knapp darüber, LLaMA-3 kommt auf etwa 1,72, während Gemini mit im Mittel 1,81 die stärkste Expansion des Zieltexts aufweist. Die Lutherübersetzung bewegt sich mit einem durchschnittlichen Längenverhältnis von etwa 1,72 nahezu auf demselben Niveau wie LLaMA-3. Praktisch bedeutet dies: Während NLLB und DeepL die hebräischen Ausgangstexte im Deutschen im Schnitt um etwa zwei Drittel verlängern, steigt diese Expansion bei Gemini auf knapp 80 %. Gemini produziert damit klar die passendsten Zieltexte, während die NMT-Systeme und LLaMA-3 näher am Längenprofil der Lutherbibel bleiben.

Noch aussagekräftiger wird das Bild durch die gemittelten **Halluzinationsraten**. Hier liegen DeepL und NLLB mit durchschnittlich etwa 67 % sehr dicht beieinander, Gemini erreicht mit rund 67,6 % den höchsten Wert, während LLaMA-3 mit etwa 63 % deutlich darunter bleibt. Die Lutherübersetzung weist mit einer mittleren Halluzinationsrate von ungefähr 70 % sogar den höchsten Wert im Vergleich auf. Zu beachten ist, dass diese Raten nicht Fehlerquoten im klassischen Sinn darstellen, sondern anzeigen, welcher Anteil der Zielwörter im Alignment keine direkte Zuordnung zu einem Quellwort erhält. Werte im Bereich von 65–70 % sind bei der Übertragung von Hebräisch ins Deutsche – insbesondere in prophetischen und poetischen Texten – strukturell erwartbar, da Kopulaverben eingefügt, Partikeln aufgelöst und implizite Relationen explizit gemacht werden müssen. Gleichwohl zeigt der Vergleich, dass LLaMA-3 im Mittel am wenigsten zusätzliche Wörter produziert, während Gemini und die Lutherübersetzung sich am oberen Rand der Explikation bewegen. Dass Luther hier höhere Werte erreicht als die MT-Systeme, unterstreicht, dass auch tradierte Bibelübersetzungen – nicht nur LLMs – in hohem Maß interpretierend und strukturauflösend arbeiten.

Die **Drop-Rate** ergänzt dieses Bild aus einer komplementären Perspektive. NLLB, DeepL und Gemini liegen im Durchschnitt nahezu gleichauf und lassen jeweils rund 46–47 % der hebräischen Wortformen im Alignment fallen. LLaMA-3 sticht hier mit einer mittleren Drop-Rate von nur etwa 39 % deutlich hervor, während Luther mit rund 51 % den höchsten Auslassungsanteil zeigt. Eine Drop-Rate in der Größenordnung von 50 % bedeutet, dass etwa jedes zweite Quellwort nicht als eigenständige lexikalische Einheit im Zieltext sichtbar wird – etwa, weil mehrere hebräische Ausdrücke zu einer deutschen Form zusammengezogen, Pronomina in Flexionsendungen absorbiert

oder elliptische Strukturen konventionalisiert werden. Im Vergleich dazu zeigt das niedrigere Drop-Profil von LLaMA-3, dass dieses Modell im Mittel mehr hebräische Lexik – wenn auch in transformierter Form – in der deutschen Übersetzung sichtbar belässt als die übrigen Systeme und sogar als Luther. Umgekehrt kombiniert Gemini eine hohe mittlere Längenexpansion mit Dropraten auf dem Niveau der NMT-Systeme, was darauf hinweist, dass hier nicht primär „alles“ aus dem Ausgangstext explizit gemacht, sondern der Text zugleich um zusätzliche Elemente ergänzt wird.
Die Aufschlüsselung nach Textgattungen bestätigt diese Tendenzen. In der narrativen Perikope Ex 3,11–15 liegen die mittleren Längenverhältnisse für DeepL, NLLB, LLaMA-3 und Gemini noch eng beieinander (ca. 1,54–1,72). Die Halluzinationsraten bewegen sich hier meist um 60–65 %, die Dropraten überwiegend zwischen 36 % und 46 %. In Jes 7,10–16 verschieben sich die Mittelwerte insbesondere bei NLLB und Gemini nach oben. In Ps 23,1–6 steigen die durchschnittlichen Halluzinationsraten bei DeepL und Gemini auf etwa 75–76 %, die Dropraten auf rund 54–56 %. Luther zeigt über die drei Perikopen hinweg einen ähnlichen Verlauf: Sowohl Halluzinations- als auch Dropraten nehmen von der narrativen über die prophetische bis hin zur poetischen Perikope deutlich zu und erreichen in Psalm 23 mit etwa 75 % (Halluzination) und 55 % (Drop) ihre höchsten Werte. Poetische Texte erweisen sich damit für alle Systeme – einschließlich der menschlichen Referenzübersetzung – als metrisch „riskanter“, da sie strukturell stärker expliziert und zugleich stärker restrukturiert werden.
Die **Token Fragmentation Rate**, die nur für NLLB und LLaMA-3 vorliegt, erlaubt schließlich einen Blick auf die Segmentierungsebene der Modelle. Über alle Perikopen hinweg beträgt die mittlere TFR für NLLB etwa 1,49, für LLaMA-3 rund 1,57; in Psalm 23 steigen die Werte auf ca. 1,56 (NLLB) bzw. 1,69 (LLaMA-3) an. Das bedeutet, dass beide Systeme die Zieltexte im poetischen Kontext in deutlich mehr Subtoken zerlegen als in der narrativen Perikope, LLaMA-3 dabei insgesamt etwas stärker als NLLB. In Verbindung mit der niedrigen Drop-Rate legt dies nahe, dass LLaMA-3 poetische Strukturen intern zwar als komplexer und segmentierungsbedürftiger verarbeitet, gleichzeitig aber bemüht ist, mehr Quelllexik in den Zieltext zu integrieren.

In der Gesamtheit zeichnen die Durchschnittswerte damit ein konsistentes Systembild:

[1] [2] DeepL und NLLB bilden im Durchschnitt ein ähnliches Profil mit moderaten Längenexpansionen und Halluzinations- bzw. Dropraten, die nahe beieinander liegen und sich in der Größenordnung klassischer Bibelübersetzungen bewegen.

[3] LLaMA-3 unterscheidet sich durch etwas höhere Satzlängen, aber zugleich geringere mittlere Halluzinations- und Drop-Werte und eine erhöhte TFR – es agiert im Vergleich semantisch flexibel, aber formnäher.

[4] Gemini markiert den Extrempol mit der stärksten Expansion, den höchsten Halluzinationsraten und Dropraten auf NMT-Niveau und entfernt sich damit strukturell am deutlichsten von der Lutherreferenz.

4.3 Qualitative, theologische Analyse

Die quantitative Auswertung bietet zunächst ein belastbares Orientierungsraster, indem sie Referenznähe, Stabilität und mögliche Ausreißer versweise sichtbar macht. Für biblische Texte ist damit jedoch nur ein Teil dessen erfasst, was Übersetzungsqualität im engeren Sinn ausmacht: Theologisch relevante Bedeutungsnuancen entstehen nicht allein durch semantische Grundäquivalenz, sondern durch die Wahl traditionsgeprägter Schlüsselbegriffe, die Einbindung in größere Sinnzusammenhänge sowie durch hermeneutische Entscheidungen, die Bildsprache, Sprecherrollen und implizite Deutungsrahmen mitprägen. Aus diesem Grund folgt auf die metrische Ebene eine qualitative, theologisch informierte Analyse, die die Übersetzungen an inhaltlich und wirkungsgeschichtlich sensiblen Punkten prüft und damit die Grenzen rein referenzorientierter Kennzahlen methodisch auffängt. Kapitel 4.3 entfaltet diese qualitative Perspektive in zwei komplementären Schritten: In 4.3.1 werden zentrale Lexeme und Formulierungen untersucht, an denen sich Übersetzungsentscheidungen besonders verdichten (z. B. bei Gottesbezeichnungen, relationalen Termini oder metaphorisch aufgeladenen Ausdrücken), einschließlich ihrer möglichen Bedeutungsfelder und der Anschlussfähigkeit an etablierte Übersetzungstraditionen. Darauf aufbauend erweitert 4.3.2 den Blick von der Einzelstelle auf den jeweiligen Textzusammenhang: Hier wird geprüft, wie konsistent die Modelle über Verse hinweg argumentieren bzw. erzählen, wie sie syntaktische und literarische Strukturen (Parallelismen, Ellipsen, Bildketten) interpretativ auflösen oder bewahren und welche theologischen Akzentverschiebungen sich daraus im Gesamtprofil ergeben. Zusammen ermöglichen beide Unterkapitel eine Auswertung, die nicht bei Abweichungen vom Referenzwortlaut stehen bleibt, sondern die Übersetzungen in ihrer theologischen Funktion und hermeneutischen Plausibilität innerhalb der Perikope beurteilt.

4.3.1 Detailanalyse ausgewählter Schlüsselbegriffe

Formale und lexikalische Stabilität

Auf der ersten Ebene werden zunächst Oberflächenphänomene, Tokenisierungsfehler, orthographische Auffälligkeiten sowie lexikalische Wahlentscheidungen sichtbar gemacht. Diese Ebene ist methodisch bewusst einfach, weil bereits kleine formale Instabilitäten in biblischen Texten – etwa bei Eigennamen, Gottesbezeichnungen oder festen liturgischen Formeln – erhebliche Anschlussprobleme für die weitere Interpretation erzeugen können. Gerade im narrativen Text von Ex 3,11–15 zeigt sich zunächst eine hohe Stabilität der Systeme, zugleich treten aber typische MT-Artefakte auf, die nicht als inhaltliche Fehlinterpretation, sondern als technische Segmentierungs- bzw. Generierungsprobleme zu lesen sind: DeepL erzeugt beispielsweise Zusammenschreibungen wie „*zumose*“ bzw. im Englischen „*tomoses*“ (Ex 3,14), was im Kontext des hebräischen Maqqef zwar erklärbar erscheint, aber in der Zielsprache die Lesbarkeit und philologische Anschlussfähigkeit deutlich reduziert. Solche Fehler sind nicht nur stilistische Kleinigkeiten, sondern markieren die Grenze einer rein statistischen Wort-/Subwortverarbeitung dort, wo Texttradition gerade durch standardisierte Namensformen Stabilität erzeugt.

Die im Anhang 8-1 ausgewiesenen aggregierten Werte – je Perikope, Sprache und Modell – verdichten die versweisen Einzelmessungen zu zwei interpretierbaren Ebenen: Der jeweilige Durchschnitt beschreibt die mittlere Referenznähe des Modells innerhalb der Perikope, während Varianzen als Stabilitätsindikator zu lesen sind. Methodisch bildet chrF primär Wortlaut-/Formnähe ab, während BERTScore stärker semantische Nähe zur Referenz approximiert und dadurch auch bei Paraphrasen oder Synonymik relativ hoch bleiben kann (Popović, 2015; Zhang et al., 2020). Gleichzeitig gilt: Beide Maße sind Referenzmetriken – sie messen Nähe zu Luther/Basis bzw. KJV/NLT, nicht direkt Treue zum Hebräischen; entsprechend sind die Ergebnisse als quantitativer Orientierungsrahmen zu verstehen, der anschließend durch die Ebenen zwei bis vier (syntaktisch-semantisch, literarisch-theologisch, intertextuell/traditionsbezogen) inhaltlich eingeordnet werden muss.

Für die narrative Perikope **Ex 3,11–15** zeigen die Mittelwerte insgesamt ein sehr hohes und zugleich stabiles Niveau: In beiden Zielsprachen liegen die Modelle überwiegend im Bereich hoher chrF- und sehr hoher BERTScore-Werte, was auf eine robuste Verarbeitung narrativer Prosa hindeutet. Unterschiede zwischen den Systemen sind hier eher graduell und betreffen vor allem die Frage, wie eng an tradierte Formulierungen angelehnt wird. Gleichzeitig macht die Streuung sichtbar, dass einzelne Modelle in der englischen Ausgabe stärker zwischen den Versen schwanken können – ein Hinweis darauf, dass Referenznähe nicht automatisch gleichmäßig über die Perikope verteilt ist, obwohl der Gesamteindruck sehr stabil bleibt.

In **Jes 7,10–16** (Prophetie) divergieren die Werte deutlich stärker. Gerade hier wird das Zusammenspiel beider Metriken interpretativ hilfreich: Einige Modelle bleiben semantisch nah an den Referenzen (hoher BERTScore), weichen aber formalseitig stärker ab (niedriger chrF) – ein typisches Profil für paraphrasierende oder registerglättende Lösungen. Umgekehrt zeigen sich auch Konstellationen, in denen die formale Nähe hoch wirkt, ohne dass die semantische Annäherung entsprechend mitzieht. Zusammen mit teils deutlich erhöhter Streuung spricht dies dafür, dass Prophetie als Textsorte häufiger zu verseweisen Instabilitäten führt (z. B. durch Ellipsen, Deixis, traditionsgeladene Lexeme), wodurch sich kritische Verse für die folgenden Ebene gut identifizieren lassen.

Am deutlichsten treten gattungsspezifische Grenzen in **Ps 23,1–6** (Poesie) hervor. Hier fällt weniger der absolute Mittelwert als vielmehr die Streuung als zentrales Signal ins Gewicht: Die Werte schwanken bei mehreren Modellen deutlich, was zu der Erwartung passt, dass poetische Bildsprache, Parallelismen und metaphorische Verdichtung nicht gleichmäßig übersetzungsstabil sind. Semantisch bleiben die Modelle zwar insgesamt weiterhin nah an den Referenzen, doch die Kombination aus geringerer formaler Nähe und erhöhter Streuung spricht dafür, dass gerade in einzelnen Versen die poetische Logik entweder paraphrasiert, verengt oder stilistisch verschoben wird. Für die qualitative Auswertung ist damit Psalm 23 besonders geeignet, um anhand weniger, gezielt ausgewählter Verse zu zeigen, wie die Modelle mit metaphorischen und traditionsprägenden Formeln umgehen.

Insgesamt lässt sich aus den aggregierten Ergebnissen ableiten, dass die Modelle in narrativen Texten am konsistentesten referenznah arbeiten, während Prophetie und insbesondere Poesie stärker zu heterogenen Leistungsprofilen führen. Die Metriken liefern damit keine abschließende Qualitätsbewertung, wohl aber ein belastbares Raster, um die anschließende textnahe Analyse zu fokussieren: Hohe Streuung markiert Interpretationsbedarf (Ausreißer), chrF vs. BERTScore zeigt, ob Abweichungen eher stilistisch/phraseologisch oder semantisch ins Gewicht fallen (Popović, 2015; Zhang et al., 2020).

Syntaktisch-semantische Kohärenz

Die zweite Ebene prüft, ob die Modelle die semantische Logik des Ausgangstextes unter den Bedingungen hebräischer Syntax (etwa Verbalsatzketten, Nominalsätze, Ellipsen und Parallelismen) in der Zielsprache kohärent abbilden. Gerade bei Exodus 3 zeigt sich, dass die scheinbar leichte Narration zwar insgesamt stabiler übersetzt wird, aber dennoch an neuralen Systemgrenzen scheitern kann, sobald Interpunktion fehlt oder syntaktische Einheiten nicht eindeutig segmentiert sind. Die Doppelung „*dasdas*“ in Ex 3,12 (bzw. „*thisthis*“ im Englischen) ist inhaltlich trivial, semantisch jedoch nicht neutral: Sie bricht den Lesefluss und markiert eine generative Instabilität, die in der MT-Forschung als Form von Halluzination bzw. Fehlgenerierung diskutiert wird (Raunak et al., 2021). In der prophetischen Perikope Jes 7,10–16 verschärft sich diese Ebene deutlich: Prophetische Sprache ist dichter, elliptischer und häufig bildhaft verdichtet. Damit steigen die Risiken semantischer Drifts, falscher Referenzbindungen und syntaktischer Glättung. In den genenrierten Übersetzungen tritt dies u. a. darin hervor, dass einzelne Modelle Personenbezüge oder Orts- und Namensformen verwechseln (z. B. fehlerhafte Eigennamen-Übertragungen) und dadurch die semantische Kette beschädigen. Auf dieser Ebene wird besonders sichtbar, dass eine metrisch ordentliche Ausgabe dennoch theologisch problematisch sein kann, wenn z. B. Kausalbeziehungen oder Zweck- und Folgestrukturen umkippen. Die zweite Ebene dient daher als Brücke zwischen rein formalen Auffälligkeiten (Ebene 1) und den tieferen literarischen bzw. theologischen Implikationen (Ebene 3).

Literarisch-theologische Dimension

Die dritte Ebene fokussiert die Frage, ob die Modelle literarische Strukturen und theologische Kernsemantik bewahren oder nivellieren. In Psalm 23 ist die Herausforderung besonders offensichtlich: Der Text arbeitet mit dichter Metaphorik („Hirte“), kulturspezifischen Wendungen und parallelen Bildketten, die nicht nur ästhetische Oberfläche sind, sondern Gottesbilder und anthropologische Grundannahmen transportieren. Gerade hier zeigen die Übersetzungen deutliche Instabilität: Schon Ps 23,1 enthält bei einzelnen Modellen unnatürliche oder semantisch verschobene Lösungen („*ich werde nicht hungernd sein*“) statt einer idiomatisch etablierten Form („*mir wird nichts mangeln*“/„*I shall not want*“). Diese Verschiebung ist mehr als Stil: Sie verengt die metaphorische Breite des hebräischen Mangels-/Fehlfeldes auf rein körperliche Bedürftigkeit und verändert damit die theologische Bildlogik des Psalms. Auch Ps 23,5 zeigt, wie schnell poetische Struktur in semantische Entgleisung kippen kann: Einzelne Ausgaben geraten in grobe Bildstörungen („*schö-*

ner Schwanz“, „*Schnauze … Schmerzen*“) oder verändern die Rollenverhältnisse (Opfer- und Feindmotive), was die intendierte Trost- und Schutzmetaphorik des Psalms unterläuft. An dieser Stelle wird besonders deutlich, warum die qualitative Ebene drei nicht durch Metriken substituiert werden kann: Selbst wenn eine Ausgabe einzelne Schlüsselwörter trifft, kann sie durch kleine metaphorische Verschiebungen ein anderes Gottesbild erzeugen. Genau hier liegt der hermeneutische Mehrwert dieser Analyseebene: Sie bewertet nicht Textähnlichkeit, sondern theologische Funktionalität im Zieltext.

Intertextuelle Kohärenz und Anschlussfähigkeit an die Übersetzungstradition

Die vierte Ebene untersucht schließlich, ob die maschinellen Übersetzungen stärker an traditionelle Formulierungen (z. B. Lutherbibel/KJV) oder an moderne Referenzen (BasisBibel/NLT) anschließen – oder ob sie eigenständige, nicht-tradierte Lesarten produzieren. Diese Ebene folgt der Einsicht, dass theologische Bedeutung häufig wirkungsgeschichtlich erschlossen wird und Übersetzungen nicht im leeren Raum stehen, sondern in Traditionslinien, die bestimmte Termini und Formeln stabilisieren (Becker, 2021). Im narrativen Kontext von Exodus zeigt sich diese Traditionsdimension besonders am Gottesnamen und an feststehenden Sprechformeln: Während Referenzübersetzungen hier teils bewusst tradierte Lösungen wählen („*Ich bin, der ich bin*“/„*I AM THAT I AM*“), erzeugen maschinelle Systeme teils hybride oder semantisch entkernte Varianten („*Ich werde euch senden*“ u. Ä.), wodurch die klassische Offenbarungsformel ihre identitätsstiftende Funktion verliert. In Jesaja 7 verstärkt sich die Traditionsproblematik, weil der Text selbst in der Rezeptionsgeschichte stark aufgeladen ist (u. a. an Schlüsselbegriffen wie עַלְמָה). Wo etablierte Übersetzungen häufig bewusst eine Linie markieren, spiegeln die Modelle eher die statistische Uneinheitlichkeit ihrer Trainingsdaten – ohne hermeneutische Differenzierung. Methodisch ist diese vierte Ebene zugleich diejenige, die am stärksten die Grenzen automatischer Evaluation aufzeigt. Mathur et al. (2020) betonen, dass Metriken insbesondere bei stilistisch-semantischen Qualitätsdimensionen und bei „meaning beyond the sentence“ systematisch unterbestimmt bleiben. Genau dieser Bereich ist für biblische Texte jedoch zentral. Die Traditionsanalyse dient daher nicht der Feststellung einer richtigen Übersetzung, sondern der Kontextualisierung: Sie macht sichtbar, ob ein Modell – bewusst oder unbewusst – tradierte Sprachformen reproduziert (z. B. archaische Register wie „*daß*“, „*shalt*“) oder moderne Register glättet und dadurch andere Rezeptionssignale setzt.

4.3.2 Kontextuelle und hermeneutische Reflexion

Die Detailanalyse einzelner Schlüsselbegriffe zeigt, wo Übersetzungsentscheidungen besonders sensibel sind. Sie erklärt jedoch noch nicht hinreichend, wie diese Entscheidungen im fortlaufenden Textzusammenhang wirken. Deshalb richtet die kontextuelle hermeneutische Reflexion den Blick auf die Perikope als kohärente Sinneinheit: Übersetzt wird nicht nur Wort für Wort, sondern immer auch ein Deutungszusammenhang, der Sprecherrollen, implizite Logik, Bildfelder und tradierte Formulierungsräume umfasst. Für die folgende Auswertung ist daher entscheidend, ob die Modelle nicht nur semantisch nahe an Referenzen bleiben, sondern ob sie die innere Textdynamik stabil abbilden – oder ob sich an Übergängen, Pronomina, Deixis, Metaphern und Registerentscheidungen kleine Verschiebungen ergeben, die hermeneutisch relevante Akzentverlagerungen erzeugen. Als Leitfragen werden dabei drei Ebenen unterschieden: Erstens die *diskursive Kohärenz* (Wer spricht? Zu wem? Mit welcher illokutionären Funktion – Zusage, Auftrag, Drohung, Trost?), zweitens die *konzeptuelle Kohärenz* (bleiben zentrale Begriffsfelder und Relationen über Verse hinweg stabil oder werden sie durch Synonymwechsel, Explikationen oder Auslassungen umgesteuert?), und drittens die *ästhetisch-theologische Kohärenz* (wird die literarische Gestalt – Parallelismus, Bildketten, Rhythmus, formelhafte Sprache – so bewahrt, dass die theologische Funktion des Textes im Zieltext plausibel bleibt?). Die quantitativen Streuungswerte aus Kapitel 4.2 dienen hierbei als pragmatischer Hinweis darauf, an welchen Versen ein Modell besonders instabil ist. Die hermeneutische Prüfung entscheidet jedoch, ob diese Instabilität lediglich stilistische Variation oder eine inhaltlich relevante Verschiebung darstellt.

Ein erstes, eher technisch anmutendes, hermeneutisch aber relevantes Problem zeigen die DeepL-Ausgaben in Ex 3,13: Durch Formatierungsartefakte wie „*zuund“*, „*untothe*“ oder „*adliberis*“ wird der Textfluss an einer inhaltlich zentralen Stelle (Moses’ Frage nach dem Gottesnamen) gestört. Solche Brüche sind nicht nur kosmetisch: In narrativen Texten funktionieren Frage-Antwort-Sequenzen als theologische Dramaturgie. Wenn ausgerechnet die Hinführung zur Namensoffenbarung sprachlich verknotet erscheint, wird die Passage weniger als autoritativer Dialog wahrgenommen, sondern wirkt wie ein fehlerhafter Rohtext. Ähnliche Dopplungen („*dasdas*“, „*thisthis*“) treten bereits in Ex 3,12 auf und markieren, dass formale Störungen punktuell die Lesbarkeit und damit die Auslegungssicherheit beeinträchtigen können. In Ex 3,15 verschiebt sich die hermeneutische Perspektive weniger durch offensichtliche Fehler, sondern durch Traditionssignale in der Gottesnamenswiedergabe. Während NLLB im Englischen „*Jehovah*“ setzt, arbeitet Gemini mit „*YHWH“*, und LLaMA-3 bleibt bei „*the Lord God*“. Damit werden unterschiedliche Rezeptionsräume aktiviert: „Jehovah“ evoziert (besonders im Englischen) spezifische kirchen- und übersetzungsgeschichtliche Kontexte; „YHWH“ wirkt philologisch-transliterationstreu; „Lord“ schließt stärker an liturgisch etablierte Ersetzungstraditionen an. In einer hermeneutischen Lesart ist dies entscheidend, weil Exodus 3 nicht nur Information über Gottes Namen gibt, sondern zugleich Identität stiftet („…mein Name für immer…“). Wenn dann zusätzlich in derselben DeepL-Ausgabe Fehler wie „*thethe*“ oder fehlende Wortgrenzen auftreten, entsteht ein Spannungsfeld zwischen semantisch

passabler Traditionsnähe und textkritisch irritierender Oberflächenqualität. Deutlicher wird das Problem der kontextuellen Kohärenz in der prophetischen Perikope. In Jes 7,11 (Aufforderung zur Zeichenbitte „in der Tiefe oder in der Höhe“) zeigen die Modelle nicht nur stilistische Unterschiede, sondern teils eine inhaltliche Entgleisung: NLLB wiederholt im Englischen auffällig oft „*deeply, deeply, or deeply*“, wodurch die zweipolige Struktur („Tiefe/Höhe“) faktisch kollabiert. LLaMA-3 verschiebt den Sinn noch stärker, indem aus der Zeichenbitte eine Suche „aus der Mitte deiner Mächtigen“ und nach „kleinem oder großem“ wird. Hermeneutisch ist das gravierend: Der Text intendiert eine rhetorische Öffnung des Erwartungshorizonts („egal wie unmöglich – Gott kann es geben“). Die Modelle erzeugen stattdessen entweder semantische Redundanz oder eine neue, kaum anschlussfähige Bildwelt. Gemini bleibt hier nah an der Struktur und verstärkt sogar die Parallelität („make the request deep … or … high above“). Ein zweites Beispiel aus Jesaja, das selten im Mittelpunkt steht, aber kulturhermeneutisch besonders ergiebig ist, findet sich in Jes 7,15 (וּדְבַשׁ תֶּאֱכָֽה). Lutherübersetzung und BasisBibel führen „*Butter und Honig*“, während Gemini „*Dickmilch und Honig*“ wählt; NLLB paraphrasiert auffällig vage („*grünes und weiches*“) und LLaMA-3 ergänzt sogar „Sirup“. Gerade solche Details steuern, ob der Vers als alltägliches Ernährungssignal, als kulturell spezifisches Milchprodukt oder als überdehnte Wohlstandsmetapher gelesen wird. Die hermeneutische Pointe bleibt zwar formal erhalten, doch die Konnotationen des Szenarios variieren deutlich – und damit auch die Anschlussfähigkeit an bekannte Auslegungsmuster. Schließlich zeigt Jes 7,16, wie stark einzelne Modellformulierungen die theologische Stoßrichtung verändern können. NLLB erzeugt im Deutschen die irritierende Wendung „*vor dem du dich vernichtest*“, während LLaMA-3 wiederholt und moralisiert *(„die Erde, die du verflucht hast…“*). Damit kippt die Aussage leicht von einer politischen Ankündigung („das Land … wird verödet/verlassen sein“) in eine psychologisierende bzw. moralisch aufgeladene Lesart. DeepL bleibt semantisch näher, verändert aber ebenfalls die Konstruktion („*Land, das du … gefürchtet hast*“). Für die hermeneutische Einordnung ist das relevant, weil der Text in seinem Kontext (Ahaz, Bedrohung durch zwei Könige) auf historische Realpolitik zielt. Eine Verschiebung hin zu „Fluch“- oder Selbstvernichtungs-Semantik kann die theologische Aussage ungewollt umkodieren.

Die poetische Perikope Psalm 23 macht schließlich sichtbar, warum rein metrische Übereinstimmungswerte die literarisch-theologische Dimension nur begrenzt erfassen. Schon Ps 23,4 kann bei NLLB in der englischen Ausgabe in eine völlig andere Bildwelt abgleiten *(„midst of trouble… thou art my shield… deliverer“*), wodurch die zentralen Psalmmetaphern (*„rod and staff“*) faktisch ersetzt werden. In Ps 23,5 zeigt NLLB im Deutschen sogar eine groteske Entgleisung (*„schönen Schwanz… Schnauze… Schmerzen“*), die nicht nur stilistisch unbrauchbar ist, sondern theologisch den Text in eine unfreiwillig parodistische Richtung verschiebt. Gerade an solchen Stellen wird deutlich: Auch wenn ein Modell in anderen Versen stabil wirken kann, reichen einzelne Ausreißer aus, um die hermeneutische Verlässlichkeit einer Übersetzung für wissenschaftliche oder liturgische Anschlussverwendungen zu unterminieren.

5 Diskussion und Abschluss

An die Auswertung schließt eine Einordnung an, die die gewonnenen Befunde in einen größeren Zusammenhang stellt. Dabei zeigt sich, wie eng technische Eigenschaften der Modelle (etwa Stabilität, Expansion oder Fehleranfälligkeit) mit theologischen Fragen nach Ambiguität, gattungstypischer Struktur und dem eröffneten Deutungsraum verbunden sind. Auf dieser Grundlage werden die Modelle anschließend differenziert gegenübergestellt und im Hinblick auf ihre Eignung für verschiedene wissenschaftliche Arbeitsschritte diskutiert. Abschließend werden daraus Perspektiven für den Einsatz von KI in der theologischen Forschung entwickelt und die zentralen Konsequenzen für die Forschungsfrage gebündelt.

5.1 Einordnung der Ergebnisse in technischer und theologischer Perspektive

Die Ergebnisse aus Kapitel 4 verdeutlichen, dass Kontexttreue bei der maschinellen Übersetzung alttestamentlicher Texte nicht als einzelnes Qualitätsmerkmal messbar ist, sondern als mehrdimensionale Zielgröße verstanden werden muss. Die quantitative Analyse erfasst hierfür vor allem strukturelle Indikatoren: das Längenverhältnis als Maß für Expansion bzw. Explizitation, Halluzinations- und Drop-Rate als Hinweise auf nicht eindeutig zuordenbare Ergänzungen bzw. Restrukturierungen sowie – für NLLB und LLaMA-3 – die Token Fragmentation Rate als Fenster in die modellinterne Segmentierung. Die qualitative, theologische Analyse ergänzt diese Perspektive, indem sie dort ansetzt, wo sich Kontexttreue tatsächlich entscheidet: in der Behandlung semantischer Mehrdeutigkeit, im Erhalt gattungstypischer Funktionen (Narration, Prophetie, Poesie) sowie in der Anschlussfähigkeit an die Übersetzungstradition und deren Wirkungsgeschichte.

Aus technischer Sicht zeigen die verdichteten Durchschnittswerte zunächst ein stabiles Gesamtbild, das jedoch klar unterscheidbare Systemprofile ausprägt. DeepL und NLLB bilden im Mittel ein ähnliches Muster mit moderaten Expansionen und vergleichbaren Halluzinations- und Drop-Werten; diese bewegen sich – je nach Perikope – in einer Größenordnung, die nicht grundsätzlich außerhalb dessen liegt, was auch bei etablierten Bibelübersetzungen als strukturbedingte Explizitation beobachtbar ist. LLaMA-3 hebt sich als LLM dadurch ab, dass es zwar tendenziell etwas längere Zieltexte erzeugt, zugleich aber geringere mittlere Halluzinations- und Drop-Werte aufweist und eine erhöhte TFR zeigt: Gerade in der poetischen Perikope steigt die Fragmentierung stärker an, was als Hinweis auf erhöhte Verarbeitungskomplexität gelesen werden kann, ohne dass dies zwingend mit größerer inhaltlicher Entfernung einhergeht. Gemini wiederum markiert den Extrempol mit der stärksten Expansion und den höchsten Halluzinationswerten; strukturell entfernt es sich damit am deutlichsten von der Lutherreferenz. Entscheidend ist jedoch, dass die Metriken selbst bereits auf ihre Interpretationsgrenzen verweisen. Zum einen sind Halluzinations- und Drop-Rate nicht deckungsgleich, da hohe Drop-Werte eher Verdichtungen und strukturelle Umbauten anzeigen, während Halluzinationswerte häufiger mit interpretativer Explizitation korrelieren. Zum anderen zeigt der Vergleich mit Referenzübersetzungen, dass auch menschliche Übersetzungen gattungsabhängig stark variieren: Poetische Texte gelten metrisch für alle Systeme – einschließlich

der Referenzen – als riskanter, weil sie Bildsprache, Parallelismen und elliptische Strukturen systematisch zu Umformungen zwingen. Damit wird deutlich: Ein hoher metrischer Wert ist nicht automatisch ein Qualitätsdefizit, sondern kann schlicht anzeigen, dass die Zielsprachlichkeit eine stärkere Restrukturierung erfordert. Zusätzlich ist methodisch zu berücksichtigen, dass Alignment-basierte Messungen (z. B. über SimAlign) aus trainingsbedingten Gründen gerade bei historischer Sprache nur Näherungen liefern; relevant bleibt daher primär der Vergleich der relativen Profile zwischen Modellen und Referenzen. Genau an dieser Stelle wird die theologische Perspektive zur notwendigen zweiten Lesart der Daten. Übersetzungen – und erst recht LLM-Outputs – sind nicht bloße Übertragungen, sondern vollziehen interpretative Setzungen: Sie entscheiden implizit über Explizitheit, Register, Bildhaftigkeit und semantische Eindeutigkeit. Die qualitative Analyse macht sichtbar, dass sich Kontexttreue gerade in poetischen und prophetischen Texten nicht an Wortnähe festmachen lässt, sondern am Umgang mit Mehrdeutigkeit und gattungstypischer Verdichtung. Wo Modelle Ambiguitäten nivellieren, metaphorische Strukturen in erklärende Paraphrasen umformen oder theologisch aufgeladene Begriffe in Richtung moderner Gebrauchskonventionen verschieben, entsteht eine Deutung, die den Text zwar verständlicher machen kann, zugleich aber den Interpretationsraum verengt. Umgekehrt kann eine formal glatte und metrisch unauffällige Übersetzung hermeneutisch problematisch sein, wenn sie tradierte Lesarten reproduziert, ohne dass diese Entscheidung als solche sichtbar bleibt. Damit bestätigt sich der methodische Grundsatz, der in der Arbeit bereits angelegt ist: Metriken leisten eine notwendige, aber keine hinreichende Qualitätsdiagnose. Kontexttreue muss hermeneutisch rückgebunden werden – insbesondere über Schlüsselbegriffe, Gattungsbezug und Traditionsanschluss.

5.2 Stärken und Schwächen der unterschiedlichen Modelle

DeepL zeigt im Vergleich der untersuchten Systeme ein insgesamt konservatives und strukturell stabiles Übersetzungsprofil. In den quantitativen Auswertungen ordnet es sich – zusammen mit NLLB – eher in den Bereich moderater Expansionen und Werte ein, die nahe an klassischen Referenzübersetzungen liegen, was besonders in der narrativen Perikope (Ex 3,11–15) sichtbar wird. Als Stärke erweist sich dabei die hohe sprachliche Glätte: DeepL produziert in der Regel gut lesbare Zieltexte, die syntaktisch kohärent wirken und im Deutschen häufig einen übersetzungstypischen Ton treffen. Für eine erste Orientierung und als Ausgangspunkt für den Vergleich mit etablierten Bibelübersetzungen ist diese Stabilität hilfreich, weil sich Abweichungen oft klar als bewusste Modellentscheidung oder als gattungstypischer Effekt interpretieren lassen.
Eine zentrale Schwäche liegt jedoch in der geringeren Robustheit gegenüber hebräischer Interpunktionsarmut. In den Beobachtungen zeigt sich DeepL als einziges Modell, das bei Satzanfängen und -enden bzw. bei der Segmentierung ohne eindeutige Satzzeichen auffällig häufiger scheitert, während NLLB, Gemini und LLaMA-3 diese Übertragungsleistung konsistenter erbringen. Damit entsteht gerade bei textkritisch und hermeneutisch sensiblen Stellen ein Risiko: Wenn die Satzgrenzen nicht zuverlässig erkannt werden, wird der Zieltext nicht nur stilistisch, sondern auch semantisch schwer anschlussfähig. Hinzu kommt, dass DeepL als kommerzielles Cloud-System methodisch weniger transparent ist: Tokenisierung, interne Kontrollparameter und Modellupdates sind nicht reproduzierbar steuerbar, was die wissenschaftliche Nachvollziehbarkeit begrenzt. Insgesamt ist DeepL damit ein starkes Instrument für sprachlich saubere, eher klassische Übertragungen, zeigt aber Grenzen dort, wo der Quelltext formale Ambiguitäten (fehlende Interpunktion, elliptische Strukturen) bewusst offenlässt und diese Offenheit für die Exegese methodisch relevant bleibt.

NLLB-200 nimmt innerhalb der Untersuchung die Rolle eines offenen, forschungsnahen NMT-Systems ein: Es ist lokal ausführbar und erlaubt damit eine deutlich höhere Kontrolle über Sprachcodes, Tokenisierung und die generelle Reproduzierbarkeit von Ergebnissen. Diese Transparenz ist eine substanzielle Stärke im wissenschaftlichen Kontext, weil sich Übersetzungen nicht nur als Textprodukte, sondern als nachvollziehbare Ergebnisse eines klar definierten Settings dokumentieren lassen. Auch in den aggregierten Befunden bildet NLLB – gemeinsam mit DeepL – ein Profil mit moderaten Längenexpansionen und insgesamt ähnlichen Halluzinations-/Dropraten. Damit eignet sich NLLB grundsätzlich gut als Baseline-System, insbesondere dann, wenn die Forschungsfrage stärker auf Vergleichbarkeit und Wiederholbarkeit als auf stilistische Feinausarbeitung zielt.
Schwächen zeigen sich dort, wo theologisch bedeutsame Nuancen nicht nur übersetzt, sondern interpretativ balanciert werden müssen. Gerade bei poetischen und prophetischen Texten können NMT-Systeme dazu tendieren, entweder zu stark zu glätten (Ambiguität zu reduzieren) oder – umgekehrt – Formulierungen zu erzeugen, die zwar formal ausreichend sind, aber weniger idiomatisch wirken und dadurch hermeneutisch schwerer zu diskutieren sind. Dass die reinen Metriken

hier keine eindeutige Qualitätsaussage zulassen, wird in Interpretation der Halluzinationswerte deutlich: Selbst wenn Kennzahlen günstig erscheinen, kann die konkrete Formulierung im Zieltext theologisch problematische Verkürzungen oder Verschiebungen erzeugen. Zudem ist NLLB – trotz lokaler Kontrollierbarkeit – weiterhin ein datengetriebenes Übersetzungssystem, das Übersetzung als Satz-zu-Satz-Abbildung behandelt; kontextübergreifende Kohärenz (z. B. konsequente Terminologie über längere Abschnitte) lässt sich methodisch weniger gut führen als bei promptbaren LLMs. Insgesamt liegt die Stärke von NLLB damit in der methodischen Sauberkeit (lokal, reproduzierbar, gut als Vergleichsniveau), während die Schwäche vor allem in der begrenzten kontextuellen Steuerbarkeit und gelegentlich geringerer stilistischer Natürlichkeit liegt.

Gemini 2.5 Pro steht in der Modellkonstellation exemplarisch für den Paradigmenwechsel von spezialisierter Übersetzung zu großskaliger, kontextsensitiver Sprachmodellierung. Technisch ist Gemini als generatives LLM konzipiert (Decoder-only, autoregressiv) und nutzt zudem Mixture-of-Experts-Mechanismen, um je nach Kontext spezialisierte Teilnetze zu aktivieren, was die Modellkapazität erhöht und semantische Differenzierung begünstigen kann. Eine klare Stärke ist damit grundsätzlich die Fähigkeit, lange Kontexte kohärent zu halten und Übersetzungen zu erzeugen, die nicht nur grammatisch, sondern oft auch erzählerisch flüssig wirken. Für theologische Arbeitsschritte, die über einzelne Verse hinausgehen (Motivketten, wiederkehrende Termini, argumentatives Fortschreiten), ist dieses Potenzial attraktiv.
Gleichzeitig markieren die quantitativen Verdichtungen Gemini als Extrempol: Es zeigt die stärkste Expansion, erhöhte Halluzinationsraten und insgesamt die deutlichste strukturelle Entfernung von der Lutherreferenz. Das ist hermeneutisch relevant, weil hier Kontexttreue schnell in Richtung Interpretationsleistung kippt: Wo der Urtext bewusst knapp ist, kann Gemini dazu neigen, implizite Beziehungen explizit zu machen (z. B. Kausalitäten, Sprecherwechsel, erklärende Ergänzungen), wodurch der Zieltext zwar verständlicher wird, aber zugleich eine bestimmte Lesart stabilisiert. Für philologisch orientierte Exegese ist das eine Schwäche, weil die Offenheit des Ausgangstextes – gerade in prophetischen und poetischen Passagen – methodischer Teil des Befundes ist. Zusätzlich bleibt Gemini als kommerzielles Cloud-System intransparent hinsichtlich Modellversionen und interner Parameter, was die Reproduzierbarkeit einschränkt. Insgesamt eignet sich Gemini daher eher für explorative Vorarbeit, weniger als primäre Übersetzungsgrundlage, wenn die Analyse gezielt an Ambiguität, knapper Form und gattungsspezifischer Struktur ansetzen soll.

LLaMA-3 nimmt in gezeigten Design eine Scharnierposition ein: Es ist ein generatives LLM wie Gemini, aber als offenes Modell lokal betreibbar und damit forschungsnäher kontrollierbar. Im Gesamtkontext unterscheidet es sich in den Durchschnittswerten von den NMT-Systemen durch etwas höhere Satzlängen, zugleich aber durch geringere mittlere Halluzinations- und Drop-Werte sowie eine erhöhte TFR, was als Hinweis auf semantische Flexibilität bei gleichzeitiger relativer Formnähe gelesen werden kann. Diese Kombination ist für theologische Fragestellungen besonders interessant: LLaMA-3 kann Bedeutungszusammenhänge über den unmittelbaren Satz hinaus plausibel organisieren, ohne dabei zwingend so stark in erklärende Expansionen zu driften.

Stärken liegen daher vor allem in der Steuerbarkeit (Prompting, Formatvorgaben, Terminologie-Listen) und der Reproduzierbarkeit im lokalen Setup: Für Forschung ist entscheidend, dass ein Modell nicht nur gut klingt, sondern dass Einstellungen dokumentiert und Ergebnisse wiederholbar erzeugt werden können. Als Schwäche bleibt jedoch das grundsätzliche LLM-Risiko bestehen, dass Übersetzung als Spezialfall generativer Textproduktion erfolgt: Auch bei niedrigen quantitativen Fehlanzeigen kann LLaMA-3 in einzelnen Fällen interpretative Verdichtungen erzeugen oder stilistisch zu glatt formulieren, wodurch ambivalente Schlüsselbegriffe in eine eindeutige Richtung geschoben werden. Diese Gefahr ist nicht primär metrisch sichtbar, sondern zeigt sich in der theologischen Detailanalyse. Insgesamt erscheint LLaMA-3 als das Modell, das methodische Kontrolle (lokal/offen) mit kontextsensitiver Textproduktion am überzeugendsten verbindet – vorausgesetzt, es wird nicht als Autorität, sondern als kontrolliertes Werkzeug in einem klaren hermeneutischen Rahmen genutzt.

Insgesamt wird deutlich, dass die untersuchten Systeme weniger entlang eines einfachen gut/schlecht-Schemas zu bewerten sind, sondern als unterschiedliche Modellprofile mit je eigenen Stärken und Risiken. Die NMT-Systeme (DeepL, NLLB) überzeugen vor allem durch stabile, formal kontrollierbare Ausgaben und eignen sich dadurch gut als reproduzierbare Ausgangsbasis; ihre Grenzen liegen dort, wo gattungsspezifische Ambiguität und theologisch relevante Nuancen nicht nur übertragen, sondern bewusst offengehalten werden sollen. Die LLMs (Gemini, LLaMA-3) bieten demgegenüber ein deutlich höheres Potenzial zur kontextuellen Kohärenz und sprachlichen Ausformulierung, verschieben Übersetzung jedoch schneller in Richtung interpretativer Explikation.

5.3 Möglichkeiten des KI-Einsatzes in theologischer Forschung

Die Ergebnisse legen nahe, dass KI in der Theologie nicht als einheitliches Werkzeug zu denken ist, sondern als Spektrum unterschiedlicher Systemtypen mit je eigener Logik: spezialisierte NMTs liefern eher stabile, nachvollziehbare Satz-zu-Satz-Übertragungen, während LLMs Text als kontextsensitives Produkt generieren und damit stärker in Richtung Deutung tendieren. Daraus ergeben sich abgestufte Einsatzmöglichkeiten, die sich nach Forschungsziel, Textgattung und notwendiger methodischer Strenge unterscheiden.

Ein naheliegendes Feld ist die *sprachliche Erschließung und Vorübersetzung*. Gerade in der Arbeit mit hebräischen, griechischen oder lateinischen Quellen kann KI als schneller Erstzugang dienen, um Textsegmente zu erschließen, Varianten nebeneinanderzustellen oder Arbeitsübersetzungen für die Diskussion im Forschungskontext zu erzeugen. Dabei ist methodisch entscheidend, welches Ziel verfolgt wird: Wo es um eine möglichst formnahe, reproduzierbare Ausgangsbasis geht (z. B. als Vergleichsniveau in der Exegese oder für didaktische Arbeitsblätter), bieten sich lokal kontrollierbare Systeme an, weil ihre Ergebnisse dokumentierbar bleiben. Wo dagegen erste Hypothesen zur Makrostruktur, zu Motivketten oder zu kohärenten Deutungszusammenhängen entwickelt werden sollen, können LLMs hilfreich sein – allerdings nicht als Übersetzung, sondern als exploratives Interpretationsinstrument, dessen Ausgaben konsequent am Quelltext rückgebunden werden müssen.

Ein zweites Feld ist die *textanalytische Unterstützung*: KI kann zur Identifikation wiederkehrender Lexeme und semantischer Felder, zur Auffindung formelhafter Strukturen, zur Clusterung thematisch verwandter Perikopen oder zur Erstellung von Vergleichskorpora genutzt werden. Hier gewinnt die Frage nach Kontexttreue nochmals ein anderes Gewicht: Eine KI-gestützte Analyse kann produktiv sein, wenn sie nicht die Deutung ersetzt, sondern Aufmerksamkeit lenkt (z. B. „Welche Begriffe variieren modellübergreifend besonders stark?“). Gerade die gewonnen Ergebnisse zeigen, dass quantitative Indikatoren allein keine Qualitätsaussage erzwingen, sondern interpretativ eingeordnet werden müssen; dieser Befund lässt sich methodisch wenden: Metriken und systematische Output-Vergleiche können in der Forschung als Heuristik dienen, um hermeneutisch sensible Stellen zu markieren, die dann klassisch-exegetisch vertieft werden.

Ein drittes Feld ist die *wissensorganisierende Arbeit* in Theologie und Religionswissenschaft, etwa beim Strukturieren großer Literaturmengen, beim Erstellen von Annotationen, beim Zusammenführen von Argumentationslinien oder beim Entwurf von Kategorienschemata für qualitative Analysen. Hier liegt ein realistischer Nutzen besonders in lokal oder kontrolliert einsetzbaren Systemen, weil Forschungsarbeit nicht nur Ergebnisproduktion, sondern auch Nachvollziehbarkeit verlangt. Die angelegte Gegenüberstellung cloudbasiert vs. lokal gewinnt damit eine praktische Konsequenz: Je stärker ein Arbeitsschritt dokumentations- und revisionspflichtig ist (z. B. Edition, Kommentararbeit, Forschungsdaten), desto eher sollten Systeme bevorzugt werden, deren Ausgaben stabil reproduzierbar sind oder deren Parameter explizit festgehalten werden können.

Viertens eröffnen sich Einsatzmöglichkeiten im Bereich *Lehre, Wissenschaftskommunikation und Transfer*: KI kann Studierenden helfen, sprachliche Hürden abzubauen, unterschiedliche Übersetzungsstrategien sichtbar zu machen oder die Wirkung theologischer Schlüsselbegriffe in verschiedenen Zielsprachen zu diskutieren. Wenn Lernende sehen, wie stark sich Übersetzungen je nach Modelltyp in Expansion, Explizitheit oder Ambiguitätserhalt unterscheiden, wird Übersetzung selbst als hermeneutischer Prozess erfahrbar. Gleichzeitig müssen die Einsatzmöglichkeiten durch klare Leitplanken begrenzt werden. Erstens ist bei allen generativen Systemen konsequent zu unterscheiden zwischen Textwiedergabe und Deutung: LLM-Ausgaben sind niemals neutral, sondern entscheiden implizit für Kohärenz, Explizitheit und stilistische Plausibilität. Zweitens zeigen die Ergebnisse, dass Halluzination im technischen Sinn nicht einfach Fehler bedeutet, weil interlinguale Strukturunterschiede – gerade in poetischen und prophetischen Texten – notwendige Ergänzungen erzwingen können; genau deshalb ist eine theologisch reflektierte Qualitätsdefinition unverzichtbar. Drittens bleibt die Frage nach Transparenz und Reproduzierbarkeit zentral: In Forschungskontexten sollte KI idealerweise so eingesetzt werden, dass Einstellungen, Datenstände und Vorgehen dokumentierbar sind, damit Ergebnisse nicht als Autoritätsargument, sondern als methodisch kontrollierbare Zwischenschritte in den hermeneutischen Prozess eingehen.

Die Einsatzmöglichkeiten von KI in der theologischen Forschung sind damit real und vielfältig – von Vorübersetzung und Erschließung über textanalytische Heuristiken bis hin zu Literaturorganisation und didaktischer Unterstützung. Zugleich zeigt sich, dass der wissenschaftliche Nutzen maßgeblich davon abhängt, wie KI eingebettet wird: produktiv ist sie dort, wo sie als Werkzeug zur Strukturierung, Hypothesenbildung und Markierung exegetisch sensibler Stellen dient, nicht aber als Instanz, die Deutung autoritativ festschreibt. Methodisch entscheidend bleiben daher dokumentierbare Rahmenbedingungen (Reproduzierbarkeit, Versionierung, Prompt-/Setting-Transparenz) und eine klare hermeneutische Disziplin, die zwischen Übersetzungsleistung, interpretativer Glättung und genuiner Sinnerschließung unterscheidet. KI kann somit weder Theolog:innen vom Erlernen der Ursprachen befreien noch theologische Forschung ersetzen, wohl aber – richtig gerahmt – Letztere effizienter, systematischer und in Teilen auch reflexiver machen und ein wertvolles Unterstützungstool sein.

5.4 Zusammenfassung der Ergebnisse

Die vorliegende Arbeit ging von der leitenden Frage aus, in welchem Umfang neuronale Übersetzungssysteme kontexttreue Übersetzungen alttestamentlicher Texte erzeugen und wie sich dabei NMT-Systeme und LLM-basierte Modelle hinsichtlich Strukturtreue, semantischer Präzision und hermeneutisch-theologischer Anschlussfähigkeit unterscheiden. Die Ergebnisse zeigen dazu in einer übergreifenden Perspektive ein klares, zugleich differenziertes Bild: Kontexttreue lässt sich weder als reine Referenznähe zu einer etablierten Bibelübersetzung noch als bloße sprachliche Korrektheit definieren, sondern nur als mehrdimensionale Zielgröße, die technische Stabilität und theologische Interpretierbarkeit zusammenbindet. Damit beantwortet die Arbeit die Forschungsfrage nicht im Sinne eines einfachen Rankings, sondern durch die Herausarbeitung konsistenter Modellprofile und ihrer Konsequenzen für exegetische Praxis.

Auf der Ebene der Strukturtreue wurde sichtbar, dass die Systeme in unterschiedlicher Weise mit den formalen Eigenheiten des biblischen Hebräisch umgehen: Während einige Modelle eher stabil und konservativ übersetzen, zeigen andere eine stärkere Tendenz zur Umformung, Explizitation und Umstrukturierung. Entscheidend ist jedoch, dass strukturelle Abweichung nicht automatisch Unzuverlässigkeit bedeutet. Gerade bei alttestamentlichen Texten – besonders in prophetischer und poetischer Rede – sind zielsprachliche Restrukturierungen häufig unvermeidlich, um Ellipsen, Parallelismen oder semantische Verdichtung in eine andere Sprachlogik zu übertragen. Die Arbeit konnte damit zeigen, dass Strukturtreue zwar ein notwendiger Indikator ist, aber nur dann aussagekräftig wird, wenn sie hermeneutisch eingeordnet wird: Nicht jede Expansion ist Interpretation, und nicht jede formale Glätte ist Treue.

Bezogen auf die semantische Präzision ergibt sich eine zweite zentrale Antwort auf die Forschungsfrage: Maschinelle Systeme können häufig sehr plausible und sprachlich überzeugende Zieltexte produzieren, doch gerade diese Plausibilität ist ambivalent. Denn sie kann verdecken, dass an einzelnen sensiblen Stellen Bedeutungsachsen verschoben, Mehrdeutigkeiten aufgelöst oder theologisch aufgeladene Schlüssellexeme in eine Richtung stabilisiert werden, die im Ausgangstext nicht zwingend festgelegt ist. Damit bestätigt die Untersuchung die hermeneutische Grundannahme der Einleitung: Übersetzung ist immer auch Deutung – und bei KI-Systemen tritt diese Deutung nicht als bewusst begründete Entscheidung, sondern als modellintern erzeugte Kohärenzleistung zutage. In Bezug auf die Forschungsfrage heißt das: Kontexttreue im semantischen Sinn ist bei allen Modelltypen nur bedingt gewährleistet; die entscheidende Differenz liegt weniger in der Existenz von interpretativen Effekten als in ihrer Häufigkeit, ihrer Vorhersagbarkeit und ihrer methodischen Kontrollierbarkeit.

Gerade hier wird die dritte Dimension der Forschungsfrage, die hermeneutisch-theologische Anschlussfähigkeit, zur Schlüsselkategorie. Die Arbeit zeigt, dass ein Zieltext theologisch anschlussfähig ist, wenn er den interpretativen Raum des Ausgangstextes nicht vorschnell schließt, gattungstypische Funktionen (Erzählführung, prophetische Zuspitzung, poetische Bildlogik) erkennbar hält und zugleich in einem Forschungsprozess transparent überprüfbar bleibt. In dieser Perspekti-

ve werden die Modellunterschiede besonders deutlich: NMT-Systeme erweisen sich insgesamt als stabilere Werkzeuge für reproduzierbare, vergleichende Arbeitsschritte, während LLM-basierte Systeme stärker zu kontextsensitiver Ausformulierung neigen und damit schneller in Richtung explizierender Deutung verschieben. Diese Eigenschaft kann für bestimmte Forschungsziele produktiv sein, ist aber zugleich genau dort kritisch, wo die Offenheit des Textes selbst Bestandteil der exegetischen Erkenntnis ist. Damit beantwortet die Arbeit die Forschungsfrage auch normativ-methodisch: Nicht die Frage, *ob* KI in der Theologie genutzt werden darf, steht im Vordergrund, sondern *unter welchen Bedingungen* sie wissenschaftlich verantwortbar ist.

Daraus folgt als abschließendes Ergebnis: Keines der untersuchten Systeme garantiert Kontexttreue im theologischen Sinn ohne methodische Sicherungen. Zugleich wird deutlich, dass KI-Übersetzung in der theologischen Forschung sinnvoll eingesetzt werden kann, wenn sie als Werkzeug und nicht als Autorität verstanden wird – also eingebettet in dokumentierte Settings, mit transparenter Vergleichsbasis und konsequenter Rückführung auf den Quelltext sowie auf die Übersetzungstradition als hermeneutischen Resonanzraum. Damit schließt die Arbeit an die Motivation der Einleitung an: Gerade weil maschinelle Übersetzung zunehmend selbstverständlich verfügbar ist, besteht die wissenschaftliche Relevanz nicht darin, ihre Existenz zu bestätigen, sondern ihren Einsatz zu qualifizieren. Die Untersuchung leistet hierzu einen Beitrag, indem sie zeigt, dass Kontexttreue in diesem Feld nicht durch Metriken allein, sondern nur durch die Verbindung technischer Profile mit theologischer Urteilskraft begründbar wird – und dass die produktive Nutzung von KI in der Exegese dort beginnt, wo ihre interpretativen Tendenzen sichtbar gemacht und methodisch kontrolliert werden.

LITERATURVERZEICHNIS

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., España-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., … Zampieri, M. (2021). Findings of the 2021 Conference on Machine Translation (WMT21). In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, & C. Monz (Hrsg.), *Proceedings of the Sixth Conference on Machine Translation* (S. 1–88). Association for Computational Linguistics.

Arnold, D., Balkan, L., Humphreys, R. L., Meijer, S., & Sadler, L. (1994). *Machine Translation: An Introductory Guide*. Blackwell Publishers.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural Machine Translation by Jointly Learning to Align and Translate*.

Becker, U. (2021). *Exegese des Alten Testaments—Ein Methoden- und Arbeitsbuch* (5. Auflage). Mohr Siebeck.

Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*.

Dao, T. (2023). *FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning*.

Decker, R. J. (2014). *Reading Koine Greek: An Introduction and Integrated Workbook*.

Donner, H., & Gesenius, W. (2007). *Hebräisches und Aramäisches Handwörterbuch über das Alte Testament*. Springer.

Dyer, C., Chahuneau, V., & Smith, N. A. (2013). *A Simple, Fast, and Effective Reparameterization of IBM Model 2*. 644–648.

Enis, M., & Hopkins, M. (2024). *From LLM to NMT: Advancing Low-Resource Machine Translation with Claude*.

Fedus, W., Zoph, B., & Shazeer, N. (2021). *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*.

Fischer, G. (1998). Bibelhandschriften. In *RELIGION IN GESCHICHTE UND GEGENWART - Handwörterbuch für Theologie und Religionswissenschaft* (Vierte, neu bearb. Auflage, Bd. 1, S. 1455–1459). Mohr Siebeck.

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the Knowledge in a Neural Network*.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*, 1735–1780.

Howard, J., & Ruder, S. (2018). *Universal Language Model Fine-tuning for Text Classification*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*.

Hutchins, J. (2004). The Georgetown-IBM Experiment Demonstrated in January 1954. In R. E. Frederking & K. B. Taylor (Hrsg.), *Machine Translation: From Real Users to Research* (S. 102–114). Springer.

Hutchins, J. (2014). *The history of machine translation in a nutshell*.

Jenni, E. (1981). *Lehrbuch der hebräischen Sprache des Alten Testaments.* (Zweite, durchgesehene Auflage). Helbing & Lichtenhan.

Koehn, P., & Knowles, R. (2017). *Six Challenges for Neural Machine Translation*.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 48–54.

Kudo, T., & Richardson, J. (2018). *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*.

Kühner, R. (1890). *Ausführliche Grammatik der Griechischen Sprache* (3. Auflage). Hahnsche Buchhandlung.

Läubli, S., Sennrich, R., & Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Hrsg.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (S. 4791–4796). Association for Computational Linguistics.

Lebedewa, J. (2007). *Mit anderen Worten—Die vollkommene Übersetzung bleibt eine Utopie*. https://www.uni-heidelberg.de/presse/ruca/ruca07-3/wort.html

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., & Chen, Z. (2020). *GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding*.

Liebeskind, S., Liebeskind, C., & Bouhnik, D. (2021). *Machine Translation for Historical Research: A case study of Aramaic-Ancient Hebrew Translations*. https://www.semantic-web-journal.net/content/machine-translation-historical-research-case-study-aramaic-ancient-hebrew-translations

Maruf, S., Saleh, F., & Haffari, G. (2021). A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys*, *54*(2), 45.

Mathur, N., Baldwin, T., & Cohn, T. (2020). Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4984–4997.

Meghouche, K. (2025). Das Tempussystem im Deutschen und Arabischen – eine kontrastive Studie – TRANS Nr. 23. *TRANS Internet-Zeitschrift für Kulturwissenschaften*. https://www.inst.at/trans/23/das-tempussystem-im-deutschen-und-arabischen-eine-kontrastive-studie/

Moghe, N., Fazla, A., Amrhein, C., Kocmi, T., Steedman, M., Birch, A., Sennrich, R., & Guillou, L. (2025). Machine Translation Meta Evaluation through Translation Accuracy Challenge Sets. *Computational Linguistics*, *51*(1), 73–137.

Müller, M., Rios, A., & Sennrich, R. (2020). *Domain Robustness in Neural Machine Translation*.

Naveen, P., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *iScience*, *27*(10).

Neef, H.-D. (2018). *Arbeitsbuch Hebräisch*. Mohr Siebeck.

Nida, E. A., & Taber, C. R. (1969). *The theory and practice of translation* (3. Auflage).

Nord, C. (1995). *Textanalyse und Übersetzen: Theoretische Grundlagen, Methode und didaktische Anwendung einer übersetzungsrelevanten Textanalyse* (3. Auflage). Julius Gross Verlag.

Pang, J., Ye, F., Wang, L., Yu, D., Wong, D. F., Shi, S., & Tu, Z. (2024). *Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models*.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, & P. Pecina (Hrsg.), *Proceedings of the Tenth Workshop on Statistical Machine Translation* (S. 392–395). Association for Computational Linguistics.

Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., & Kaur, R. (2023). Neural Machine Translation for Low-resource Languages: A Survey. *ACM Comput. Surv.*, *55*(11).

Rapacz, M., & Smywiński-Pohl, A. (2025). Low-Resource Interlinear Translation: Morphology-Enhanced Neural Models for Ancient Greek. In H. Hettiarachchi, T. Ranasinghe, P. Rayson, R. Mitkov, M. Gaber, D. Premasiri, F. A. Tan, & L. Uyangodage (Hrsg.), *Proceedings of the First Workshop on Language Models for Low-Resource Languages* (S. 145–165). Association for Computational Linguistics.

Raunak, V., Menezes, A., & Junczys-Dowmunt, M. (2021). *The Curious Case of Hallucinations in Neural Machine Translation*.

Rösel, M. (2006). *Bibelkunde des Alten Testaments: Die kanonischen und apokryphen Schriften*. Neukirchener Verlag.

Schwanke, M. (1991). Geschichte der Maschinellen Übersetzung. In M. Schwanke (Hrsg.), *Maschinelle Übersetzung: Ein Überblick über Theorie und Praxis* (S. 69–95). Springer.

Sennrich, R., Haddow, B., & Birch, A. (2016). *Improving Neural Machine Translation Models with Monolingual Data*.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*.

Siebenthal, H. (2008). *Grundkurs Neutestamentliches Griechisch: Grammatik - Grundwortschatz - Übersetzungsmethodik*. Brunnen Verlag GmbH.

Sizov, F., España-Bonet, C., Van Genabith, J., Xie, R., & Dutta Chowdhury, K. (2024). Analysing Translation Artifacts: A Comparative Study of LLMs, NMTs, and Human Translations. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Hrsg.), *Proceedings of the Ninth Conference on Machine Translation* (S. 1183–1199). Association for Computational Linguistics.

Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I., & de Freitas, N. (2023). Machine Learning for Ancient Languages: A Survey. *Computational Linguistics*, *49*(3), 703–747.

Stahlberg, F. (2020). Neural Machine Translation: A Review. *Journal of Artificial Intelligence Research*, *69*, 343–418.

Stampf, J. (2012). *Maschinelle Übersetzung—Ein kritischer Vergleich* [Masterarbeit, Universtität Wien]. https://phaidra.univie.ac.at/download/o:1282356

Stein, D. (2009). Maschinelle Übersetzung—Ein Überblick. *Journal for Language Technology and Computational Linguistics*, *24*(Heft 3). https://jlcl.org/article/download/119/117/124

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to Sequence Learning with Neural Networks*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Conference on Neural Information Processing System*.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2021). *Finetuned Language Models Are Zero-Shot Learners*.

Wu, X., Liu, H., Zhou, J., Zhao, X., Xu, L., Wang, L., Luo, W., & Zhang, K. (2025). *Challenging Multilingual LLMs: A New Taxonomy and Benchmark for Unraveling Hallucination in Translation*.

Wu, Z., Wei, D., Chen, X., Shang, H., Guo, J., Li, Z., Luo, Y., Yang, J., Rao, Z., & Yang, H. (2025). *Combining the Best of Both Worlds: A Method for Hybrid NMT and LLM Translation*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating Text Generation with BERT*.

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). *Calibrate Before Use: Improving Few-Shot Performance of Language Models*.

ANHANG

Anhangsverzeichnis

Anhang 1-1 Exodus 3, 11 – Griechisch | Latein

BHS	וַיֹּאמֶר מֹשֶׁה אֶל־הָאֱלֹהִים מִי אָנֹכִי כִּי־אֵלֵךְ אֶל־פַּרְעֹה וְכִי אוֹצִיא אֶת־בְּנֵי יִשְׂרָאֵל מִמִּצְרָיִם
unvokalisiert	ויאמר משה אל האלהים מי אנכי כי אלך אל פרעה וכי אוציא את בני ישראל ממצרים

LXX αὶ εἶπεν Μωυσῆς πρὸς τὸν θεόν Τίς εἰμι, ὅτι πορεύσομαι πρὸς Φαραω βασιλέα Αἰγύπτου, καὶ ὅτι ἐξάξω τοὺς υἱοὺς Ισραηλ ἐκ γῆς Αἰγύπτου;

DeepL	---
NLLB-200	---
Gemini	καὶ εἶπεν Μωυσῆς πρὸς τὸν θεόν· Τίς εἰμι ἐγώ, ὅτι πορεύσομαι πρὸς Φαραὼ καὶ ὅτι ἐξάξω τοὺς υἱοὺς Ἰσραὴλ ἐξ Αἰγύπτου;
LLaMA3	Εἶπεν Μωυσῆς πρὸς τὸν Θεὸν Τίς εἰμὶ ἐγὼ ὅτι ἀπέλθω πρὸς Φαραὼ καὶ ὅτι ἐξαγαγῶ τοὺς υἱοὺς Ἰσραήλ ἐκ Μισοραίμ.

VUL dixit Moses ad Deum quis ego sum ut vadam ad Pharaonem et educam filios Israhel de Aegypto

DeepL	Et dixit Moyses ad Deum: "Quis ego sum, ut eat ad Pharaonem, et ut educam filios Israel de Aegypto?"
NLLB-200	---
Gemini	Et ait Moyses ad Deum: Quis sum ego, ut eam ad Pharaonem et ut educam filios Israel ex Aegypto?
LLaMA3	Quis ego sum, ut ascendam ad Pharaonem, et ut eduxim filios Israel de Aegypto?

Anhang 1-2 **Exodus 3, 11** – Deutsch | Englisch

LUTH Mose sprach zu Gott: Wer bin ich, dass ich zum Pharao gehe und führe die Israeliten aus Ägypten?

DeepL	und Mose sprach zu Gott: Wer bin ich, dass ich zum Pharao gehen und die Kinder Israel aus Ägypten führen soll?
NLLB-200	Und Mose sprach zu Gott: Wer bin ich, daß ich zu Pharao gehe und die Kinder Israel aus Ägypten herausbringe?
Gemini	Und Mose sprach zu Gott: Wer bin ich, dass ich zum Pharao gehe und dass ich die Söhne Israels aus Ägypten herausführe?
LLaMA3	Und Moses sprach zu Gott: Wer bin ich, dass ich hingehe zu Pharao und dass ich die Söhne Israels aus Ägypten führe?

BASIS Mose sagte zu Gott: »Wer bin ich denn, dass ich einfach zum Pharao gehe? Und wie soll ich die Israeliten aus Ägypten führen?

KJV And Moses said unto God, Who am I, that I should go unto Pharaoh, and that I should bring forth the children of Israel out of Egypt?

DeepL	and Moses said unto God, Who am I, that I should go unto Pharaoh, and that I should bring forth the children of Israel out of Egypt?
NLLB-200	And Moses said to God, Who am I, that I should go to Pharaoh, and that I should bring forth the children of Israel out of Egypt?
Gemini	And Moses said to God, 'Who am I, that I should go to Pharaoh and that I should bring out the sons of Israel from Egypt?'
LLaMA3	And Moses said to God, "Who am I that I should go to Pharaoh and that I should bring out the sons of Israel from Egypt?

NLT And Moses said unto God, Who am I, that I should go unto Pharaoh, and that I should bring forth the children of Israel out of Egypt?

Anhang 1-3 **Exodus 3, 12** – Griechisch | Latein

BHS	וַיֹּאמֶר כִּי־אֶהְיֶה עִמָּךְ וְזֶה־לְּךָ הָאוֹת כִּי אָנֹכִי שְׁלַחְתִּיךָ בְּהוֹצִיאֲךָ אֶת־הָעָם מִמִּצְרַיִם תַּעַבְדוּן אֶת־הָאֱלֹהִים עַל הָהָר הַזֶּה
unvokalisiert	ויאמר כי אהיה עמך וזה לך האות כי אנכי שלחתיך בהוציאך את העם ממצרים תעבדון את האלהים על ההר הזה

LXX εἶπεν δὲ ὁ θεὸς Μωυσεῖ λέγων ὅτι Ἔσομαι μετὰ σοῦ, καὶ τοῦτό σοι τὸ σημεῖον ὅτι ἐγώ σε ἐξαποστέλλω· ἐν τῷ ἐξαγαγεῖν σε τὸν λαόν μου ἐξ Αἰγύπτου καὶ λατρεύσετε τῷ θεῷ ἐν τῷ ὄρει τούτῳ.

DeepL	---
NLLB-200	---
Gemini	καὶ εἶπεν· Ἐγὼ ἔσομαι μετὰ σοῦ, καὶ τοῦτό σοι τὸ σημεῖον ὅτι ἐγώ σε ἀπέστειλα· ἐν τῷ ἐξαγαγεῖν σε τὸν λαὸν ἐξ Αἰγύπτου, λατρεύσετε τῷ θεῷ ἐν τῷ ὄρει τούτῳ.
LLaMA3	Καὶ εἶπεν ὅτι ἔσομαι μετὰ σοῦ καὶ τοῦτο τὸ σημεῖον ἔσται σοὶ ὅτι ἐγὼ ἀπέστειλα σὲ ἵνα ἐξαγαγῇς τὸν λαὸν ἐκ τῆς Αἰγύπτου καὶ δουλεύσουσιν τῷ Θεῷ ἐπὶ τῷ ὄρει τούτῳ.

VUL qui dixit ei ero tecum et hoc habebis signum quod miserim te cum eduxeris populum de Aegypto immolabis Deo super montem istum

DeepL	Et dixit, "Ego ero tecum, et hocsignum quod ego te misi: cum populum ex Aegypto eduxeris, Deum in hoc monte adorabitis.
NLLB-200	---
Gemini	Ait: Ego ero tecum, et hoc tibi erit signum quod ego te misi: cum eduxeris populum de Aegypto, servietis Deo super montem hunc.
LLaMA3	Dixit quia ego ero tecum et haec signum tibi erit quia ego misi te ut educerem populum de Aegypto ut serviant Deo in hoc monte.

Anhang 1-4 Exodus 3, 12 – Deutsch | Englisch

LUTH Er sprach: Ich will mit dir sein. Und das soll dir das Zeichen sein, dass ich dich gesandt habe: Wenn du mein Volk aus Ägypten geführt hast, werdet ihr Gott dienen auf diesem Berge.

DeepL	und er sprach: Ich werde mit dir sein, und dasdas ist für dich das Zeichen, dass ich dich gesandt habe: Wenn du das Volk aus Ägypten herausführst, werdet ihr Gott auf diesem Berg dienen
NLLB-200	Und er sprach: Ich will mit dir sein, und siehe, das ist dein Zeichen; denn ich habe dich gesandt, als du das Volk aus Ägypten führst; ihr werdet Gott dienen auf diesem Berg.
Gemini	Und er sprach: Ich werde mit dir sein, und dies sei dir das Zeichen, dass ich dich gesandt habe: Wenn du das Volk aus Ägypten herausgeführt hast, werdet ihr Gott auf diesem Berg dienen.
LLaMA3	Ich werde bei dir sein und dies ist dir das Zeichen, dass ich dich gesandt habe, um das Volk aus Ägypten zu führen, dass sie den Gott auf diesem Berge dienen werden.

BASIS Gott antwortete: »Ich werde bei dir sein! Daran wirst du sehen, dass ich dich gesandt habe: Wenn du das Volk aus Ägypten geführt hast, sollt ihr mir an diesem Berg dienen.«

KJV And he said, Certainly I will be with thee; and this shall be a token unto thee, that I have sent thee: When thou hast brought forth the people out of Egypt, ye shall serve God upon this mountain.

DeepL	and he said, "I will be with you, and thisthis shall be the sign to you that it is I who have sent you: when you have brought the people out of Egypt, you shall serve God upon this mountain
NLLB-200	And he said, For I will be with thee; and behold, this shall be a sign to thee, for I have sent thee; for when thou hast brought forth the people out of Egypt, ye shall serve God upon this mountain.
Gemini	And he said, 'I will be with you, and this shall be the sign for you that I have sent you: when you have brought the people out of Egypt, you shall serve God on this mountain.'
LLaMA3	And he said, When I am with you, and this shall be the sign to you, that I have sent you, when you have brought forth the people from the land of Egypt, you shall serve the God on this mountain.

NLT God answered, “I will be with you. And this is your sign that I am the one who has sent you: When you have brought the people out of Egypt, you will worship God at this very mountain.”

Anhang 1-5 Exodus 3, 13 – Griechisch | Latein

BHS	אֹמֶר מֹשֶׁה אֶל־הָאֱלֹהִים הִנֵּה אָנֹכִי בָא אֶל־בְּנֵי יִשְׂרָאֵל וְאָמַרְתִּי לָהֶם אֱלֹהֵי אֲבוֹתֵיכֶם שְׁלָחַנִי אֲלֵיכֶם וְאָמְרוּ־לִי מַה־שְּׁמוֹ מָה אֹמַר אֲלֵהֶם
unvokalisiert	ויאמר משה אל האלהים הנה אנכי בא אל בני ישראל ואמרתי להם אלהי אבותיכם שלחני אליכם ואמרו לי מה שמו מה אמר אלהם

LXX καὶ εἶπεν Μωυσῆς πρὸς τὸν θεόν Ἰδοὺ ἐγὼ ἐλεύσομαι πρὸς τοὺς υἱοὺς Ισραηλ καὶ ἐρῶ πρὸς αὐτούς Ὁ θεὸς τῶν πατέρων ὑμῶν ἀπέσταλκέν με πρὸς ὑμᾶς, ἐρωτήσουσίν με Τί ὄνομα αὐτῷ; τί ἐρῶ πρὸς αὐτούς;

DeepL	---
NLLB-200	---
Gemini	Καὶ εἶπε Μωυσῆς πρὸς τὸν θεόν· Ἰδοὺ ἐγὼ ἐλεύσομαι πρὸς τοὺς υἱοὺς Ἰσραὴλ καὶ ἐρῶ αὐτοῖς· Ὁ θεὸς τῶν πατέρων ὑμῶν ἀπέσταλκέ με πρὸς ὑμᾶς. καὶ ἐροῦσί μοι· Τί ὄνομα αὐτῷ; τί ἐρῶ πρὸς αὐτούς;
LLaMA3	Καὶ εἶπεν Μωυσῆς πρὸς τὸν Θεὸν ἰδοὺ ἔρχομαι πρὸς υἱοὺς Ισραηλ καὶ ἐροῦν τῷ λαῷ Θεὸν τῶν πατέρων ὑμῶν ἀπέσταλκέν με πρὸς ὑμᾶς καὶ ἐροῦσιν μοι τί ἐρῶ αὐτοῖς

VUL ait Moses ad Deum ecce ego vadam ad filios Israhel et dicam eis Deus patrum vestrorum misit me ad vos si dixerint mihi quod est nomen eius quid dicam eis

DeepL	Et dixit Moyses ad Deum: Ecce, venio adliberis Israel, et dicam eis: 'Deus patrum vestrorum me ad vos misit,' et dicunt mihi: 'Quod est nomen eius?' Quid dicam eis?
NLLB-200	---
Gemini	Dixit Moyses ad Deum: "Ecce, ego veniam ad filios Israel et dicam eis: Deus patrum vestrorum misit me ad vos. Et dicent mihi: Quid est nomen eius? Quid dicam eis?"
LLaMA3	Ecce venio ad filios Israel et dico eis Deum patrum vestrorum misit me ad vos et dicent mihi: Quid est nomen eius quid dico eis?

Anhang 1-6 Exodus 3, 13 – Deutsch | Englisch

LUTH Mose sprach zu Gott: Siehe, wenn ich zu den Israeliten komme und spreche zu ihnen: Der Gott eurer Väter hat mich zu euch gesandt!, und sie mir sagen werden: Wie ist sein Name?, was soll ich ihnen sagen?

DeepL	und Moses sprach zu Gott: Siehe, ich komme zuund sage zu ihnen: Der Gott eurer Väter hat mich zu euch gesandt. Und sie werden mich fragen: Wie heißt er? Was soll ich ihnen sagen?
NLLB-200	Und Mose sprach zu Gott: Siehe, ich komme zu den Kindern Israel und sage ihnen: Der Gott eurer Väter hat mich zu euch gesandt. Und siehe, sie werden mich fragen: Was ist sein Name? Was soll ich ihnen sagen?
Gemini	Da sprach Mose zu Gott: „Siehe, ich komme zu den Söhnen Israels und sage zu ihnen: ‚Der Gott eurer Väter hat mich zu euch gesandt.‘ Und sie werden zu mir sagen: ‚Was ist sein Name?‘ Was soll ich ihnen sagen?“
LLaMA3	Und Moses sprach zu Gott: Siehe, ich gehe zu den Söhnen Israels hin und sage ihnen: Götter eurer Väter haben mich gesandt zu euch, und sie werden zu mir sagen: Was ist sein Name, was sage ich ihnen?

BASIS Mose antwortete Gott: »Ich werde zu den Israeliten gehen und ihnen sagen: ›Der Gott eurer Väter schickt mich zu euch.‹ Was ist, wenn sie mich fragen: ›Wie heißt er?‹ Was soll ich ihnen dann sagen?«

KJV And Moses said unto God, Behold, when I come unto the children of Israel, and shall say unto them, The God of your fathers hath sent me unto you; and they shall say to me, What is his name? what shall I say unto them?

DeepL	and Moses said unto God, Behold, I come untothe children of Israel, and say to them, 'The God of your fathers has sent me to you,' and they say to me, 'What is His name?' What shall I say to them?"
NLLB-200	And Moses said to God, Behold, I will come to the children of Israel, and say to them, The God of your fathers hath sent me to you; and they will say to me, What is his name? what shall I say to them?
Gemini	And Moses said to God, "Behold, I am going to the children of Israel, and I will say to them, 'The God of your fathers has sent me to you.' And they will say to me, 'What is his name?' What shall I say to them?"
LLaMA3	And Moses said to God, Behold, I am going to go to the children of Israel and I will say to them, God of their fathers has sent me to you, and they will say to me, What is his name, what shall I say to them?

NLT But Moses protested, “If I go to the people of Israel and tell them, ‘The God of your ancestors has sent me to you,’ they will ask me, ‘What is his name?’ Then what should I tell them?”

Anhang 1-7 **Exodus 3, 14** – Griechisch | Latein

BHS	וַיֹּאמֶר אֱלֹהִים אֶל־מֹשֶׁה אֶהְיֶה אֲשֶׁר אֶהְיֶה וַיֹּאמֶר כֹּה תֹאמַר לִבְנֵי יִשְׂרָאֵל אֶהְיֶה שְׁלָחַנִי אֲלֵיכֶם
unvokalisiert	ויאמר אלהים אל משה אהיה אשר אהיה ויאמר כה תאמר לבני ישראל אהיה שלחני אליכם

LXX καὶ εἶπεν ὁ θεὸς πρὸς Μωυσῆν Ἐγώ εἰμι ὁ ὤν· καὶ εἶπεν Οὕτως ἐρεῖς τοῖς υἱοῖς Ισραηλ Ὁ ὢν ἀπέσταλκέν με πρὸς ὑμᾶς.

DeepL	---
NLLB-200	---
Gemini	καὶ εἶπεν ὁ θεὸς πρὸς Μωυσῆν· Ἐγώ εἰμι ὁ ὤν. καὶ εἶπεν· Οὕτως ἐρεῖς τοῖς υἱοῖς Ἰσραήλ· Ὁ ὢν ἀπέσταλκέν με πρὸς ὑμᾶς.
LLaMA3	Καὶ εἶπεν ὁ Θεός πρὸς Μωυσῆν ἔσομαι ὅστις ἔσομαι καὶ εἶπεν οὕτως εἴπες τοῖς υἱοῖς Ισραηλ ἔσομαι ἀποστείλαντί ὑμῖν.

VUL dixit Deus ad Mosen ego sum qui sum ait sic dices filiis Israhel qui est misit me ad vos

DeepL	Et Deus dixit adMoses, "Ero quod ero, " et dixit, "Sic dices filiis Israel: 'Ero me misit ad vos.'"
NLLB-200	---
Gemini	Dixit Deus ad Moysen: Ego sum qui sum. Ait: Sic dices filiis Israel: Qui sum misit me ad vos.
LLaMA3	Et dixit Deus ad Moysen: Eheu quod erit, quod erit, et dixit: Sic loqueris filiis Israel: Eheu quod misi vobis.

Anhang 1-8 Exodus 3, 14 – Deutsch | Englisch

LUTH Gott sprach zu Mose: חֶמְאָה Und sprach: So sollst du zu den Israeliten sagen: »Ich werde sein«, der hat mich zu euch gesandt.

DeepL	und Gott sprach zumose: „Ich werde sein, was ich sein werde.“ Und er sprach: „So sollst du zu den Kindern Israels sagen: Ich werde euch senden.“
NLLB-200	Und Gott sprach zu Mose: Ich werde sein, der ich bin. Und er sprach: So sollst du zu den Kindern Israel sagen: Ich werde mich zu euch gesandt haben.
Gemini	Und Gott sprach zu Mose: Ich werde sein, der ich sein werde. Und er sprach: So sollst du zu den Söhnen Israels sagen: ›Ich werde sein‹ hat mich zu euch gesandt.
LLaMA3	Und Gott sprach zu Mose: Ich werde sein, der ich sein werde, und er sprach: So sollst du zu den Söhnen Israels sprechen: Ich werde senden, der ich senden werde.

BASIS Da sprach Gott zu Mose: »›Ich werde sein, der ich sein werde.‹ Das sollst du den Israeliten sagen: Der ›Ich-werde-sein‹ hat mich zu euch geschickt.«

KJV And God said unto Moses, I AM THAT I AM: and he said, Thus shalt thou say unto the children of Israel, I AM hath sent me unto you.

DeepL	and God said tomoses, "I will be what I will be," and he said, "Thus you shall say to the children of Israel, 'I will be has sent me to you.'"
NLLB-200	And God said to Moses, I will be he that I will be. And he said, Thus shalt thou say to the children of Israel, I will be hath sent me to you.
Gemini	And God said to Moses, 'I will be what I will be.' And he said, 'Thus you shall say to the sons of Israel, “I WILL BE has sent me to you.”'
LLaMA3	And God said to Moses, "I will be who I will be, and He said, Thus you shall say to the children of Israel, I will be sent to you.

NLT God replied to Moses, “ I AM WHO I AM . Say this to the people of Israel: I AM has sent me to you.”

Anhang 1-9 **Exodus 3, 15** – Griechisch | Latein

BHS	וַיֹּאמֶר עוֹד אֱלֹהִים אֶל־מֹשֶׁה כֹּה־תֹאמַר אֶל־בְּנֵי יִשְׂרָאֵל יְהוָה אֱלֹהֵי אֲבֹתֵיכֶם אֱלֹהֵי אַבְרָהָם אֱלֹהֵי יִצְחָק וֵאלֹהֵי יַעֲקֹב שְׁלָחַנִי אֲלֵיכֶם זֶה־שְּׁמִי לְעֹלָם וְזֶה זִכְרִי לְדֹר דֹּר
unvokalisiert	ויאמר עוד אלהים אל משה כה תאמר אל בני ישראל יהוה אלהי אבתיכם אלהי אברהם אלהי יצחק ואלהי יעקב שלחני " אליכם זה שמי לעלם וזה זכרי לדר דר

LXX καὶ εἶπεν ὁ θεὸς πάλιν πρὸς Μωυσῆν Οὕτως ἐρεῖς τοῖς υἱοῖς Ισραηλ Κύριος ὁ θεὸς τῶν πατέρων ὑμῶν, θεὸς Αβρααμ καὶ θεὸς Ισαακ καὶ θεὸς Ιακωβ, ἀπέσταλκέν με πρὸς ὑμᾶς· τοῦτό μού ἐστιν ὄνομα αἰώνιον καὶ μνημόσυνον γενεῶν γενεαῖς.

DeepL	---
NLLB-200	---
Gemini	καὶ εἶπεν ὁ θεὸς πάλιν πρὸς Μωυσῆν· Οὕτως ἐρεῖς τοῖς υἱοῖς Ισραηλ· Κύριος ὁ θεὸς τῶν πατέρων ὑμῶν, θεὸς Αβρααμ καὶ θεὸς Ισαακ καὶ θεὸς Ιακωβ, ἀπέσταλκέν με πρὸς ὑμᾶς· τοῦτό μού ἐστιν ὄνομα αἰώνιον καὶ μνημόσυνόν μου γενεαῖς γενεῶν.
LLaMA3	Και εἶπεν ὁ θεὸς πάλιν Μωυσῆ Τούτο εἶπεν εἰς υἱοὺς Ισραηλ Κύριος ὁ θεὸς τῶν πατέρων ὑμῶν ὁ θεὸς Ἀβραάμ ὁ θεὸς Ἰσαάκ καὶ ὁ θεὸς Ιακώβ ἀπέστειλεν με πρὸς ὑμᾶς τοῦτο τὸ ὄνομμά μου εἰς τὸν αἰῶνα καὶ τοῦτο τὸ μνημόσυνον ἐπὶ πάντα τὰ ἔτη.

VUL dixitque iterum Deus ad Mosen haec dices filiis Israhel Dominus Deus patrum vestrorum Deus Abraham Deus Isaac et Deus Iacob misit me ad vos hoc nomen mihi est in aeternum et hoc memoriale meum in generationem et generatione

DeepL	Et dixit Deus ad Mosen, "Sic dices adFiliis Israel: Dominus, Deus patrum vestrorum, Deus Abrahae, Deus Isaac, et Deus Iacob, misit me ad vos.Hoc est nomen meum in aeternum, et hoc est memoriale meum in omnes generationes."
NLLB-200	---
Gemini	Dixitque iterum Deus ad Moysen: Haec dices filiis Israel: Dominus, Deus patrum vestrorum, Deus Abraham, Deus Isaac et Deus Iacob, misit me ad vos. Hoc nomen mihi est in aeternum, et hoc memoriale meum in generationem et generationem.
LLaMA3	Dixit Deus iterum Moysi: Hæc dices filiis Israel: Dominus Deus patrum tuorum Deus Abraham Deus Isaac et Deus Jacob misit me vobis hæc nomen meum in æternum et hæc memoria mea in generationem generationem.

Anhang 1-10 **Exodus 3, 15** – Deutsch | Englisch

LUTH Und Gott sprach weiter zu Mose: So sollst du zu den Israeliten sagen: Der Herr, der Gott eurer Väter, der Gott Abrahams, der Gott Isaaks, der Gott Jakobs, hat mich zu euch gesandt. Das ist mein Name auf ewig, mit dem man mich anrufen soll von Geschlecht zu Geschlecht.

DeepL	und Gott sprach weiter zu Mose: So sollst du zu denden Kindern Israels: Der Herr, der Gott eurer Väter, der Gott Abrahams, der Gott Isaaks und der Gott Jakobs, hat mich zu euch gesandtmein Name ist für immer, und dies ist mein Gedächtnis für alle Generationen
NLLB-200	Und Gott sprach noch zu Mose: So sollst du zu den Kindern Israel sagen: Ich, Jehova, der Gott eurer Väter, der Gott Abrahams, der Gott Isaaks und der Gott Jakobs, habe euch gesandt. Dieser Name ist für immer und ich werde mich an euch erinnern von Generation zu Generation.
Gemini	Und Gott sprach weiter zu Mose: So sollst du zu den Söhnen Israels sprechen: JHWH, der Gott eurer Väter, der Gott Abrahams, der Gott Isaaks und der Gott Jakobs, hat mich zu euch gesandt. Dies ist mein Name für ewig, und dies mein Gedenken von Generation zu Generation.
LLaMA3	Und sprach Gott zu Mose: So sprich zu den Söhnen Israels: Der HERR, der Gott eurer Väter, der Gott Abrahams, der Gott Isaaks und der Gott Jakobs, hat mich gesandt zu euch. Dies ist mein Name für immer, und dies ist mein Erinnerung für alle Generationen.

BASIS Weiter sprach Gott zu Mose: »Das sollst du den Israeliten sagen: ›Der Herr hat mich zu euch geschickt, der Gott eurer Väter Abraham, Isaak und Jakob.‹ So heiße ich schon immer, und so will ich bei all ihren Nachkommen genannt werden.

KJV And God said moreover unto Moses, Thus shalt thou say unto the children of Israel, The LORD God of your fathers, the God of Abraham, the God of Isaac, and the God of Jacob, hath sent me unto you: this is my name for ever, and this is my memorial unto all generations.

DeepL	and God said to Moses, "Thus shalt thou say unto thethe children of Israel: The LORD, the God of your fathers, the God of Abraham, the God of Isaac, and the God of Jacob, has sent me to youthis is my name forever, and this is my memorial unto all generations
NLLB-200	And God said also to Moses, Thus shalt thou say to the children of Israel: Jehovah, the God of your fathers, the God of Abraham, the God of Isaac, and the God of Jacob, hath sent me unto you. This is my name for ever, and this shall be remembered of me for all generations.
Gemini	And God said further to Moses: Thus you shall say to the sons of Israel: YHWH, the God of your fathers, the God of Abraham, the God of Isaac, and the God of Jacob, has sent me to you. This is my name forever, and this is my memorial to all generations.
LLaMA3	And God said to Moses, Thus you shall say to the children of Israel, The Lord God of your fathers, the God of Abraham, the God of Isaac, and the God of Jacob, has sent me to you. This is my name forever, and this is how I am to be remembered from generation to generation.

NLT God also said to Moses, “Say this to the people of Israel: Yahweh, the God of your ancestors—the God of Abraham, the God of Isaac, and the God of Jacob—has sent me to you. This is my eternal name, my name to remember for all generations.

Anhang 2-1 **Jesaja 7, 10** – Griechisch | Latein

BHS	וַיֹּוסֶף יְהוָה דַּבֵּר אֶל־אָחָז לֵאמֹר
unvokalisiert	ויוסף יהוה דבר אל אחז לאמר

LXX Καὶ προσέθετο κύριος λαλῆσαι τῷ Αχαζ λέγων

DeepL	---
NLLB-200	---
Gemini	Καὶ προσέθετο Κύριος λαλῆσαι τῷ Ἄχαζ λέγων
LLaMA3	Καὶ ἐλάλησεν Κύριος πρὸς Ὀζεία λεγὼν

VUL et adiecit Dominus loqui ad Ahaz dicens

DeepL	Et locutus est Dominus iterum ad Achaz, dicens:
NLLB-200	---
Gemini	Et adiecit Dominus loqui ad Achaz, dicens:
LLaMA3	Sed dixit Dominus ad Achaz dicens.

Anhang 2-2 Jesaja 7, 10 – Deutsch | Englisch

LUTH Und der Herr redete abermals zu Ahas und sprach:

DeepL	und der Herr sprach weiter zu Ahas und sagte
NLLB-200	Und Jehova sprach zu Ohaz:
Gemini	Und der HERR redete weiter zu Ahas und sprach:
LLaMA3	Und der Herr sprach zu Achab

BASIS Weiter ließ der Herr Ahas sagen:

KJV Moreover the LORD spake again unto Ahaz, saying,

DeepL	and the Lord spoke again to Ahaz, saying,
NLLB-200	And Jehovah added the word to Ahaz, saying,
Gemini	And the LORD spoke again to Ahaz, saying,
LLaMA3	And the Lord spoke to Ahaz saying.

NLT Later, the LORD sent this message to King Ahaz:

Anhang 2-3 Jesaja 7, 11 – Griechisch | Latein

BHS	שְׁאַל־לְךָ אוֹת מֵעִם יְהוָה אֱלֹהֶיךָ הַעְמֵק שְׁאָלָה אוֹ הַגְבֵּהַ לְמָעְלָה
unvokalisiert	שאל לך אות מעם יהוה אלהיך העמק שאלה או הגבה למעלה

LXX Αἴτησαι σεαυτῷ σημεῖον παρὰ κυρίου θεοῦ σου εἰς βάθος ἢ εἰς ὕψος.

DeepL	---
NLLB-200	---
Gemini	Αἴτησαι σεαυτῷ σημεῖον παρὰ κυρίου τοῦ θεοῦ σου· βάθυνον τὸ αἴτημα, ἢ ὕψωσον εἰς τὸ ἄνω.
LLaMA3	Σέο ἀπὸ Κύριον τὸν Θεόν σου ἐκτενέω ἢ ἐπὶ τὸ ὕψος.

VUL pete tibi signum a Domino Deo tuo in profundum inferni sive in excelsum supra

DeepL	Pete signum a Domino Deo tuo, sive in profundis sive in altis.
NLLB-200	---
Gemini	Pete tibi signum a Domino Deo tuo; fac petitionem profundam, aut exalta in excelsum.
LLaMA3	Soli tibi a Domino Deo tuo petimus auxilium, vel in profundum, vel in altum.

Anhang 2-4 Jesaja 7, 11 – Deutsch | Englisch

LUTH Fordere dir ein Zeichen vom Herrn, deinem Gott, es sei drunten in der Tiefe oder droben in der Höhe!

DeepL	frage den Herrn, deinen Gott, um ein Zeichen, sei es tief oder hoch
NLLB-200	Du sollst dich von dem Volk, von dem Jehova, dein Gott, bitten, die Tiefe zu bitten, oder die Tiefe zu bitten,
Gemini	Fordere dir ein Zeichen von Jahwe, deinem Gott; mache die Bitte tief oder mache sie hoch nach oben.
LLaMA3	Sage zu dir von dem Herrn deinem Gott aus dem Tiefland Sage zu dir oder höre auf hinauf.

BASIS »Bitte den Herrn, deinen Gott, zur Bestätigung um ein Zeichen. Du selbst kannst wählen, woher es kommen soll: vom Totenreich unten bis zum Himmel oben.«

KJV Ask thee a sign of the LORD thy God; ask it either in the depth, or in the height above.

DeepL	ask a sign from the Lord your God, whether it be deep or high
NLLB-200	Ask thee, O people, of Jehovah thy God, deeply, deeply, or deeply,
Gemini	Ask for yourself a sign from Yahweh your God; make the request deep, or make it high above.
LLaMA3	Seek for yourself from the Lord your God, and from the midst of your mighty ones, seek a small thing, or seek a great thing.

NLT “Ask the LORD your God for a sign of confirmation, Ahaz. Make it as difficult as you want—as high as heaven or as deep as the place of the dead.”

Anhang 2-5 Jesaja 7, 12 – Griechisch | Latein

BHS	וַיֹּאמֶר אָחָז לֹא־אֶשְׁאַל וְלֹא־אֲנַסֶּה אֶת־יְהוָה
unvokalisiert	ויאמר אחז לא אשאל ולא אנסה את יהוה

LXX καὶ εἶπεν Αχαζ Οὐ μὴ αἰτήσω οὐδ᾽ οὐ μὴ πειράσω κύριον.

DeepL	---
NLLB-200	---
Gemini	καὶ εἶπεν Αχας· Οὐ μὴ αἰτήσω, οὐδὲ πειράσω τὸν κύριον.
LLaMA3	Εἶπεν Ἀχάζ Οὐκ ἐρωτήσω καὶ οὐκ ἐπερωτήσω Κύριον.

VUL et dixit Ahaz non petam et non temptabo Dominum"

DeepL	Et Achaz dixit: "Non rogabo, neque probabo Dominum.”
NLLB-200	---
Gemini	Et dixit Achaz: Non petam, et non tentabo Dominum.
LLaMA3	Non rogabo, neque inquiram Dominum.

Anhang 2-6 Jesaja 7, 12 – Deutsch | Englisch

LUTH Aber Ahas sprach: Ich will’s nicht fordern, damit ich den Herrn nicht versuche.

DeepL	und Achaz sprach: Ich werde nicht fragen und den Herrn nicht versuchen
NLLB-200	Und Achaz sprach: Ich will nicht unrecht tun und nicht den HERRN vergessen.
Gemini	Und Ahas sprach: Ich will nicht bitten und will den HERRN nicht versuchen.
LLaMA3	Ich will nicht bitten und ich will nicht forschen nach dem HERRN.

BASIS Ahas aber antwortete: »Ich will nicht um ein Zeichen bitten, um den Herrn nicht auf die Probe zu stellen.«

KJV But Ahaz said, I will not ask, neither will I tempt the LORD.

DeepL	and Ahaz said, "I will not ask, nor will I test the Lord."
NLLB-200	And Ahaz said, I will not do wickedly, neither will I forget Jehovah.
Gemini	But Ahaz said, 'I will not ask, and I will not test the LORD.'
LLaMA3	And Azariah said I will not ask and I will not try the Lord.

NLT But the king refused. “No,” he said, “I will not test the LORD like that.”

Anhang 2-7 Jesaja 7, 13 – Griechisch | Latein

BHS	וַיֹּאמֶר שִׁמְעוּ־נָא בֵּית דָּוִד הַמְעַט מִכֶּם הַלְאוֹת אֲנָשִׁים כִּי תַלְאוּ גַּם אֶת־אֱלֹהָי
unvokalisiert	ויאמר שמעו נא בית דוד המעט מכם הלאות אנשים כי תלאו גם את אלהי

LXX καὶ εἶπεν Ἀκούσατε δή, οἶκος Δαυιδ· μὴ μικρὸν ὑμῖν ἀγῶνα παρέχειν ἀνθρώποις; καὶ πῶς κυρίῳ παρέχετε ἀγῶνα;

DeepL	---
NLLB-200	---
Gemini	Καὶ εἶπεν· Ἀκούσατε δή, οἶκος Δαυίδ. Μὴ μικρὸν ὑμῖν κοποῦν ἀνθρώπους, ὅτι κοποῦτε καὶ τὸν θεόν μου;
LLaMA3	Φωνήσας εἶπεν Σίμεων οἴκου Δαυίδ μὴ ὀλιγὼν ἀπὸ ὑμῶν ἄνδρες ὅτι καὶ τὸν Θεὸν ὑμᾶς ἀναλαμβάνετε.

VUL et dixit audite ergo domus David numquid parum vobis est molestos esse hominibus quia molesti estis et Deo meo

DeepL	Et dixit: Audite nunc, domus David: num parum vobis est quod homines laborantes facitis, ut etiam Deum meum laborantem faciatis?
NLLB-200	---
Gemini	Et dixit: Audite nunc, domus David. Numquid parum vobis est fatigare homines, quia fatigatis etiam Deum meum?
LLaMA3	Audite, domus David, paucos habetis homines, qui sustinent vos, etiam Deum sustinete.

Anhang 2-8 Jesaja 7, 13 – Deutsch | Englisch

LUTH Da sprach Jesaja: Wohlan, so hört, ihr vom Hause David: Ist’s euch zu wenig, dass ihr Menschen müde macht? Müsst ihr auch meinen Gott müde machen?

DeepL	und er sprach: Hört nun, ihr vom Hause Davids! Ist es euch zu wenig, Menschen zu verfolgen, dass ihr auch meinen Gott verfolgt?
NLLB-200	Und er sprach: Hört zu, ihr Haus Davids, daß euch die Schwäche der Menschen fehlt, die auch den Gott erfüllt.
Gemini	Und er sprach: Hört nun, Haus David! Ist es euch zu wenig, Menschen zu ermüden, dass ihr auch meinen Gott ermüdet?
LLaMA3	Hört zu, Haus Davids, nur wenige von euch haben mich gesehen, dass auch die Männer ihre Götter lasten.

BASIS Da sprach Jesaja: »Hört, ihr vom Haus David! Genügt es euch nicht, Menschen müde zu machen? Müsst ihr auch noch meinen Gott müde machen?

KJV And he said, Hear ye now, O house of David; Is it a small thing for you to weary men, but will ye weary my God also?

DeepL	and he said, Hearken now, O house of David: Is it too little for you to weary men, that ye weary my God also?
NLLB-200	And he said, Hearken now, O house of David, where ye are lacking men, that ye may fill even the house of God.
Gemini	And he said, 'Hear now, O house of David! Is it too small a thing for you to weary men, that you weary my God also?'
LLaMA3	And he said Hear now, the house of David, a little while ago you were not as many as the days of old, nor have they increased to the days of yesterday.

NLT Then Isaiah said, “Listen well, you royal family of David! Isn’t it enough to exhaust human patience? Must you exhaust the patience of my God as well?

Anhang 2-9 Jesaja 7, 14– Griechisch | Latein

BHS	לָכֵן יִתֵּן אֲדֹנָי הִוּא לָכֶם אוֹת הִנֵּה הָעַלְמָה הָרָה וְיֹלֶדֶת בֵּן וְקָרָאת שְׁמוֹ עִמָּנוּ אֵל
unvokalisiert	לכן יתן אדני הוא לכם אות הנה העלמה הרה וילדת בן וקראת שמו עמנו אל

LXX διὰ τοῦτο δώσει κύριος αὐτὸς ὑμῖν σημεῖον· ἰδοὺ ἡ παρθένος ἐν γαστρὶ ἕξει καὶ τέξεται υἱόν, καὶ καλέσεις τὸ ὄνομα αὐτοῦ Εμμανουηλ

DeepL	---
NLLB-200	---
Gemini	Διὰ τοῦτο δώσει Κύριος αὐτὸς ὑμῖν σημεῖον· ἰδοὺ ἡ παρθένος ἐν γαστρὶ ἕξει καὶ τέξεται υἱόν, καὶ καλέσει τὸ ὄνομα αὐτοῦ Ἐμμανουήλ.
LLaMA3	Τότε δώσει Κύριος ὁ Θεός ἡμῖν σημεῖον ἰδοὺ ἡ παρθένος ἡ ἔχουσα καὶ τίκτουσα υἱόν καὶ καλέσει τὸ ὄνομα αὐτοῦ μεθ' ἡμῶν ἐπὶ ἡμᾶς

VUL propter hoc dabit Dominus ipse vobis signum ecce virgo concipiet et pariet filium et vocabitis nomen eius Emmanuhel

DeepL	Propter hoc dabit Dominus vobis signum: ecce, virgo concipiet et pariet filium, et vocabunt nomen eius Emmanuel.
NLLB-200	---
Gemini	Propterea dabit Dominus ipse vobis signum: Ecce, virgo concipiet, et pariet filium, et vocabit nomen eius Emmanuel.
LLaMA3	Itaque dabit Dominus noster vobis signum Ecce mulier gravida et pariet filium et vocabitur nomen eius Immanu El.

Anhang 2-10 Jesaja 7, 14 – Deutsch | Englisch

LUTH Darum wird euch der Herr selbst ein Zeichen geben: Siehe, eine Jungfrau ist schwanger und wird einen Sohn gebären, den wird sie nennen Immanuel.

DeepL	darum wird euch der Herr ein Zeichen geben: Siehe, die Jungfrau wird schwanger werden und einen Sohn gebären, und sie wird ihm den Namen Immanuel geben
NLLB-200	So soll der Herr dir geben, daß sie, die du siehst, schwanger wird und einen Sohn gebären und seinen Namen Gott mit uns nennen wird.
Gemini	Darum wird der Herr selbst euch ein Zeichen geben: Siehe, die junge Frau ist schwanger und wird einen Sohn gebären, und sie wird seinen Namen Immanuel nennen.
LLaMA3	Dann wird mein Herr geben euch ein Zeichen siehe die Frau ist schwanger und gebiert einen Sohn und rufen ihn Namen mit uns ist er

BASIS Darum wird euch der Herr von sich aus ein Zeichen geben. Ihr werdet sehen: Die junge Frau wird schwanger werden und einen Sohn zur Welt bringen. Den wird sie Immanuel, ›Gott mit uns‹, nennen.

KJV Therefore the Lord himself shall give you a sign; Behold, a virgin shall conceive, and bear a son, and shall call his name Immanuel.

DeepL	therefore, the Lord will give you a sign: Behold, the virgin shall conceive and bear a son, and shall call his name Immanuel
NLLB-200	And now, behold, the Lord hath given thee the widow, and she shall conceive, and bear a son, and call his name, Jehovah our God.
Gemini	Therefore the Lord himself will give you a sign: Behold, the young woman is pregnant and shall bear a son, and she shall call his name Immanuel.
LLaMA3	Therefore, my Lord will give you a sign: Behold, the virgin shall conceive and bear a son, and you shall call his name Immanuel.

NLT All right then, the Lord himself will give you the sign. Look! The virgin will conceive a child! She will give birth to a son and will call him Immanuel.

Anhang 2-11 Jesaja 7, 15 – Griechisch | Latein

BHS	חֶמְאָה וּדְבַשׁ יֹאכֵל לְדַעְתּוֹ מָאוֹס בָּרָע וּבָחוֹר בַּטּוֹב
unvokalisiert	חמאה ודבש יאכל לדעתו מאוס ברע ובחור בטוב

LXX βούτυρον καὶ μέλι φάγεται· πρὶν ἢ γνῶναι αὐτὸν ἢ προελέσθαι πονηρὰ ἐκλέξεται τὸ ἀγαθόν

DeepL	---
NLLB-200	---
Gemini	βούτυρον καὶ μέλι φάγεται, τοῦ γνῶναι αὐτὸν ἀπωθεῖσθαι πονηρὸν καὶ αἱρεῖσθαι ἀγαθόν.
LLaMA3	Γλυκύ καὶ μέλισσα ἐσθίειν γινώσκων ἀπείθαρχος ἐν τῷ κακῷ καὶ ἐν τῷ ἀγαθῷ

VUL butyrum et mel comedet ut sciat reprobare malum et eligere bonum

DeepL	Butyrum et mel edet, malum reiciens et bonum eligens.
NLLB-200	---
Gemini	Butyrum et mel comedet, ut sciat reprobare malum et eligere bonum.
LLaMA3	Mellis et melle dulcis manducat, scire facit odium et amorem.

Anhang 2-12 Jesaja 7, 15 – Deutsch | Englisch

LUTH Butter und Honig wird er essen, bis er weiß, Böses zu verwerfen und Gutes zu erwählen.

DeepL	er wird Butter und Honig essen, um zu erkennen, was gut ist, und das Böse verachten und das Gute wählen
NLLB-200	Er wird grünes und weiches essen, um zu erkennen, was gut ist und was schlecht.
Gemini	Dickmilch und Honig wird er essen, bis er weiß, das Böse zu verwerfen und das Gute zu wählen.
LLaMA3	Honig und Sirup isst er, wenn er weiß, was ist böse und was ist gut.

BASIS Butter und Honig wird er essen, bis er Böses und Gutes unterscheiden kann.

KJV Butter and honey shall he eat, that he may know to refuse the evil, and choose the good.

DeepL	he shall eat butter and honey, knowing how to reject evil and choose good
NLLB-200	He shall eat raw and dried up, to discern between the evil and the good.
Gemini	Curds and honey he will eat, until he knows to refuse the evil and choose the good.
LLaMA3	Honey and honeycomb, he shall eat to know evil and to choose the bad.

NLT By the time this child is old enough to choose what is right and reject what is wrong, he will be eating yogurt and honey.

Anhang 2-13 Jesaja 7, 16 – Griechisch | Latein

BHS	כִּי בְּטֶרֶם יֵדַע הַנַּעַר מָאֹס בָּרָע וּבָחֹר בַּטּוֹב תֵּעָזֵב הָאֲדָמָה אֲשֶׁר אַתָּה קָץ מִפְּנֵי שְׁנֵי מְלָכֶיהָ
unvokalisiert	כי בטרם ידע הנער מאס ברע ובחר בטוב תעזב האדמה אשר אתה קץ מפני שני מלכיה

LXX διότι πρὶν ἢ γνῶναι τὸ παιδίον ἀγαθὸν ἢ κακὸν ἀπειθεῖ πονηρίᾳ τοῦ ἐκλέξασθαι τὸ ἀγαθόν, καὶ καταλειφθήσεται ἡ γῆ, ἣν σὺ φοβῇ ἀπὸ προσώπου τῶν δύο βασιλέων.

DeepL	---
NLLB-200	---
Gemini	διότι πρὶν ἢ γνῶναι τὸ παιδίον ἀπωθεῖσθαι τὸ κακὸν καὶ ἐκλέγεσθαι τὸ ἀγαθόν, ἐγκαταλειφθήσεται ἡ γῆ, ἣν σὺ βδελύσσῃ ἀπὸ προσώπου τῶν δύο βασιλέων αὐτῆς.
LLaMA3	ὅτι πρὸ τοῦ γνῶναι τὸν παῖδα τὸ κακὸν καὶ τὸ ἀγαθὸν ἀποστραφεῖ ἡ γῆ ἣν ἔκτισας ἐκ πλευρᾶς ἀπὸ πρὸς πρὸς ἡμέρας βασιλέων αὐτῆς

VUL quia antequam sciat puer reprobare malum et eligere bonum derelinquetur terra quam tu detestaris a facie duum regum suorum

DeepL	am antequam puer sciat malum reicere et bonum eligere, terra quam relinquitis derelinquetur propter duos reges suos.
NLLB-200	---
Gemini	Quia antequam sciat puer reprobare malum et eligere bonum, derelinquetur terra, quam tu detestaris, a facie duorum regum eius.
LLaMA3	Cum antequam sciat puer malum et bonum, relinquet terra quam tu creasti a facie annorum regni ejus.

Anhang 2-14 Jesaja 7, 16 – Deutsch | Englisch

LUTH Denn ehe der Knabe weiß, Böses zu verwerfen und Gutes zu erwählen, wird das Land verödet sein, vor dessen zwei Königen dir graut.

DeepL	denn bevor der Junge das Böse verabscheut und das Gute wählt, wird das Land, das du vor seinen beiden Königen gefürchtet hast, verlassen werden
NLLB-200	Denn bevor der Junge weiß, was gut und was schlecht ist, wird das Land, vor dem du dich vernichtest, verlassen sein vor den beiden Königen.
Gemini	Denn bevor der Knabe weiß, das Böse zu verwerfen und das Gute zu erwählen, wird das Land, vor dessen beiden Königen dir graut, verlassen sein.
LLaMA3	Denn bevor der Knabe weiß, was Böses ist, und bevor er weiß, was Gut ist, wird die Erde, die du verflucht hast, verlassen werden, die du verflucht hast, aufgrund der Jahre ihrer Königreiche.

BASIS Denn verwüstet wird das Land der beiden Könige, die dich in Angst und Schrecken versetzen. Das wird geschehen, bevor der Junge Böses und Gutes unterscheiden kann.

KJV For before the child shall know to refuse the evil, and choose the good, the land that thou abhorrest shall be forsaken of both her kings.

DeepL	for before the boy knows how to reject the evil and choose the good, the land which you are dispossessing will be forsaken, because of the two kings of it
NLLB-200	For before the child knoweth the difference between the wicked and the righteous, the land which thou hast destroyed from the presence of the two kings thereof shall be forsaken.
Gemini	For before the boy knows to refuse the evil and choose the good, the land that you dread because of its two kings will be forsaken.
LLaMA3	For before the boy knows how to refuse the evil and choose the good, the land whose God is the Lord will be destroyed from being a land.

NLT For before the child is that old, the lands of the two kings you fear so much will both be deserted.

Anhang 3-1 Psalm 23, 1 – Griechisch | Latein

BHS	מִזְמוֹר לְדָוִד יְהוָה רֹעִי לֹא אֶחְסָר
unvokalisiert	מזמור לדוד יהוה רעי לא אחסר

LXX Ψαλμὸς τῷ Δαυιδ. Κύριος ποιμαίνει με, καὶ οὐδέν με ὑστερήσει.

DeepL	---
NLLB-200	---
Gemini	Ψαλμὸς τῷ Δαυίδ. Κύριος ποιμαίνει με, καὶ οὐδέν με ὑστερήσει.
LLaMA3	Ποίημα τῷ Δαυίδ· Κύριος ὁ ποιμήν μου οὐκ ἐκκλέψω.

VUL canticum David Dominus pascit me nihil mihi deerit

DeepL	Psalmus Davidis. Dominus est pastor meus; nihil mihi deerit.
NLLB-200	---
Gemini	Psalmus David. Dominus regit me, et nihil mihi deerit.
LLaMA3	Dominus pastor meus, non deerit.

Anhang 3-2 Psalm 23, 1 – Deutsch | Englisch

LUTH Ein Psalm Davids. Der Herr ist mein Hirte, mir wird nichts mangeln.

DeepL	ein Psalm Davids. Der Herr ist mein Hirte, mir wird nichts mangeln
NLLB-200	Psalm Davids. Der Herr ist mein Hirte. Ich werde nicht fehlen.
Gemini	Ein Psalm Davids. Der HERR ist mein Hirte, mir wird nichts mangeln.
LLaMA3	Psalmlied für David Der Herr ist mein Hirte, ich werde nicht hungernd sein.

BASIS EIN PSALM, VON DAVID. Der Herr ist mein Hirte. Mir fehlt es an nichts.

KJV A Psalm of David. The LORD is my shepherd; I shall not want.

DeepL	a psalm of David. The Lord is my shepherd; I shall not want
NLLB-200	{A Psalm of David.} Jehovah is my shepherd; I shall not want.
Gemini	A Psalm of David. The LORD is my shepherd; I shall not want.
LLaMA3	A Psalm of David. The Lord is my shepherd, I shall not want.

NLT The LORD is my shepherd; I have all that I need.

Anhang 3-3 Psalm 23, 2 – Griechisch | Latein

BHS	בִּנְאוֹת דֶּשֶׁא יַרְבִּיצֵנִי עַל־מֵי מְנֻחוֹת יְנַהֲלֵנִי
unvokalisiert	בנאות דשא ירביצני על מי מנחות ינהלני

LXX εἰς τόπον χλόης, ἐκεῖ με κατεσκήνωσεν, ἐπὶ ὕδατος ἀναπαύσεως ἐξέθρεψέν με,

DeepL	---
NLLB-200	---
Gemini	ἐν τόπῳ χλόης ἀνακλίνει με, ἐπὶ ὑδάτων ἀναπαύσεως ἄγει με.
LLaMA3	Γέενταί δέσποτα ἐρβίτζενί μοι ἐπὶ ὕδασι νηχῶν ἡλαλήνεί

VUL in pascuis herbarum adclinavit me super aquas refectionis enutrivit me

DeepL	Facit me iacere in viridibus pascuis; ducit me ad aquas quietas.
NLLB-200	---
Gemini	In pascuis virentibus me collocat, ad aquas requietis deducit me.
LLaMA3	In aedibus Dei sedebit me in rivis quietis inebriabit me.

Anhang 3-4 Psalm 23, 2 – Deutsch | Englisch

LUTH Er weidet mich auf einer grünen Aue und führet mich zum frischen Wasser.

DeepL	er lässt mich an stillen Wassern lagern, er führt mich
NLLB-200	Ich will dich nicht mehr sehen, ich will dich nicht mehr sehen.
Gemini	Auf grasreichen Weiden lässt er mich lagern, an Wassern der Ruhe leitet er mich.
LLaMA3	Ich werde in grünen Hainen umhergetragen werden auf den Wassern der Ruhe mich erquicken lassen.

BASIS Auf saftig grünen Weiden lässt er mich lagern. Er leitet mich zu Ruheplätzen am Wasser,

KJV He maketh me to lie down in green pastures: he leadeth me beside the still waters.

DeepL	he makes me lie down in green pastures; He leads me beside quiet waters
NLLB-200	In the sight of the Lord, he will multiply me above the waters of the waters of life.
Gemini	He makes me lie down in green pastures; he leads me beside still waters.
LLaMA3	In the shadow of your wings I will rejoice; upon the waters of your sanctuary I will meditate.

NLT He lets me rest in green meadows; he leads me beside peaceful streams.

Anhang 3-5 Psalm 23, 3 – Griechisch | Latein

BHS	נַפְשִׁי יְשׁוֹבֵב יַנְחֵנִי בְמַעְגְּלֵי־צֶדֶק לְמַעַן שְׁמוֹ
unvokalisiert	נפשי ישובב ינחני במעגלי צדק למען שמו

LXX τὴν ψυχήν μου ἐπέστρεψεν. ὡδήγησέν με ἐπὶ τρίβους δικαιοσύνης ἕνεκεν τοῦ ὀνόματος αὐτοῦ.

DeepL	---
NLLB-200	---
Gemini	τὴν ψυχήν μου ἐπέστρεψεν, ὡδήγησέν με ἐπὶ τρίβους δικαιοσύνης ἕνεκεν τοῦ ὀνόματος αὐτοῦ.
LLaMA3	Πνεῦμα μου ἐπιστρέφει καὶ ὁδηγεῖ με ἐν ὁδῷ δικαιοσύνης ἵνα κληρὸς ὄνομα αὐτοῦ ᾗ

VUL animam meam refecit duxit me per semitas iustitiae propter nomen suum

DeepL	nima mea revertetur et reducet me in circulos iustitiae propter nomen eius.
NLLB-200	---
Gemini	Animam meam convertit, deduxit me super semitas iustitiae propter nomen suum.
LLaMA3	Anima mea reuertetur in semitis iustitiae, ad nomen eius.

Anhang 3-6 **Psalm 23, 3** – Deutsch | Englisch

LUTH Er erquicket meine Seele. Er führet mich auf rechter Straße um seines Namens willen.

DeepL	eine Seele wird mich in Kreisen der Gerechtigkeit führen, um seines Namens willen
NLLB-200	Mein Leben ist wieder zurückgekehrt. Er führt mich auf den Wegen der Gerechtigkeit, um seines Namens willen.
Gemini	Er stellt meine Seele wieder her, er leitet mich auf Pfaden der Gerechtigkeit um seines Namens willen.
LLaMA3	Meine Seele wird mich umkehren, mich leiten auf Pfaden der Gerechtigkeit, um Seinen Namen.

BASIS dort erfrischt er meine Seele. Er führt mich gerecht durchs Leben. Dafür steht er mit seinem Namen ein.

KJV He restoreth my soul: He leadeth me in the paths of righteousness for his name's sake.

DeepL	my soul shall return and guide me in the circles of righteousness for the sake of His name
NLLB-200	My soul turneth round. He leadeth me in the paths of righteousness for his name's sake.
Gemini	He restores my soul; he leads me in paths of righteousness for his name's sake.
LLaMA3	My soul turns back to Your judgments in the paths of righteousness for the sake of Your name.

NLT He renews my strength. He guides me along right paths, bringing honor to his name.

Anhang 3-7 Psalm 23, 4 – Griechisch | Latein

BHS	גַּם כִּי-אֵלֵךְ בְּגֵיא צַלְמָוֶת לֹא-אִירָא רָע כִּי-אַתָּה עִמָּדִי שִׁבְטְךָ וּמִשְׁעַנְתֶּךָ הֵמָּה יְנַחֲמֵנִי
unvokalisiert	גם כי אלך בגיא צלמות לא אירא רע כי אתה עמדי שבטך ומשענתך המה ינחמני

LXX ἐὰν γὰρ καὶ πορευθῶ ἐν μέσῳ σκιᾶς θανάτου, οὐ φοβηθήσομαι κακά, ὅτι σὺ μετ᾽ ἐμοῦ εἶ· ἡ ῥάβδος σου καὶ ἡ βακτηρία σου, αὐταί με παρεκάλεσαν.

DeepL	---
NLLB-200	---
Gemini	καὶ γὰρ ἐὰν πορευθῶ ἐν φάραγγι σκιᾶς θανάτου, οὐ φοβηθήσομαι κακόν, ὅτι σὺ μετ’ ἐμοῦ εἶ· ἡ ῥάβδος σου καὶ ἡ βακτηρία σου, αὗταί με παρακαλοῦσιν.
LLaMA3	Καὶ ἐὰν ἔλθω εἰς τὸν χῶρον τοῦ θανάτου οὐ φοβήσομαι τὸ πονηρόν ὅτι σὺ μετὰ ἐμαυτοῦ ἡ σέβασμα σου καὶ ἡ σωτηρία σου ἡ αὐταὶ θάλπουσίν με.

VUL sed et si ambulavero in valle mortis non timebo malum quoniam tu mecum es virga tua et baculus tuus ipsa consolabuntur me

DeepL	Etsi ambulaverim per vallem umbrae mortis, non timebo malum, quia tu mecum es; virga tua et baculus tuus, me consolantur.
NLLB-200	---
Gemini	Nam etsi ambulavero in valle umbrae mortis, non timebo malum, quoniam tu mecum es; virga tua et baculus tuus, ipsa me consolantur.
LLaMA3	Etiam si ambulavero in umbra mortis non timebo malum quia tu mecum es et baculus tuus confortat me.

Anhang 3-8 Psalm 23, 4 – Deutsch | Englisch

LUTH Und ob ich schon wanderte im finstern Tal, fürchte ich kein Unglück; denn du bist bei mir, dein Stecken und Stab trösten mich.

DeepL	auch wenn ich durch die Tal der Schatten des Todes gehe, fürchte ich kein Übel, denn du bist bei mir, dein Stab und dein Stecken trösten mich
NLLB-200	Du bist mein Schutz und mein Erlöser, du tröstest mich."
Gemini	Auch wenn ich gehe im Tal des Todesschattens, fürchte ich kein Unheil, denn du bist bei mir; dein Stecken und dein Stab, sie trösten mich.
LLaMA3	Weil auch ich wandere durch das Tal der Schatten des Todes, fürchte ich kein Unglück, denn du bist bei mir, dein Stab und deine Steigerung, die mich trösten.

BASIS Und muss ich durch ein finsteres Tal, fürchte ich kein Unglück. Denn du bist an meiner Seite! Dein Stock und dein Stab schützen und trösten mich.

KJV Yea, though I walk through the valley of the shadow of death, I will fear no evil: for thou art with me; Thy rod and thy staff they comfort me.

DeepL	even though I walk through the valley of the shadow of death, I will fear no evil, for you are with me; your rod and your staff, they comfort me
NLLB-200	Though I walk in the midst of trouble, I see no evil; for thou art my shield, and my deliverer: thou comforts me.
Gemini	Even though I walk through the valley of the shadow of death, I will fear no evil, for you are with me; your rod and your staff, they comfort me.
LLaMA3	Even when I walk through the valley of death, I will not fear evil, for You are with me, Your rod and Your staff, they comfort me.

NLT Even when I walk through the darkest valley, I will not be afraid, for you are close beside me. Your rod and your staff protect and comfort me.

Anhang 3-9 Psalm 23, 5 – Griechisch | Latein

BHS	תַּעֲרֹךְ לְפָנַי שֻׁלְחָן נֶגֶד צֹרְרָי דִּשַּׁנְתָּ בַשֶּׁמֶן רֹאשִׁי כּוֹסִי רְוָיָה
unvokalisiert	תערך לפני שלחן נגד צררי דשנת בשמן ראשי כוסי רויה

LXX ἡτοίμασας ἐνώπιόν μου τράπεζαν ἐξ ἐναντίας τῶν θλιβόντων με· ἐλίπανας ἐν ἐλαίῳ τὴν κεφαλήν μου, καὶ τὸ ποτήριόν σου μεθύσκον ὡς κράτιστον.

DeepL	---
NLLB-200	---
Gemini	Ἑτοιμάζεις ἐνώπιόν μου τράπεζαν ἔναντι τῶν ἐχθρῶν μου· ἐλίπανας ἐν ἐλαίῳ τὴν κεφαλήν μου· τὸ ποτήριόν μου ὑπερχεῖται.
LLaMA3	Και ἔθηκα πρὸ προσώπου μου πίνακα ἐνώπιον ἐχθρῶν μου ἐν ἐλαίῳ κεφαλῆς μου ἐχρῖσμά μου ἐν ἐλαιῷ

VUL pones coram me mensam ex adverso hostium meorum inpinguasti oleo caput meum calix meus inebrians

DeepL	Parasti mensam ante me in conspectu inimicorum meorum; unxisti caput meum oleo; et poculum meum superfluit.
NLLB-200	---
Gemini	Paras coram me mensam adversus inimicos meos; unxisti oleo caput meum; calix meus redundat.
LLaMA3	Preparabo mensam coram me ante adversarios meos, et in oleo capiti meo ungam capillum meum.

Anhang 3-10 Psalm 23, 5 – Deutsch | Englisch

LUTH Du bereitest vor mir einen Tisch im Angesicht meiner Feinde. Du salbest mein Haupt mit Öl und schenkest mir voll ein.

DeepL	du bereitest mir einen Tisch vor meinen Feinden, du salbst mein Haupt mit Öl, mein Becher fließt über
NLLB-200	Du hast mir einen schönen Schwanz gegeben, und meine Schnauze ist voll von Schmerzen.
Gemini	Du bereitest vor mir einen Tisch angesichts meiner Feinde; du hast mein Haupt mit Öl gesalbt, mein Becher fließt über.
LLaMA3	Ich richte ein Tisch vor dir, vor meinem Feind, mit Öl bestreut, mein Haupt hat ein Ölbad.

BASIS Du deckst für mich einen Tisch vor den Augen meiner Feinde. Du salbst mein Haar mit duftendem Öl und füllst mir den Becher bis zum Rand.

KJV Thou preparest a table before me in the presence of mine enemies: Thou anointest my head with oil; my cup runneth over.

DeepL	you prepare a table before me in the presence of my enemies; You anoint my head with oil; my cup overflows
NLLB-200	Thou hast anointed my face with oil, my head with a pure oil.
Gemini	You prepare a table before me in the presence of my enemies; you have anointed my head with oil; my cup overflows.
LLaMA3	You shall spread a table before me in my presence, with my enemies as a sacrifice, and with the fat of my head running over my cup.

NLT You prepare a feast for me in the presence of my enemies. You honor me by anointing my head with oil. My cup overflows with blessings.

Anhang 3-11 Psalm 23, 6 – Griechisch | Latein

BHS	אַךְ טוֹב וָחֶסֶד יִרְדְּפוּנִי כָּל-יְמֵי חַיָּי וְשַׁבְתִּי בְּבֵית-יְהוָה לְאֹרֶךְ יָמִים
unvokalisiert	אך טוב וחסד ירדפוני כל ימי חיי ושבתי בבית יהוה לאורך ימים

LXX καὶ τὸ ἔλεός σου καταδιώξεταί με πάσας τὰς ἡμέρας τῆς ζωῆς μου, καὶ τὸ κατοικεῖν με ἐν οἴκῳ κυρίου εἰς μακρότητα ἡμερῶν.

DeepL	---
NLLB-200	---
Gemini	ἀγαθωσύνη καὶ ἔλεος καταδιώξουσίν με πάσας τὰς ἡμέρας τῆς ζωῆς μου, καὶ κατοικήσω ἐν οἴκῳ κυρίου εἰς μακρότητα ἡμερῶν.
LLaMA3	Καλὸν καὶ χάριν διώξω με πάντα τὰ ἡμέραι τῆς ζωῆς μου καὶ ἐν οἴκῳ κυρίου εἰς τὸν αἰῶνα.

VUL sed et benignitas et misericordia subsequetur me omnibus diebus vitae meae et habitabo in domo Domini in longitudine dierum

DeepL	Bona autem et misericordia sequentur me omnibus diebus vitae meae, et in domo Domini habitabo in aeternum.
NLLB-200	---
Gemini	Certe bonitas et misericordia subsequentur me omnibus diebus vitae meae, et habitabo in domo Domini in longitudinem dierum.
LLaMA3	Bonum et misericordia consequentur me omni die vita mea, et in domo Domini in aeternum habitabo.

Anhang 3-12 Psalm 23, 6 – Deutsch | Englisch

LUTH Gutes und Barmherzigkeit werden mir folgen mein Leben lang, und ich werde bleiben im Hause des Herrn immerdar.

DeepL	aber Güte und Gnade werden mich verfolgen alle Tage meines Lebens, und ich werde bleiben im Hause des Herrn immerdar
NLLB-200	Nur Gnade und Güte werden mich alle Tage meines Lebens begleiten, und ich werde ewig in dem Hause Jehovas wohnen.
Gemini	Fürwahr, Güte und Gnade werden mir folgen alle Tage meines Lebens, und ich werde wohnen im Haus des HERRN für die Länge der Tage.
LLaMA3	Aber Güte und Barmherzigkeit verfolgen mich alle Tage meines Lebens und ich wohne im Hause Jahwes für immer.

BASIS Nichts als Liebe und Güte begleiten mich alle Tage meines Lebens. Mein Platz ist im Haus des Herrn. Dort möchte ich mein Leben lang sein.

KJV Surely goodness and mercy shall follow me all the days of my life: And I will dwell in the house of the LORD for ever.

DeepL	but goodness and mercy shall follow me all the days of my life, and I shall dwell in the house of the Lord for ever
NLLB-200	Surely goodness and mercy shall follow me all the days of my life; and I will dwell in the house of Jehovah for ever.
Gemini	Surely goodness and steadfast love shall pursue me all the days of my life, and I shall dwell in the house of the LORD for length of days.
LLaMA3	Good and kindness shall follow me all the days of my life, and I shall dwell in the house of the Lord for length of days.

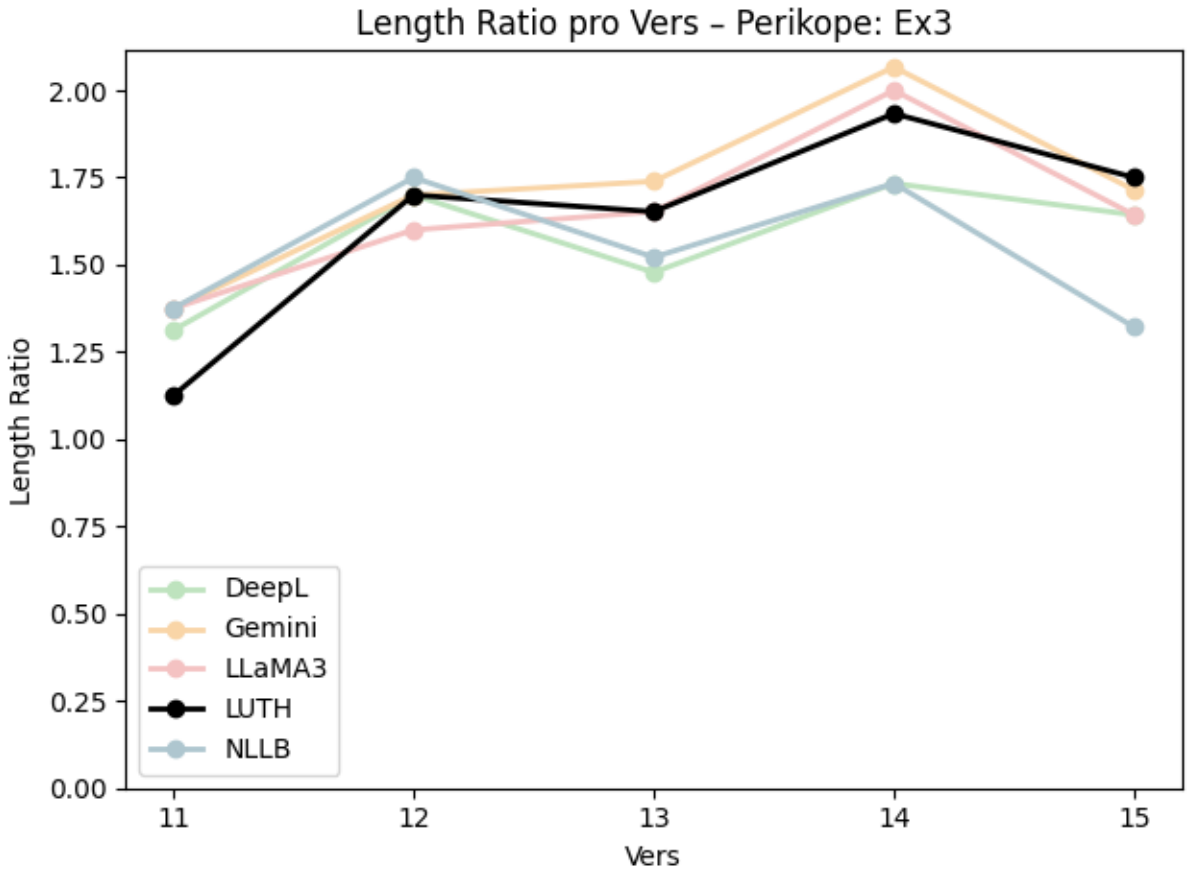
NLT Surely your goodness and unfailing love will pursue me all the days of my life, and I will live in the house of the LORD forever.

Anhang 4-1 Metriken | Exodus 3, 11 – 15

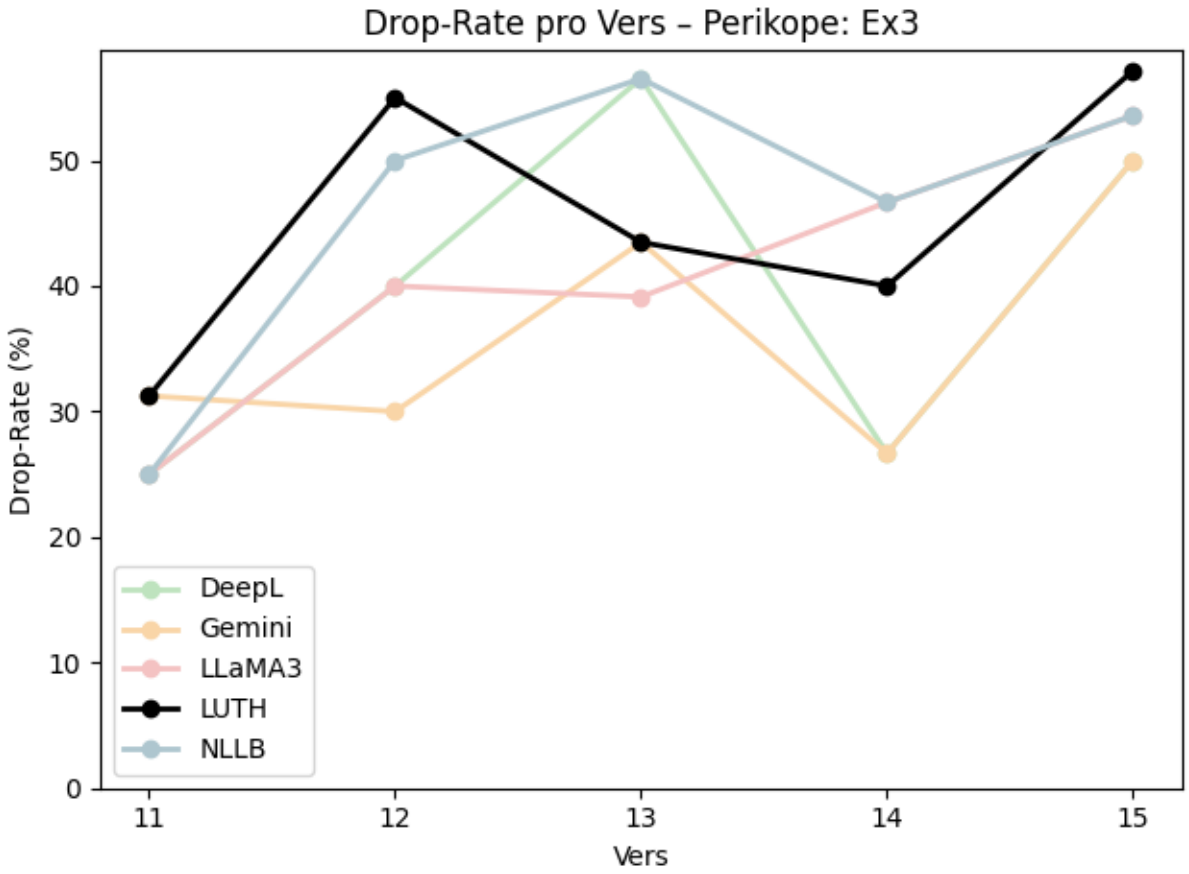
Perikope	**Buch**	**Kapitel**	**Vers**	**Modell**	**Wortzahl** (Quelle)	**Wortzahl** (Ziel)	**Längenverhältnis**	**Subtokenanzahl**	**TFR**	
pericope	book	chapter	verse	model	src_word_count	dst_word_count	length_ratio	subtoken_count	act_tfr	
Ex3,11-15	Exodus	3	11	DeepL	16	21	1.3125	-	-	
Ex3,11-15	Exodus	3	12	DeepL	20	34	1.7000	-	-	
Ex3,11-15	Exodus	3	13	DeepL	23	34	1.4783	-	-	
Ex3,11-15	Exodus	3	14	DeepL	15	26	1.7333	-	-	
Ex3,11-15	Exodus	3	15	DeepL	28	46	1.6429	-	-	
Ex3,11-15	Exodus	3	11	NLLB	16	22	1.3750	29	1.3182	
Ex3,11-15	Exodus	3	12	NLLB	20	35	1.7500	47	1.3429	
Ex3,11-15	Exodus	3	13	NLLB	23	35	1.5217	49	1.4000	
Ex3,11-15	Exodus	3	14	NLLB	15	26	1.7333	37	1.4231	
Ex3,11-15	Exodus	3	15	NLLB	28	37	1.3214	57	1.5405	
Ex3,11-15	Exodus	3	11	Gemini	16	22	1.3750	-	-	
Ex3,11-15	Exodus	3	12	Gemini	20	34	1.7000	-	-	
Ex3,11-15	Exodus	3	13	Gemini	23	40	1.7391	-	-	
Ex3,11-15	Exodus	3	14	Gemini	15	31	2.0667	-	-	
Ex3,11-15	Exodus	3	15	Gemini	28	48	1.7143	-	-	
Ex3,11-15	Exodus	3	11	LLaMA3	16	22	1.3750	35	1.5909	
Ex3,11-15	Exodus	3	12	LLaMA3	20	32	1.6000	44	1.3750	
Ex3,11-15	Exodus	3	13	LLaMA3	23	38	1.6522	58	1.5263	
Ex3,11-15	Exodus	3	14	LLaMA3	15	30	2.0000	48	1.6000	
Ex3,11-15	Exodus	3	15	LLaMA3	28	46	1.6429	74	1.6087	

	fehlend im Quelltext (Auslassungen)	**fehlend im Zieltext** (Halluzinationen)	**Quell-Ziel-Wortpaare**	**Halluzinationsrate**	**Anteil: fehlend im Quelltext**	**Modell**	**Stelle**
	src_unalign_count	dst_unalign_count	align_pair_count	hallucination_rate	drop_rate	model	passage
	4	9	12	42.86 %	25.00 %	DeepL	Ex3,11
	8	22	12	64.71 %	40.00 %	DeepL	Ex3,12
	13	24	10	70.59 %	56.52 %	DeepL	Ex3,13
	4	15	11	57.69 %	26.67 %	DeepL	Ex3,14
	14	32	14	69.57 %	50.00 %	DeepL	Ex3,15
	4	10	12	45.45 %	25.00 %	NLLB	Ex3,11
	10	25	10	71.43 %	50.00 %	NLLB	Ex3,12
	13	25	10	71.43 %	56.52 %	NLLB	Ex3,13
	7	18	8	69.23 %	46.67 %	NLLB	Ex3,14
	15	24	13	64.86 %	53.57 %	NLLB	Ex3,15
	5	11	11	50.00 %	31.25 %	Gemini	Ex3,11
	6	20	14	58.82 %	30.00 %	Gemini	Ex3,12
	10	27	13	67.50 %	43.48 %	Gemini	Ex3,13
	4	20	11	64.52 %	26.67 %	Gemini	Ex3,14
	14	34	14	70.83 %	50.00 %	Gemini	Ex3,15
	4	10	12	45.45 %	25.00 %	LLaMA3	Ex3,11
	8	20	12	62.50 %	40.00 %	LLaMA3	Ex3,12
	9	24	14	63.16 %	39.13 %	LLaMA3	Ex3,13
	7	22	8	73.33 %	46.67 %	LLaMA3	Ex3,14
	15	33	13	71.74 %	53.57 %	LLaMA3	Ex3,15

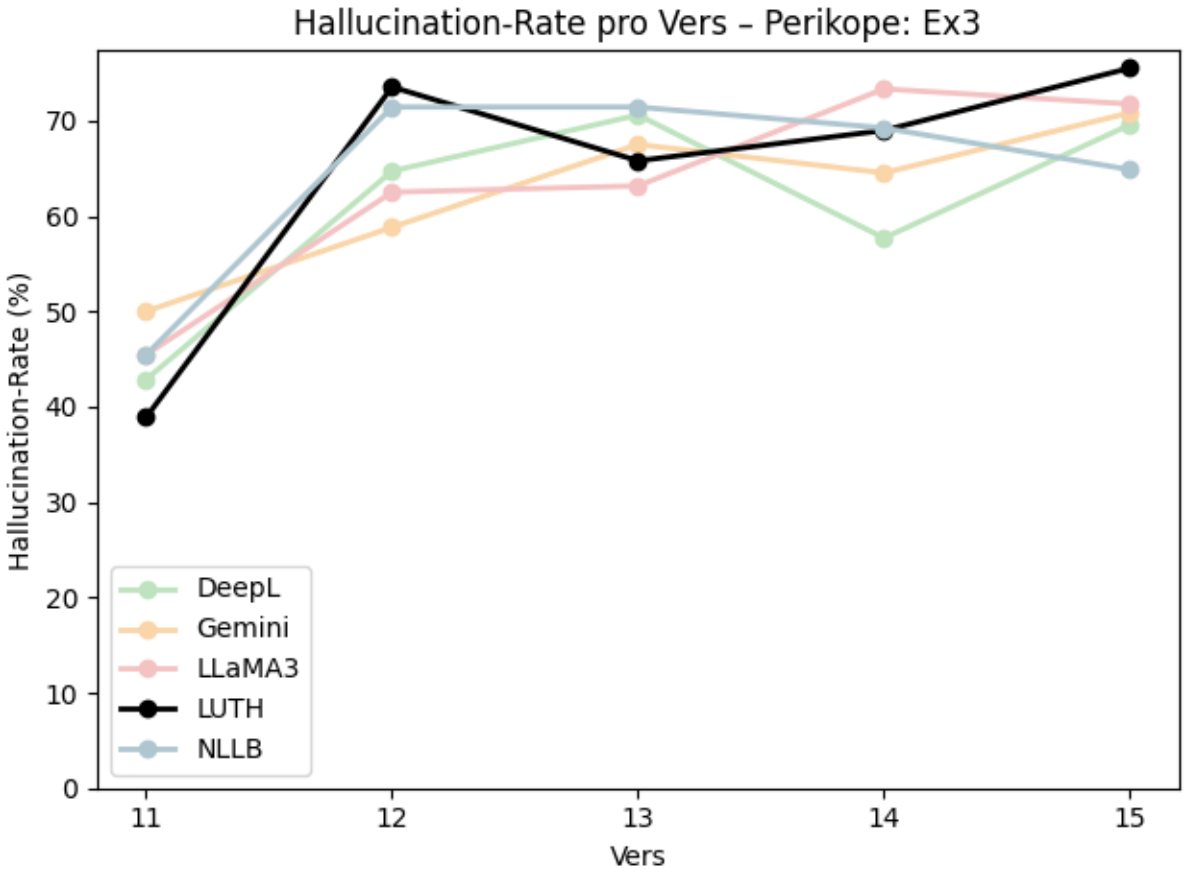
Anhang 4-2 Liniendiagramm | Ex3 – Length Ratio



Anhang 4-3 Liniendiagramm | Ex3 – Halluzinationsrate



Anhang 4-4 Liniendiagramm | Ex3 – Drop-Rate

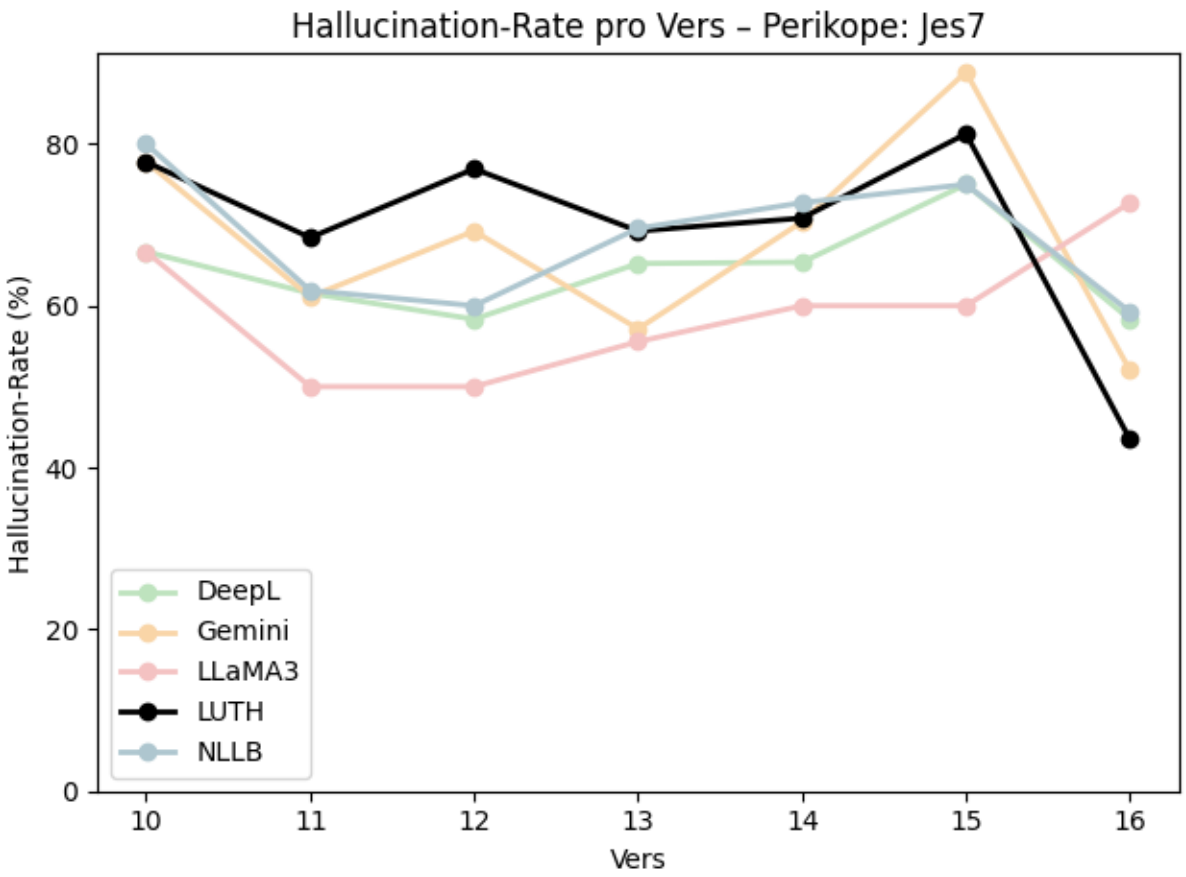


Anhang 5-1 Metriken | Jesaja 7, 10 – 16

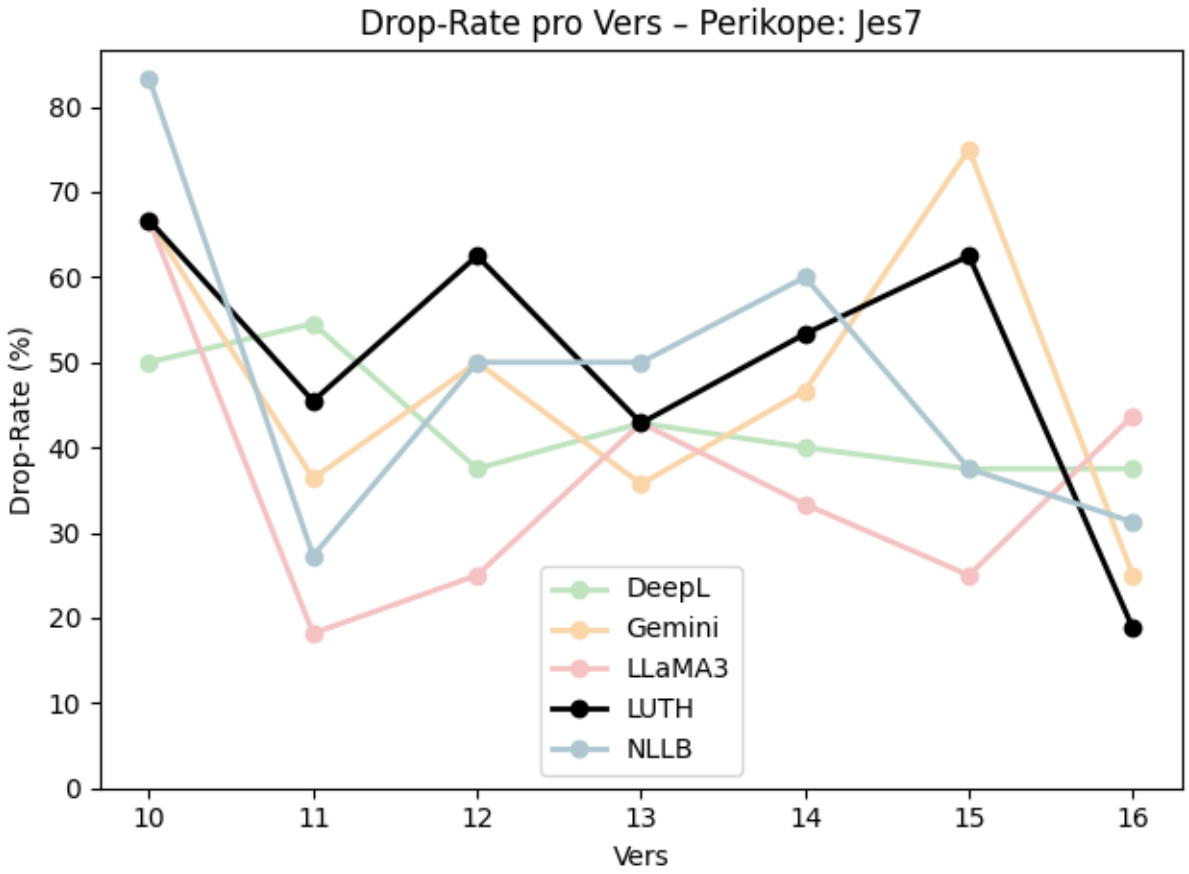
Perikope	Buch	Kapitel	Vers	Modell	Wortzahl (Quelle)	Wortzahl (Ziel)	Längenverhältnis	Subtokenanzahl	TFR
pericope	book	chapter	verse	model	src_word_count	dst_word_count	length_ratio	subtoken_count	act_tfr
Jes7,10-16	Jesaja	7	10	DeepL	6	9	1.5000	-	-
Jes7,10-16	Jesaja	7	11	DeepL	11	13	1.1818	-	-
Jes7,10-16	Jesaja	7	12	DeepL	8	12	1.5000	-	-
Jes7,10-16	Jesaja	7	13	DeepL	14	23	1.6429	-	-
Jes7,10-16	Jesaja	7	14	DeepL	15	26	1.7333	-	-
Jes7,10-16	Jesaja	7	15	DeepL	8	20	2.5000	-	-
Jes7,10-16	Jesaja	7	16	DeepL	16	24	1.5000	-	-
Jes7,10-16	Jesaja	7	10	NLLB	6	5	0.8333	7	1.4000
Jes7,10-16	Jesaja	7	11	NLLB	11	21	1.9091	33	1.5714
Jes7,10-16	Jesaja	7	12	NLLB	8	10	1.2500	17	1.7000
Jes7,10-16	Jesaja	7	13	NLLB	14	23	1.6429	35	1.5217
Jes7,10-16	Jesaja	7	14	NLLB	15	22	1.4667	30	1.3636
Jes7,10-16	Jesaja	7	15	NLLB	8	20	2.5000	29	1.4500
Jes7,10-16	Jesaja	7	16	NLLB	16	27	1.6875	40	1.4815
Jes7,10-16	Jesaja	7	10	Gemini	6	9	1.5000	-	-
Jes7,10-16	Jesaja	7	11	Gemini	11	18	1.6364	-	-
Jes7,10-16	Jesaja	7	12	Gemini	8	13	1.6250	-	-
Jes7,10-16	Jesaja	7	13	Gemini	14	21	1.5000	-	-
Jes7,10-16	Jesaja	7	14	Gemini	15	27	1.8000	-	-
Jes7,10-16	Jesaja	7	15	Gemini	8	18	2.2500	-	-
Jes7,10-16	Jesaja	7	16	Gemini	16	25	1.5625	-	-
Jes7,10-16	Jesaja	7	10	LLaMA3	6	6	1.0000	9	1.5000
Jes7,10-16	Jesaja	7	11	LLaMA3	11	18	1.6364	27	1.5000
Jes7,10-16	Jesaja	7	12	LLaMA3	8	12	1.5000	16	1.3333
Jes7,10-16	Jesaja	7	13	LLaMA3	14	18	1.2857	29	1.6111
Jes7,10-16	Jesaja	7	14	LLaMA3	15	25	1.6667	32	1.2800
Jes7,10-16	Jesaja	7	15	LLaMA3	8	15	1.8750	22	1.4667
Jes7,10-16	Jesaja	7	16	LLaMA3	16	33	2.0625	57	1.7273

	fehlend im Quelltext (Auslassungen)	**fehlend im Zieltext** (Halluzinationen)	**Quell-Ziel-Wortpaare**	**Halluzinationsrate**	**Anteil: fehlend im Quelltext**	**Modell**	**Stelle**
	src_unalign_count	dst_ unalign_count	align_pair_count	hallucination_rate	drop_rate	model	passage
	3	6	3	66.67 %	50.00 %	DeepL	Jes7,10
	6	8	5	61.54 %	54.55 %	DeepL	Jes7,11
	3	7	5	58.33 %	37.50 %	DeepL	Jes7,12
	6	15	8	65.22 %	42.86 %	DeepL	Jes7,13
	6	17	9	65.38 %	40.00 %	DeepL	Jes7,14
	3	15	8	75.00 %	37.50 %	DeepL	Jes7,15
	6	14	10	58.33 %	37.50 %	DeepL	Jes7,16
	5	4	1	80.00 %	83.33 %	NLLB	Jes7,10
	3	13	8	61.90 %	27.27 %	NLLB	Jes7,11
	4	6	4	60.00 %	50.00 %	NLLB	Jes7,12
	7	16	7	69.57 %	50.00 %	NLLB	Jes7,13
	9	16	6	72.73 %	60.00 %	NLLB	Jes7,14
	3	15	5	75.00 %	37.50 %	NLLB	Jes7,15
	5	16	11	59.26 %	31.25 %	NLLB	Jes7,16
	4	7	2	77.78 %	66.67 %	Gemini	Jes7,10
	4	11	7	61.11 %	36.36 %	Gemini	Jes7,11
	4	9	4	69.23 %	50.00 %	Gemini	Jes7,12
	5	12	9	27.14 %	35.71 %	Gemini	Jes7,13
	7	19	8	70.37 %	46.67 %	Gemini	Jes7,14
	6	16	2	88.89 %	75.00 %	Gemini	Jes7,15
	4	13	12	52.00 %	25.00 %	Gemini	Jes7,16
	4	4	2	66.67 %	66.67 %	LLaMA3	Jes7,10
	2	9	9	50.00 %	18.18 %	LLaMA3	Jes7,11
	2	6	6	50.00 %	25.00 %	LLaMA3	Jes7,12
	6	10	8	55.56 %	42.86 %	LLaMA3	Jes7,13
	5	15	10	60.00 %	33.33 %	LLaMA3	Jes7,14
	2	9	6	60.00 %	25.00 %	LLaMA3	Jes7,15
	7	24	9	72.73 %	43.75 %	LLaMA3	Jes7,16

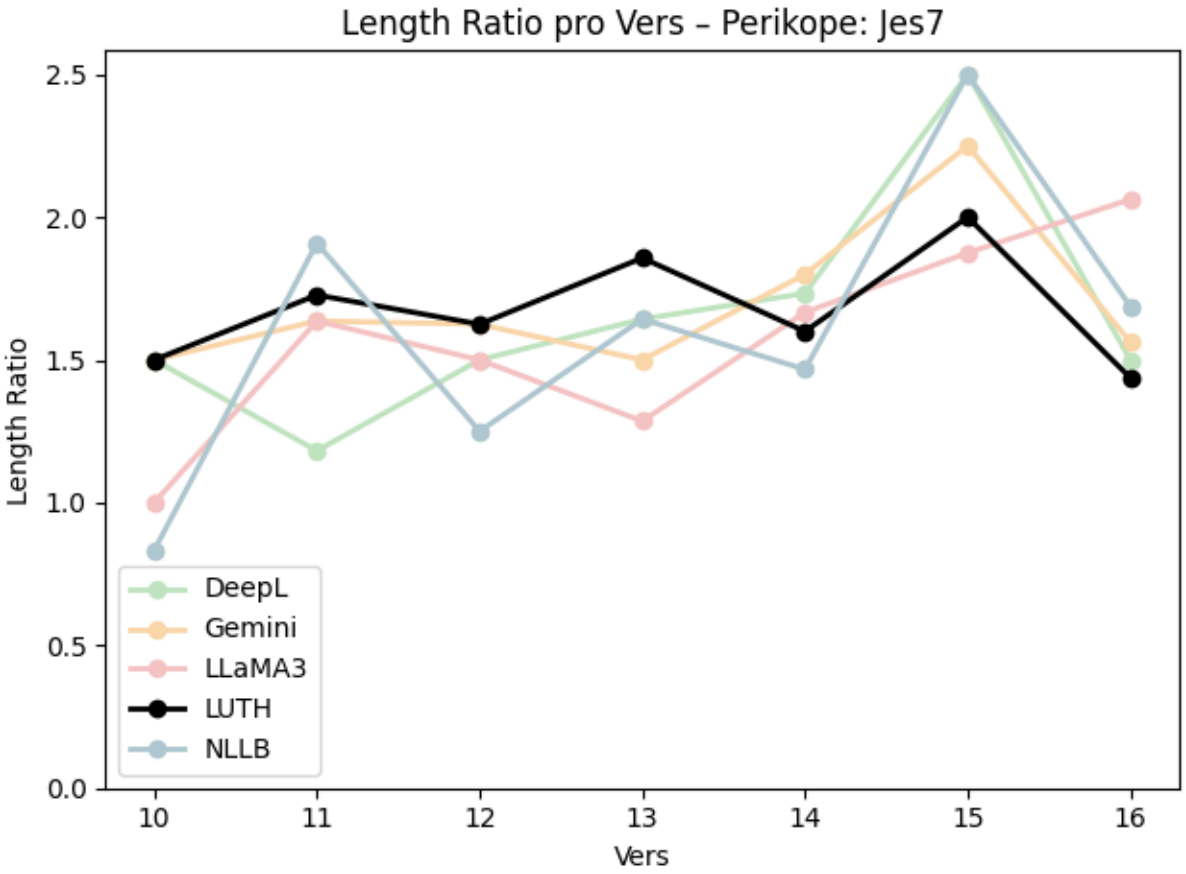
Anhang 5-2 Liniendiagramm | Jes7 – Length Ratio



Anhang 5-3 Liniendiagramm | Jes7 – Halluzinationsrate



Anhang 5-4 Liniendiagramm | Jes7 – Drop-Rate

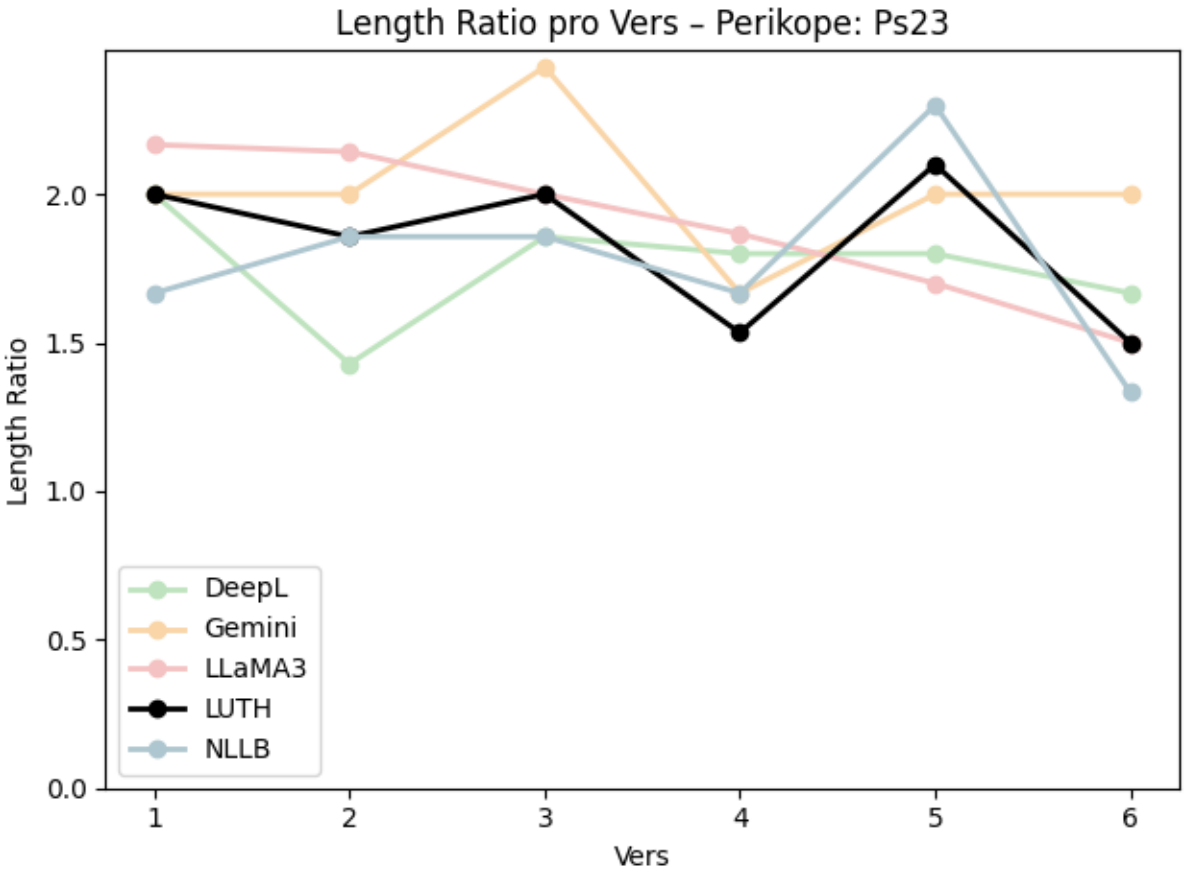


Anhang 6-1 Metriken | Psalm 23, 1 – 6

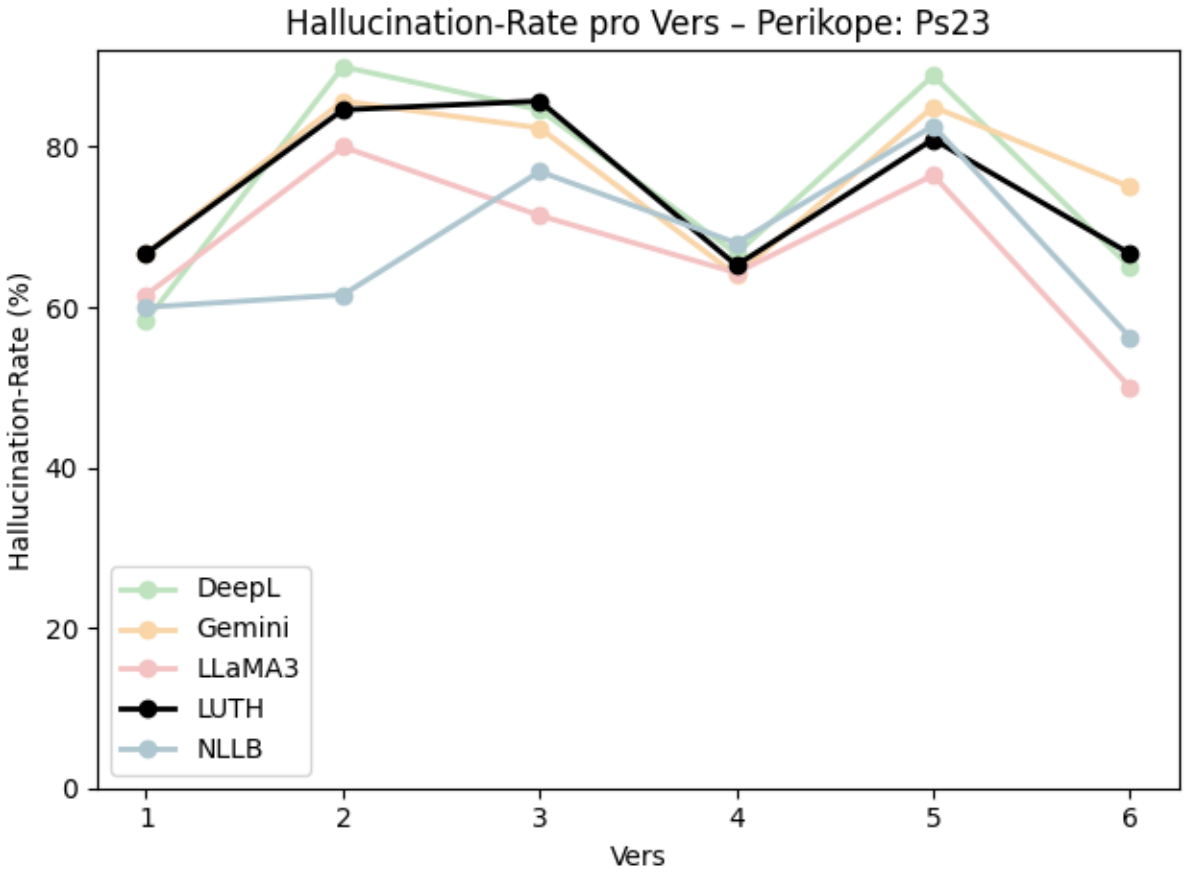
Perikope	**Buch**	**Kapitel**	**Vers**	**Modell**	**Wortzahl** (Quelle)	**Wortzahl** (Ziel)	**Längenverhältnis**	**Subtokenanzahl**	**TFR**
pericope	book	chapter	verse	model	src_word_count	dst_word_count	length_ratio	subtoken_count	act_tfr
Ps23,1-6	Psalm	23	1	DeepL	6	12	2.0000	-	-
Ps23,1-6	Psalm	23	2	DeepL	7	10	1.4286	-	-
Ps23,1-6	Psalm	23	3	DeepL	7	13	1.8671	-	-
Ps23,1-6	Psalm	23	4	DeepL	15	27	1.8000	-	-
Ps23,1-6	Psalm	23	5	DeepL	10	18	1.8000	-	-
Ps23,1-6	Psalm	23	6	DeepL	12	20	1.6667	-	-
Ps23,1-6	Psalm	23	1	NLLB	6	10	1.6667	15	1.5000
Ps23,1-6	Psalm	23	2	NLLB	7	13	1.8571	19	1.4615
Ps23,1-6	Psalm	23	3	NLLB	7	13	1.8571	20	1.5385
Ps23,1-6	Psalm	23	4	NLLB	15	25	1.6667	38	1.5200
Ps23,1-6	Psalm	23	5	NLLB	10	23	2.3000	43	1.8696
Ps23,1-6	Psalm	23	6	NLLB	12	16	1.3000	24	1.5000
Ps23,1-6	Psalm	23	1	Gemini	6	12	2.0000	-	-
Ps23,1-6	Psalm	23	2	Gemini	7	14	2.0000	-	-
Ps23,1-6	Psalm	23	3	Gemini	7	17	2.4286	-	-
Ps23,1-6	Psalm	23	4	Gemini	15	25	1.6667	-	-
Ps23,1-6	Psalm	23	5	Gemini	10	20	2.0000	-	-
Ps23,1-6	Psalm	23	6	Gemini	12	24	2.0000	-	-
Ps23,1-6	Psalm	23	1	LLaMA3	7	13	2.1667	19	1.4615
Ps23,1-6	Psalm	23	2	LLaMA3	7	15	2.1429	26	1.7333
Ps23,1-6	Psalm	23	3	LLaMA3	7	14	2.0000	27	1.9286
Ps23,1-6	Psalm	23	4	LLaMA3	15	28	1.8667	45	1.6071
Ps23,1-6	Psalm	23	5	LLaMA3	10	17	1.7000	27	1.5822
Ps23,1-6	Psalm	23	6	LLaMA3	12	18	1.8333	33	1.8333

	fehlend im Quelltext (Auslassungen)	**fehlend im Zieltext** (Halluzinationen)	**Quell-Ziel-Wortpaare**	**Halluzinationsrate**	**Anteil: fehlend im Quelltext**	**Modell**	**Stelle**
	src_unalign_count	dst_unalign_count	align_pair_count	hallucination_rate	drop_rate	model	passage
	1	7	5	58.33 %	16.67 %	DeepL	Ps23,1
	6	9	1	90.00 %	85.71 %	DeepL	Ps23,2
	5	11	2	84.62 %	71.43 %	DeepL	Ps23,3
	6	18	9	66.67 %	40.00 %	DeepL	Ps23,4
	8	16	2	88.89 %	80.00 %	DeepL	Ps23,5
	5	13	7	65.00 %	41.67 %	DeepL	Ps23,6
	2	6	4	60.00 %	33.33 %	NLLB	Ps23,1
	2	8	5	61.54 %	28.57 %	NLLB	Ps23,2
	4	10	3	76.92 %	57.14 %	NLLB	Ps23,3
	7	17	8	68.00 %	46.67 %	NLLB	Ps23,4
	6	19	4	82.61 %	60.00 %	NLLB	Ps23,5
	5	9	7	56.25 %	41.67 %	NLLB	Ps23,6
	2	8	4	66.67 %	33.33 %	Gemini	Ps23,1
	5	12	2	85.71 %	71.43 %	Gemini	Ps23,2
	4	14	3	82.35 %	57.14 %	Gemini	Ps23,3
	6	16	9	64.00 %	40.00 %	Gemini	Ps23,4
	7	17	3	85.00 %	70.00 %	Gemini	Ps23,5
	6	18	6	75.00 %	50.00 %	Gemini	Ps23,6
	1	8	5	61.54 %	16.67 %	LLaMA3	Ps23,1
	4	12	3	80.00 %	57.14 %	LLaMA3	Ps23,2
	3	10	4	71.43 %	42.86 %	LLaMA3	Ps23,3
	5	18	10	64.29 %	33.33 %	LLaMA3	Ps23,4
	6	13	4	76.47 %	60.00 %	LLaMA3	Ps23,5
	3	9	9	50.00 %	25.00 %	LLaMA3	Ps23,6

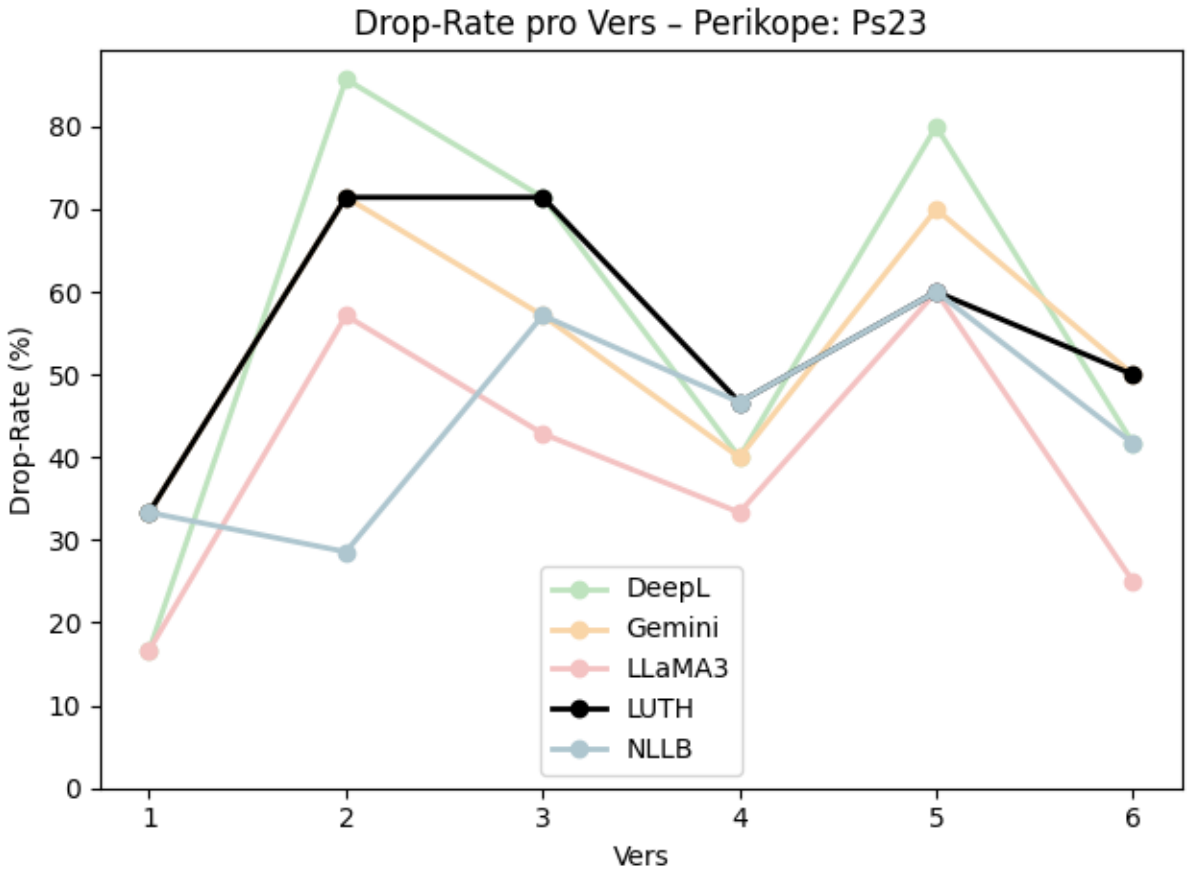
Anhang 6-2 Liniendiagramm | Ps23 – Length Ratio



Anhang 6-3 Liniendiagramm | Ps23 – Halluzinationsrate



Anhang 6-4 Liniendiagramm | Ps23 – Drop-Rate

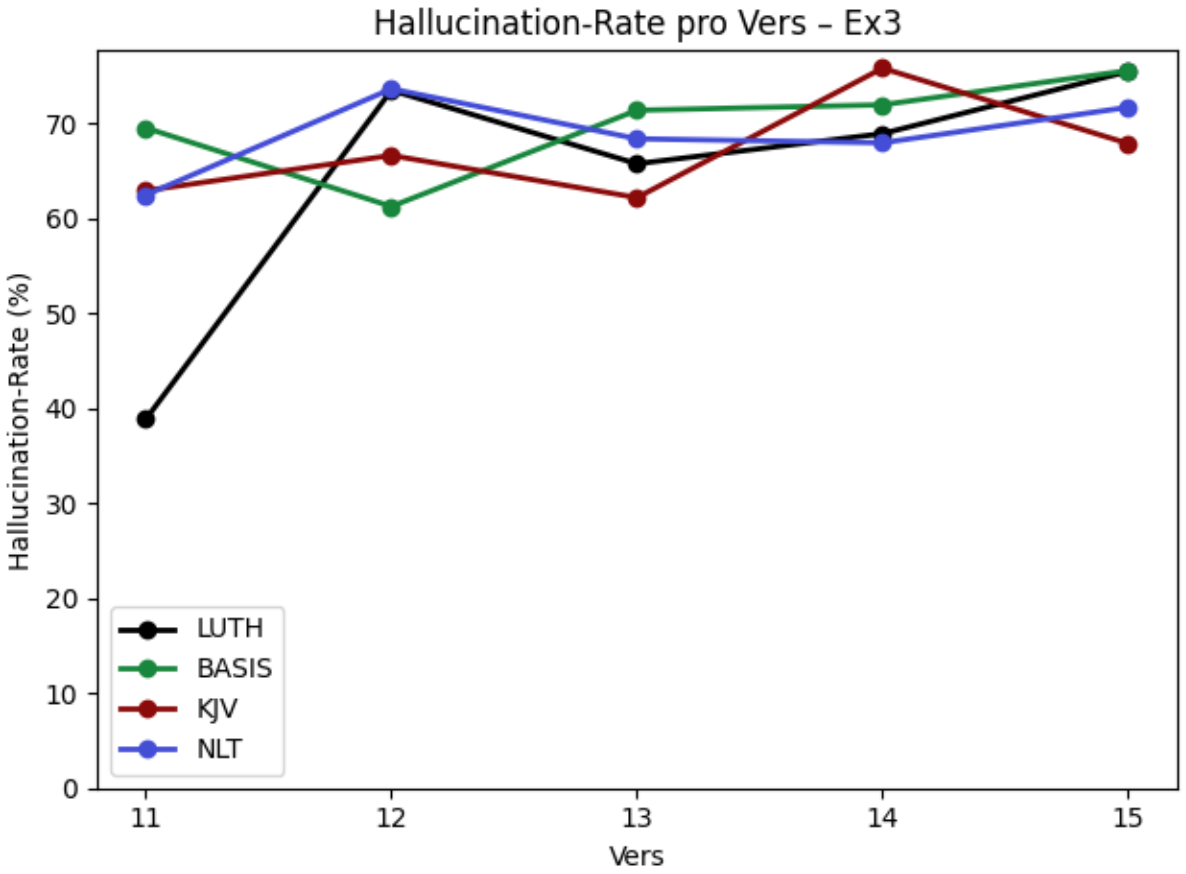


Anhang 7-1 Metriken | Referenzübersetzung – Luther

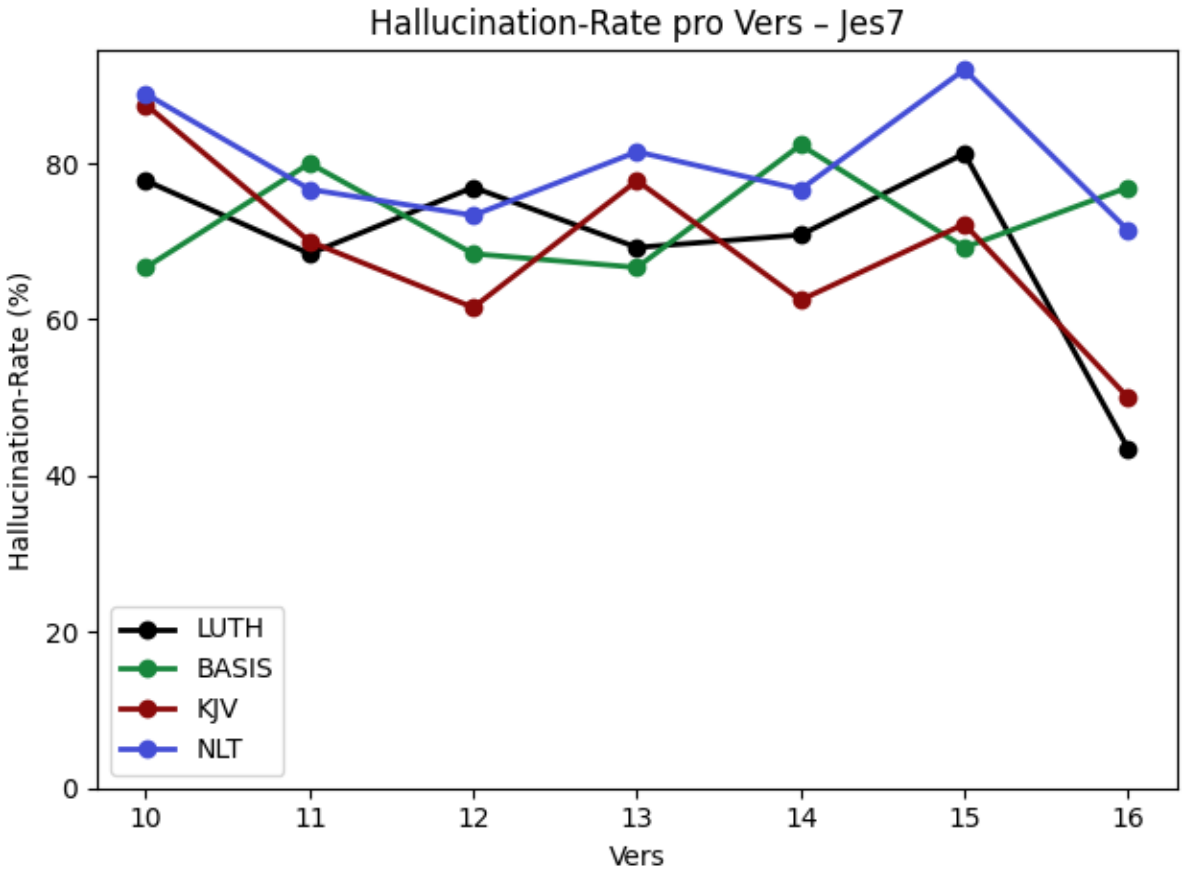
Perikope	**Buch**	**Kapitel**	**Vers**	**Ausgabe**	**Wortzahl** (Quelle)	**Wortzahl** (Ziel)	**Längenverhältnis**	**Subtokenanzahl**	**TFR**
pericope	book	chapter	verse	edition	src_word_count	dst_word_count	length_ratio	subtoken_count	act_tfr
Ex3,11-15	Exodus	3	11	Luther	16	18	1.1250	-	-
Ex3,11-15	Exodus	3	12	Luther	20	34	1.7000	-	-
Ex3,11-15	Exodus	3	13	Luther	23	38	1.6522	-	-
Ex3,11-15	Exodus	3	14	Luther	15	29	1.9333	-	-
Ex3,11-15	Exodus	3	15	Luther	28	49	1.7500	-	-
Jes7,10-16	Jesaja	7	10	Luther	6	9	1.5000	-	-
Jes7,10-16	Jesaja	7	11	Luther	11	19	1.7273	-	-
Jes7,10-16	Jesaja	7	12	Luther	8	13	1.650	-	-
Jes7,10-16	Jesaja	7	13	Luther	14	26	1.8571	-	-
Jes7,10-16	Jesaja	7	14	Luther	15	24	1.6000	-	-
Jes7,10-16	Jesaja	7	15	Luther	8	16	2.0000	-	-
Jes7,10-16	Jesaja	7	16	Luther	16	23	1.4375	-	-
Ps23,1-6	Psalm	23	1	Luther	6	12	2.000	-	-
Ps23,1-6	Psalm	23	2	Luther	7	13	1.8571	-	-
Ps23,1-6	Psalm	23	3	Luther	7	14	2.0000	-	-
Ps23,1-6	Psalm	23	4	Luther	15	23	1.5333	-	-
Ps23,1-6	Psalm	23	5	Luther	10	21	2.1000	-	-
Ps23,1-6	Psalm	23	6	Luther	12	18	1.5000	-	-

	fehlend im Quelltext (Auslassungen)	**fehlend im Zieltext** (Halluzinationen)	**Quell-Ziel-Wortpaare**	**Halluzinationsrate**	**Anteil: fehlend im Quelltext**	**Ausgabe**	**Stelle**
	src_unalign_count	dst_unalign_count	align_pair_count	hallucination_rate	drop_rate	edition	passage
	5	7	11	38.89 %	31.25 %	Luther	Ex3,11
	11	25	9	73.53 %	55.00 %	Luther	Ex3,12
	10	25	13	65.79 %	43.48 %	Luther	Ex3,13
	6	20	9	68.98 %	40.00 %	Luther	Ex3,14
	16	37	12	75.51 %	57.14 %	Luther	Ex3,15
	4	7	2	77.78 %	66.67 %	Luther	Jes7,10
	5	13	6	68.42 %	45.45 %	Luther	Jes7,11
	5	10	3	76.92 %	62.50 %	Luther	Jes7,12
	6	18	8	69.23 %	42.86 %	Luther	Jes7,13
	8	17	7	70.83 %	53.33 %	Luther	Jes7,14
	5	13	3	81.25 %	62.50 %	Luther	Jes7,15
	3	10	13	43.48 %	18.75 %	Luther	Jes7,16
	2	8	4	66.67 %	33.33 %	Luther	Ps23,1
	5	11	2	84.62 %	71.43 %	Luther	Ps23,2
	5	12	2	85.71 %	71.43 %	Luther	Ps23,3
	7	15	8	65.22 %	46.67 %	Luther	Ps23,4
	6	17	4	80.95 %	60.00 %	Luther	Ps23,5
	6	12	6	66.67 %	50.00 %	Luther	Ps23,6

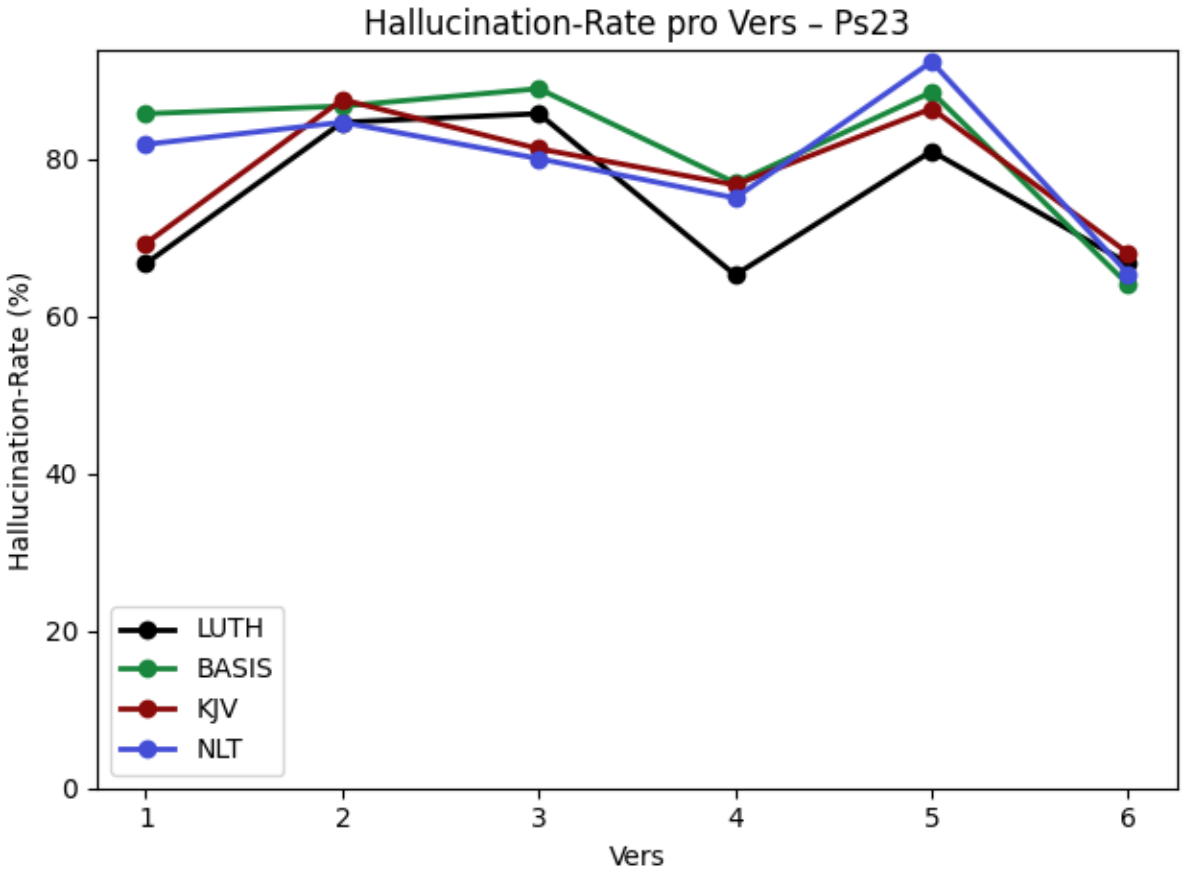
Anhang 7-2 Liniendiagramm | Ex3 – Halluzinationsrate (Referenzen)



Anhang 7-3 Liniendiagramm | Jes7 – Halluzinationsrate (Referenzen)



Anhang 7-4 Liniendiagramm | Ps23 – Halluzinationsrate (Referenzen)



Anhang 8-1 Qualitative Metriken: chrF u. BERTScore

Perikope	**Sprache**	**Modell**	**chrF** (Durchschnitt)	**chrF** (Varianz)	**BERTScore [F1]** (Durchschnitt)	**BERTScore [F1]** (Varianz)	**BERTScore [P]** (Durchschnitt)	**BERTScore [P]** (Varianz)
pericope	language	model	chrf_avrg	chrf_vari	bert_avrg	bert_vari	bert_avrg	bert_vari
Ex3,11-15	DE	DeepL	71	6	0,952	0,015	0,952	0,015
Ex3,11-15	DE	NLLB	71	4	0,961	0,009	0,961	0,010
Ex3,11-15	DE	Gemini	78	6	0,967	0,011	0,963	0,014
Ex3,11-15	DE	LLaMA3	66	5	0,950	0,006	0,953	0,002
Ex3,11-15	EN	DeepL	76	16	0,952	0,035	0,951	0,035
Ex3,11-15	EN	NLLB	82	7	0,970	0,012	0,973	0,014
Ex3,11-15	EN	Gemini	66	9	0,945	0,021	0,945	0,029
Ex3,11-15	EN	LLaMA3	67	6	0,942	0,015	0,951	0,013
Jes7,10-16	DE	DeepL	54	11	0,917	0,025	0,927	0,024
Jes7,10-16	DE	NLLB	34	14	0,868	0,086	0,871	0,097
Jes7,10-16	DE	Gemini	37	7	0,944	0,033	0,950	0,028
Jes7,10-16	DE	LLaMA3	64	13	0,887	0,027	0,892	0,029
Jes7,10-16	EN	DeepL	59	15	0,935	0,026	0,941	0,025
Jes7,10-16	EN	NLLB	45	10	0,908	0,024	0,908	0,029
Jes7,10-16	EN	Gemini	64	14	0,947	0,024	0,950	0,024
Jes7,10-16	EN	LLaMA3	46	16	0,914	0,037	0,918	0,040
Ps23,1-6	DE	DeepL	61	20	0,927	0,032	0,933	0,028
Ps23,1-6	DE	NLLB	38	18	0,893	0,040	0,899	0,038
Ps23,1-6	DE	Gemini	64	14	0,935	0,030	0,933	0,031
Ps23,1-6	DE	LLaMA3	46	10	0,901	0,027	0,898	0,027
Ps23,1-6	EN	DeepL	72	11	0,950	0,019	0,959	0,015
Ps23,1-6	EN	NLLB	57	27	0,935	0,044	0,942	0,038
Ps23,1-6	EN	Gemini	78	10	0,969	0,016	0,972	0,015
Ps23,1-6	EN	LLaMA3	57	19	0,930	0,043	0,933	0,046

Anhang 9-1 Zugriff auf Arbeitsdateien

Zusätzlich zu den im Anhang abgedruckten Ergebnissen werden alle im Rahmen dieser Arbeit verwendeten Arbeitsdateien (u. a. die aufbereiteten Textkorpora, Skripte, Ausgabedateien sowie ergänzende Materialien) zentral über eine begleitende Projektwebseite bereitgestellt. Dort stehen die Dateien gebündelt zum Download zur Verfügung, um die Nachvollziehbarkeit der Arbeitsschritte und die Reproduzierbarkeit der Auswertungen zu unterstützen. Die Ablage umfasst sowohl die Rohdaten als auch die daraus abgeleiteten Zwischen- und Endstände, jeweils in den in der Arbeit beschriebenen Formaten. So können die einzelnen Verarbeitungsschritte (Digitalisierung, Strukturierung, Übersetzung und Auswertung) transparent eingesehen und bei Bedarf eigenständig überprüft werden.

Die begleitende Projektwebseite ist entweder direkt über die URL https://masterarbeit.baldeweg.me erreichbar oder bequem über den im Anhang abgebildeten QR-Code. Über beide Zugangswege gelangen Leser:innen zur Download-Übersicht mit den bereitgestellten Arbeitsdateien und ergänzenden Materialien.

https://masterarbeit.baldeweg.me

Eidesstattliche Erklärung

INTERNATIONALE HOCHSCHULE

Hiermit versichere ich an Eides statt, dass ich die Abschlussarbeit selbständig und ohne Inanspruchnahme fremder Hilfe angefertigt habe. Ich habe dabei nur die angegebenen Quellen und Hilfsmittel verwendet und die aus diesen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen. Ich erkläre mich damit einverstanden, dass die Arbeit mit Hilfe eines Plagiatserkennungsdienstes auf enthaltene Plagiate überprüft wird.

Langenweddekath, 19. Januar 2026
Ort, Datum

Josie Baldewig
Unterschrift