universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Few-Shot Prompting vs. Chain-of-Dictionary Prompting of Large Language Models for Machine Translation with a Multilingual Knowledge Graph

verfasst von | submitted by

Carina Obster B.A. M.A.

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 587
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Joint-Masterstudium Multilingual Technologies
Betreut von | Supervisor:	Miguel Angel Rios Gaona PhD

Contents

1 Introduction

Large Language Models (LLMs) have recently demonstrated strong performance across a wide range of Natural Language Processing (NLP) tasks, including Machine Translation (MT). Owing to their scale, multilingual training data, and In-Context Learning (ICL) capabilities, LLMs are able to perform translation in zero-shot and few-shot settings, in some cases rivaling or even surpassing traditional Neural Machine Translation (NMT) systems for high-resource language pairs (Vilar et al., 2023; Lyu et al., 2024; Zhu et al., 2024b).

However, despite these advances, LLM-based MT faces substantial challenges. In particular, translation quality often degrades in domain-specific and low-resource contexts, where specialized terminology, institutional conventions, and subtle semantic distinctions play a crucial role. This limitation is especially pronounced in the legal, medical, or technical domains, where mistranslations can lead to serious consequences.

Knowledge Graphs (KGs), typically expressed as triples, address this limitation of LLMs by providing a structured representation of real-world knowledge in the form of entities and their relations (Kaffee et al., 2023; Pan et al., 2024). By organizing information in an explicit and interpretable graph structure, KGs enable efficient access to factual and domain-specific knowledge and so have been widely used in downstream applications such as Question Answering (QA), Information Retrieval (IR), and Recommendation systems (Ibrahim et al., 2024).

Multilingual Knowledge Graphs (MKGs) extend this idea by linking concepts across languages, either through multilingual labels or interconnected, language-specific subgraphs (Kaffee et al., 2023). As a result, MKGs can encode cross-lingual correspondences and domain-specific terminology in a way that is independent of any single language, making them a promising resource for addressing terminology-related weaknesses in MT systems.

Prior research has largely focused on integrating KGs into traditional NMT models through embeddings, retrieval mechanisms, or joint training strategies (Pan et al., 2024). While these approaches have shown improvements in translation quality, they often require costly retraining, complex architectures, and limited flexibility when new knowledge is added. In contrast, recent work on LLMs has shifted attention toward inference-time methods, particularly prompting (Pan et al., 2024), which allows structured external knowledge to be injected into a model without modifying its parameters. In this context, KG-based prompting

has emerged as an appealing alternative, as structured knowledge from a KG can be transformed into textual prompts that guide a model's translation behavior (Pan et al., 2024). Early studies suggest that KG-derived prompts can function as domain-specific glossaries or translation memories, improving the handling of named entities and specialized terminology (Moussallem et al., 2019).

Building on these developments, this work investigated the use of MKGs as a source of in-context examples employed to improve domain-specific MT using LLMs. Rather than relying solely on sentence-level parallel data, we explored how structured multilingual terminology from a KG can be leveraged to construct prompts that better activate LLMs' cross-lingual and reasoning capabilities. In particular, we focused on prompting strategies inspired by Chain-of-Thought (CoT) methods, which assume that auxiliary, high-resource languages can help bridge semantic gaps between source and target languages, especially low-resource languages (Lu et al., 2024b). By grounding prompts in an MKG, this approach aims to combine the strengths of structured knowledge representation and flexible LLM inference, offering a lightweight and adaptable solution for domain-adapted MT. Our research question thus is: To what extent does knowledge-based ICL compare to CoT prompting for domain-specific MT in LLMs? We focused on MT from English into German, in the domain of EU terminology. The results suggest that while the few-shot baseline method proves more reliable for established EU glossary terminology, the proposed CoT method produces more idiomatic and nuanced translations, though neither approach consistently outperforms the other across all models and subdomains.

This thesis is organized as follows. Chapter 3 provides the background for this research, with Chapter 3.1 providing an introduction to KGs. Chapter 3.2 discusses augmentation of LLMs with KGs, particularly with respect to enhancement of LLM inference. Chapter 3.3 discusses the role of KGs in MT, including their integration via embeddings, retrieval, and prompting. Chapter 3.4 provides the various techniques used in MT, including Rule-Based MT (RBMT), Statistical MT (SMT), and NMT, and concludes with LLMs. Chapter 4 presents related work relevant to the research. Chapter 4.1 discusses the role of prompting in LLMs, and Chapter 4.2 provides strategies for retrieving ICL examples.

Respectively, Chapters 5, 6, and 7 provide the methodology employed in the study, a report of the results obtained, and a discussion of the implication of these results. Chapter 8 concludes the thesis with a discussion of the study limitations and suggestions for future work.

2 Background

2.1 Introduction to KGs

KGs are an effective means of storing information as graph-structured data (Kaffee et al., 2023; Pan et al., 2024). They are an effective way to store information, as graph traversal is usually quick even as a KG grows in size (Ibrahim et al., 2024). As shown below, in their mathematical form, KGs are based on triples comprised of a head entity (h), a tail entity (t), and their relationship (r), where E denotes the set of entities, fundamental units that represent real-world concepts, and R is the set of these entities' relations.

$$KG = \{(h, r, t) \subseteq E \times R \times E\}$$

Thus, for the statement "Paris is the capital of France", h = "Paris", r = "is the capital of", and t is "France". Graphically, entities are denoted by *nodes* and the relations between them by *edges*; features or qualities of nodes and edges are called *properties* (Ibrahim et al., 2024). The edges connecting nodes are not necessarily symmetric; in most cases, head entities point to tail entities via the edge (Peng et al., 2023a).

Data models commonly used for KGs are Resource Description Framework (RDF) and Labeled Property Graphs (LPGs). RDF uses a simple triple structure (subject-predicate-object) and is widely used for the Semantic Web (Peng et al., 2023a). LPGs provide rich information regarding properties of nodes and edges, making them popular in graph databases for more complex, real-world applications, such as social networks and recommendation systems. The following are prerequisites that allow the extraction of prompts from a KG, which can then be fed to the LLM: Entities, classes, and properties must be labelled with a natural language label (Kaffee et al., 2023). The language of an entity has to be labelled with a language tag. It is, of course, also important that each entity has labels in the relevant languages. So-called "monolingual islands", i.e., parts of the KG that are only labelled in one language, should be avoided for this task. In addition, a single preferred label should be used per entity to avoid ambiguity.

Pan et al. (2024) distinguish four types of KGs (see Figure 1): The most well-known kind of KG is the *encyclopedic KG*, which stores common knowledge. *Commonsense KGs*, on the other hand, store information on everyday concepts and

their relationships, such as objects or events. *Domain-specific KGs* store knowledge of specific domains, such as medicine, finance or law. Finally, *multimodal KGs* contain image-, sound- and video-based information along with textual information. As we are particularly interested in the question how KGs can enable LLMs to better adapt to a certain domain for MT, we focused on domain-specific KGs in this work.

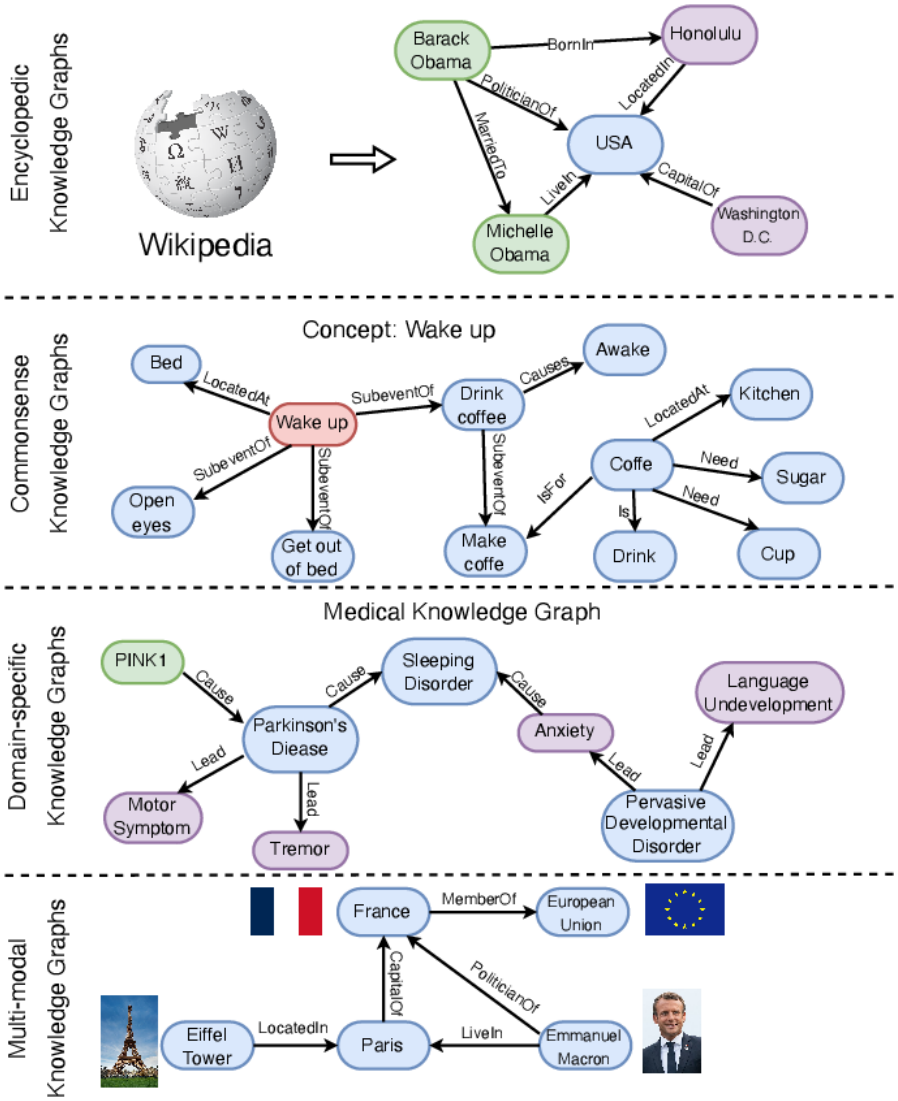

Figure 1: Categories of KGs. Reprinted from Pan et al. (2024) , © 2024 IEEE.

In some KGs, called MKGs, stored information is available in several languages. There are different ways of storing multilingual information in KGs: A KG entity

might, for example, feature labels and descriptions in multiple languages (Wikidata KG, see Figure 2). In other cases, a sub-KG is created for every single language, and these sub-KGs are interlinked to form one comprehensive KG (DBPedia KG) (Kaffee et al., 2023).

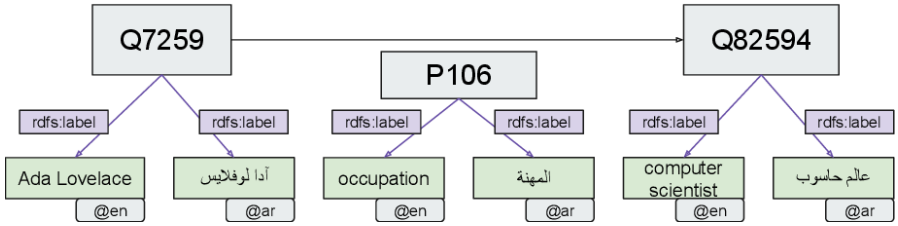

Figure 2: A unified KG with labels in several languages. From Kaffee et al. (2023), licensed under CC BY 4.0.

MKGs encapsulate details about entities regardless of whether this information is explicitly available in every language, ensuring that it remains accessible even in languages where it might not otherwise exist (Kaffee et al., 2023). We focus on the first type of MKG—a unified KG with labels and descriptions in several languages, since it lends itself to extracting the bilingual prompts needed to prompt an LLM for MT in a later step.

MKGs are being applied to a wide variety of downstream tasks, such as Multilingual Knowledge Graph QA, which uses facts stored in a KG to answer user queries in different languages. Another popular, related field of application is search and recommendation systems (Ibrahim et al., 2024). In a relatively new development, the integration of KGs into NMT models and LLMs for MT is of particular use when translating domain-specific knowledge or named entities (Kaffee et al., 2023).

In Chapter 4, we look more closely at how KGs can be integrated in LLMs for MT. Of particular interest are how LLMs can be used to prompt KGs, the design of these prompts, and the extent to which the similarity of the ICL Examples to the prompt's input can play a role.

2.2 KGs for LLMs

Pan et al. (2024) argue that KGs offer the practical knowledge that LLMs—despite being trained on large amounts of data—often lack. As a consequence, KGs can aid in reducing LLM hallucinations (Ibrahim et al., 2024) and provide LLMs with contextual reasoning, thereby supporting them in maintaining coherence and grasping subtleties. This is particularly helpful in the context of personal or domain-specific applications (Ibrahim et al., 2024). Additionally, KGs can mitigate the perception of LLMs as "black boxes" by showing a clear, organized view of the information used during the LLM's reasoning, thus making it easier to understand and trace where the information came from and so improving the LLM's transparency and explainability (Ibrahim et al., 2024). All of this leads to increased reliability, a crucial concern in fields where misinformation can have serious consequences, such as healthcare, finance, and law (Ibrahim et al., 2024).

2.2.1 Fields of Application

Recent research has focused on using the structured information contained in KGs to guide LLMs' reasoning process, thus reducing hallucinations and improving multi-hop reasoning. Existing methods treat KGs as reasoning maps for LLMs, construct explicit reasoning paths for LLMs to follow, or reformat KG data into LLM-friendly structures and apply reasoning (Ji et al., 2024).

Other researchers have improved LLM's factual accuracy and retrieval efficiency using KGs by, e.g., providing the LLM with optimized KG subgraphs or by adapting the query to the KG structure (Li et al., 2024).

2.2.2 KG-Enhanced LLM Pre-training

KGs can enhance LLMs during either the pre-training or the inference stage. As the focus of our work was the inference stage, we only briefly touch upon the subtopic of pre-training. Within this area, some researchers focus on knowledge-aware training of the LLM, i.e., using a KG to expose the model to a greater amount of knowledge during pre-training through such techniques as masking important entities. Other methods help the LLM connect input text with text from the KG by, for instance, aligning words with KG entities (Pan et al., 2024).

Integrating KGs into LLM inputs involves introducing relevant knowledge subgraphs or triples into the inputs of LLMs, and various approaches have been proposed to combine textual descriptions of entities and relations from KGs with LLM representations (Pan et al., 2024).

KG instruction-tuning helps train LLMs to better understand KGs and follow user instructions to complete complex tasks. Thus, KG instruction-tuning uses

both facts and KGs' structures to improve the model's reasoning skills (Pan et al., 2024).

However, a limitation of such LLM pre-training approaches is the need for a complete retraining of the model to incorporate new knowledge. Moreover, the resulting pre-trained LLM may not generalize well to new, previously unseen knowledge.

2.2.3 KG-Enhanced LLM Inference

Enhancing the LLM with KGs during inference is a way to keep the knowledge and the text space separate (Pan et al., 2024).

One way to accomplish this is to retrieve knowledge from the KG and then integrate this retrieved knowledge into the LLM. The most well-known approach, *Retrieval Augmented Generation* (RAG), searches for knowledge in the KG and feeds the retrieved pieces of information to the LLM as hidden variables (Pan et al., 2024).

In contrast, KG *Prompting* (see Figure 3) constructs prompts that convert structured knowledge from the KG into text sequences that are then fed to the LLM. According to Pan et al. (2024), KGs help create better, more diverse prompts compared to manual methods. They have been shown to improve both single-turn and multi-turn prompts—prompts requiring more reasoning and deeper graph traversal—with minimal effort. Adding explicit knowledge from KGs (like attributes and relationships) further improves LLM predictions (Brate et al., 2022).

Pan et al. (2024) state that some research directions in KG-enhanced LLM inference require further exploration, including how KGs can enforce rules to ensure that LLMs generate reliable, domain-specific information. Utilizing the inference approach thus enables LLMs to have up-to-date knowledge and thus be better able to process novel topics and domains. However, since the model is not fully trained on utilizing the new knowledge provided during inference, it may encounter certain performance issues.

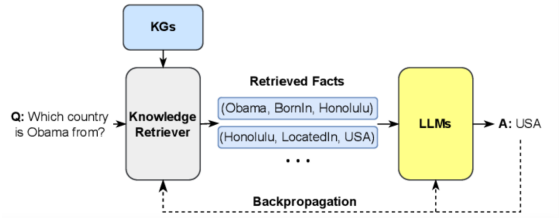

Figure 3: Illustration of the KG Prompting process. Reprinted from Pan et al. (2024), © 2024 IEEE

2.3 KGs Used for MT

To date, KGs employed in MT have primarily been incorporated into traditional NMT models. The most important motivation for injection of knowledge from KGs into NMT models and LLMs is domain adaptation—that is, enabling models to more effectively handle specialized terminology and context-specific meanings. LLMs in particular can perform poorly when tasked with translating domain-specific texts (Zhang et al., 2023b).

2.3.1 KG Type and Preparation

Different strategies have been developed to prepare KGs for integration into MT systems. For instance, Zhao et al. (2020b) aligned two monolingual KGs by converting them into a shared semantic space, thereby facilitating cross-lingual knowledge transfer. Zhao et al. (2020a) used a source-language KG and a target-language KG to construct a new bilingual KG, thus enhancing the model's understanding of semantic equivalence across languages. Zhang et al. (2023b) and Conia et al. (2024a) both utilized an MKG.

2.3.2 Integration Methods

Two general methods are employed to integrate KGs into NMTs and LLMs.

Embeddings

Moussallem et al. (2019) converted a KG into embeddings, which were then added to the NMT system's embedding layer to enrich the model with structured knowledge. Combining these embeddings with links generated through Entity Linking (EL) allowed the system to align input tokens with corresponding entities in the KG and so incorporate semantic context during translation.

During the past few years, however, the trend has been to move away from embeddings and pre-training and towards direct retrieval from KGs.

Retrieval

Zhao et al. (2020a) propose a method to pick out relevant information from a KG by matching key terms in the source sentence with components of the KG, using partial matching and a high-frequency filter strategy—both of which perform best among their selection methods. The researchers found that a reinforcement learning approach to dynamically select the most useful triples from the KG leads to better results than random selection. Their method next ranks the selected triples of information with respect to relevance using a neural network. With the

help of a dedicated knowledge encoder and KG attention mechanisms, the triples found most relevant can then be selected and used to guide an NMT system.

Conia et al. (2024a) utilized Wiki-MT, a demo system that integrates the Wikidata KG with an MT model. It features a knowledge retriever that identifies the most relevant terms from the KG by computing the cosine similarity between vector embeddings, and a knowledge-enhanced translator that generates the target text while incorporating the retrieved terms to improve translation accuracy and contextual relevance.

The same researchers also introduced XC-Translate, a benchmark designed to evaluate the performance of MT systems on texts containing culturally nuanced entity names (Conia et al., 2024b). To address the challenges posed by such texts, the authors proposed KG-MT, an end-to-end method that integrates information from an MKG into an NMT model using a dense retrieval mechanism. The aim of this approach is thus to enhance the translation of culturally specific entities by leveraging structured knowledge from KGs.

In another approach to entity representation, Zhao et al. (2020a) proposed the Byte-Pair Encoding (BPE) method, which breaks down entities from both the dataset and the KG into sub-entities. To the resulting fine-grained structure, they introduce a multi-task learning framework that jointly trains the translation model and reinforces the semantic representation of these sub-entities as an auxiliary task. This approach enables the model to better capture the meaning of entities, even those that are rare or fragmented, thus leading to more accurate handling of entity names.

The retrieval methods summarized above are difficult to implement and incur high computational costs for training the NMT model. Moreover, they cannot be updated continuously as when, for instance, new terminology is added to the KG.

2.3.3 Prompting (LLMs)

In the recent past, prompting LLMs with additional information has garnered increased attention. Zhang et al. (2023b) propose a method in which entities are first extracted from their source sentence. The MKG is then queried to retrieve bilingual entity pairs corresponding to the extracted entities. The triples introduced into the prompt are not used as traditional ICL examples, but rather as term-translation pairs, thereby effectively treating the MKG as a domain-specific glossary or Translation Memory (TM) to guide the model's output.

2.4 Approaches to MT

Four primary approaches to MT have been employed: Rule-Based MT, Statistical MT, Neural MT, and MT in LLMs. Brief discussions of these follow.

2.4.1 RBMT

RBMT is an early MT approach based on grammatical rules, interlingual transformational rules, and bilingual dictionaries. The general RBMT process is as follows: First, a source sentence is parsed into its constituent elements through Part-Of-Speech (POS)-tagging and phrase identification. Then, the individual elements and the sentence structure are analysed. Finally, transformational rules and dictionaries are applied to the source structure in order to generate the target structure (Slocum, 1985). This process involves numerous linguists writing rules and building dictionaries and can take two to three years to create the resources needed to generate one language pair.

There are several types of RBMT. *Direct Translation* refers to a word-by-word translation with minimal syntactic processing. Systran, an early commercial MT system, is an example of Direct Translation (Slocum, 1985). The MT transfer approach, by contrast, first analyzes a sentence in the source language and then converts its meaning into an intermediate syntactic representation, i.e. a form that suits the target language, before generating the final translation. To ensure accurate language adaptation, this process consists of three steps: Analysis, Transfer, and Synthesis (Slocum, 1985). In contrast, the *Interlingua Approach* to translation first converts a sentence into a representation having universal meaning and so not tied to any specific language. This representation is then used to generate the translation, based on the assumption that all languages share common underlying meanings (Slocum, 1985).

2.4.2 SMT

Inspired by the success of statistical methods in speech recognition, IBM Research introduced the concept of SMT in the late 1980s. The general idea underlying SMT is to build statistical models, i.e., mathematical representations, of training data. Initial SMT models regarded words as units and sentences as entities containing these units. Accordingly, aligning units from the source sentence to the units from the target sentence is often the first step of SMT methods (Koehn, 2010). One way to accomplish this is by employing the expectation maximization algorithm, which first estimates how likely different word alignments are and then counts how often they actually occur.

However, a more successful approach is usage of phrase-based models. These phrases can be any group of contiguous words not necessarily forming a proper grammatical unit. In this method, the input sentence is split into phrases, which are then mapped one-to-one to output phrases that may then be rearranged (Koehn, 2010).

SMT models also employ a decoding strategy. Probabilistic models first assign a score to every possible translation of a sentence. Decoding is then performed to identify the translation with the highest score. During decoding, the translation is built word by word, from beginning to end (Koehn, 2010).

Language models are a key part of SMT and help make the output sound natural. They affect word choice, reordering, and other decisions. From a mathematical perspective, they assign a probability to each sentence, showing how likely it is to appear in real text. Traditional SMT systems make frequent use of n-gram language models, which are based on the Markov assumption (the probability of a sentence equals the product of the probability of each word, given the history of the preceding words; (Koehn, 2010)). This probability is assessed based on the frequency of these words' contiguity in the text used to train the model.

In general, while RBMT models may be more accurate, SMT models tend to produce more fluent output than RBMT models. In addition, SMT models can be created much faster than rule-based ones, and are also quite robust to errors in the training data (Koehn, 2010).

2.4.3 NMT

A neural network is a machine learning method that processes multiple inputs to predict outputs. Neural Network Models (NNMS) enhance the sharing of statistical information among similar words and so supply a more comprehensive context yielding better translation quality (Koehn, 2020), in particular Deep Learning models, NNMs having multiple layers enable learning of advanced patterns. In addition, a nonlinear activation function enables these models to learn complex relationships. While NNMs had already been researched in the 1980s and 1990s, they were not adopted for MT until the mid-2000s, when they began being used to support traditional SMT models in numerous ways, from enhancing scoring mechanisms and translation tables to refining, reordering and pre-ordering models (Koehn, 2020).

Eventually, more ambitious efforts shifted toward use of full NMT (see Figure 4), completely replacing traditional statistical methods. Early approaches included Recurrent Neural Networks (RNNs), convolutional models, and sequence-to-sequence models, which performed well on short sentences but struggled with longer ones. Further refinements, such as incorporation of Byte-Pair Encoding (BPE) (Sennrich et al., 2016) to handle large vocabularies and back-translation of monolingual

target-language data ultimately established NMT as the new State of the Art (SOTA) (Koehn, 2020). Within two years, neural approaches completely dominated MT research, outperforming and replacing SMT models.

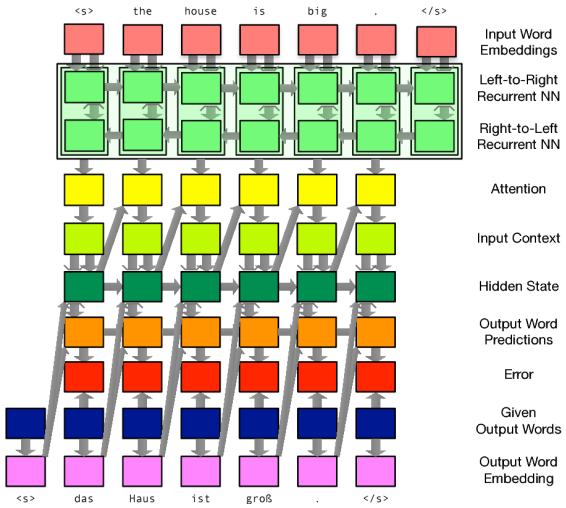

Figure 4: The full step-by-step process of training an NMT model on a sentence ('the house is big '). The model calculates and then sums an error for each translated word. During training, the model uses the correct previous words to help predict the next word. From Koehn (2017), licensed under CC BY-NC-SA 4.0.

Another major breakthrough was the introduction of the attention mechanism by Bahdanau et al. (2014), allowing the MT model to focus on different parts of the input sentence dynamically (see Figure 5). The Transformer model (Vaswani et al., 2017) extended this idea by replacing recurrence with self-attention as the primary mechanism for encoding and decoding sequences. Self-attention enables the model to compute relationships between all words in a sequence in parallel, rather than sequentially as in RNNs, making Transformers more efficient and better able to capture long-range dependencies. Transformer-based models, like Google's BERT and OpenAI's GPT, further improved translation quality.

In recent years, some attempts have been made to further improve NMT. Model optimization techniques have enhanced how a model learns, trains, and generates translations, and Ensemble Decoding uses multiple models instead of a single one, averaging all their individual probabilities to obtain better predictions. Guided Alignment Training improves how the model learns word alignments, thus increasing fluency and faithfulness in translation. Modeling Coverage ensures that all words in the source sentence receive adequate attention, thereby addressing an NMT model's tendency to sometimes ignore parts of an input sentence (Koehn, 2020).

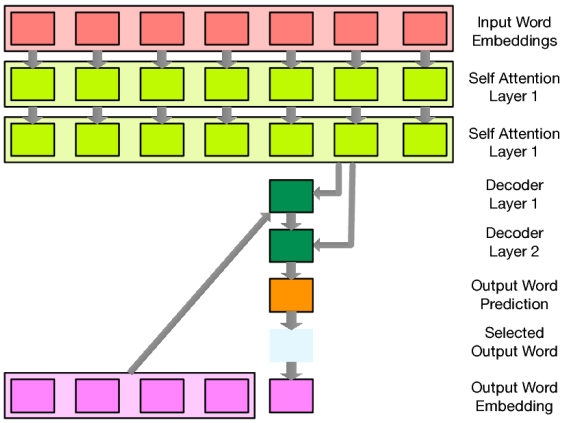

Figure 5: Attention-based MT model. From Koehn (2017), licensed under CC BY-NC-SA 4.0.

As in the attention-based MT model shown in Figure 5, the input sentence is encoded through multiple layers of self-attention, allowing the model to capture relationships between all words in the sentence. The decoder then generates the translation by attending to these encoded representations across several layers, using previously generated word embeddings as input.

Data augmentation strategies use additional or synthetic data to improve translation quality. The Back-Translation method generates synthetic parallel data for better training, helping in cases where a large amount of unlabeled text is available in the target language. Domain Adaptation fine-tunes models for specialized domains, helping NMT models that are trained on generic data. Multilingual Training improves model performance across multiple languages, enabling zero-shot translation (translating between language pairs that lack explicit training data) (Koehn, 2020).

As mentioned above, techniques like BPE have been developed to handle large vocabularies and rare words. BPE splits words into subword units to improve translation of rare, out-of-vocabulary, and morphologically challenging words (Koehn, 2020).

Lastly, strategies have been developed to incorporate linguistic knowledge into NMT models. Adding linguistic annotations, e.g. syntactic features, helps to improve grammar and sentence structure accuracy (Koehn, 2020).

Despite the efforts described above, NMT still faces challenges of various types. Data-related challenges stem from limitations in data availability, quality, and distribution, such as the availability of limited training data in low-resource languages. Further challenges arise from the model architecture including, e.g., issues involving word alignment, which—in contrast to SMT models—NMT models do not explicitly address. Still other challenges relate to optimization and generalization,

with overfitting, i.e., the tendency of NMT models to overfit to training data and then struggle with real-world variations comprising a prominent example.

2.4.4 LLMs for MT

In recent years, the advent of LLMs such as GPT-3 and GPT-4 has significantly transformed various aspects of the MT landscape. Like MT systems, NLP systems in general have transitioned from the rule-based systems used in the mid-90s to statistical systems in the late 1990s, and then to neural systems in the early 2000s (Kumar, 2024).

LLMs are scaled Pretrained Language Models (PLMs), which have been found to show performance improvements regarding downstream tasks. Both PLMs and LLMs are based on the Transformer architecture, which is developed to process long-range dependencies within sequences. It enables a model to weigh the importance of words in a sequence, effectively capturing contextual relationships. In addition, Transformers process all tokens simultaneously, leading to faster training times and improved scalability (Kumar, 2024).

The emergence of LLMs such as GPT (OpenAI), Gemini (Google), and Claude (Anthropic) is largely due to the exponential increase in large datasets and advancements in computational hardware. Increases in model size and training data volume consistently adhere to observed scaling laws, enhancing model capabilities. Because of their superior performance in understanding and generating human-like text compared to smaller models, LLMs have become central to AI research. Self-Supervised Learning is employed in pre-training an LLM, meaning that the model is trained on massive unlabeled corpora. The model can then be used to predict either the next token (autoregressive, e.g., GPT) or missing words (Masked Language Modeling, BERT). To fine-tune an LLM for a specific task, it is trained using labeled data, where each example consists of input-output pairs. Here, the objective is for the model to learn a function that maps inputs to outputs while minimizing a loss function. Reinforcement Learning (RL) is also used to fine-tune some LLMs during post-training using reward signals rather than labels (Kumar, 2024).

Research has also revealed that, as models increase in size, they develop such emergent abilities as ICL (Wei et al., 2021), driving continuous efforts to scale up LLMs (Brown et al., 2020), which, as a consequence, then demonstrate increasing levels of performance across various NLP tasks, including MT. Even if not optimized on multilingual data, LLMs show surprisingly high performance levels in MT for high-resource language pairs (Vilar et al., 2023). Several authors have found that the zero-shot translation capability of LLMs rivals that of highly competitive fully supervised MT systems (Lyu et al., 2024). Moreover, the translation capabilities of LLMs are constantly evolving; GPT-4 now surpasses GPT 3.5 (Zhu et al., 2024b).

In addition, LLMs have opened new possibilities to deal with challenging source texts, such as long documents having complex structures. Interestingly, LLMs have demonstrated strong potential in handling exceptionally long texts with intricate discourse structures, making them promising candidates for advancing document-level translation (Wang et al., 2023). With the introduction of larger context windows, translating longer documents with LLMs, such as novels and books, is becoming increasingly feasible.

Numerous new approaches are fueled by the possibility to prompt an LLM. There have been attempts at Stylised Translation with LLMs, i.e. to generate translations that align with a specific style or genre, made possible by the availability of large multi-parallel corpora. The zero-shot capabilities of LLMs now enable these tasks with ease, among others by directly prompting the model to translate in a specified style (Lyu et al., 2024).

Another approach is to utilize LLMs in a kind of TM function. For decades, Computer-Aided Translation (CAT) systems have employed TMs to assist human translators. In TM-based MT, the process typically involves searching for similar translations using fuzzy matching techniques and then refining or editing the retrieved translation to produce a high-quality result (Lyu et al., 2024). With the emergence of LLMs and their ICL capability—allowing them to learn specific tasks from examples provided in the prompt—the use of retrieved similar or relevant sentence pairs naturally aligns with few-shot prompting techniques for LLMs (Vilar et al., 2023; Moslem et al., 2023). These studies typically use sentence-level embeddings generated by external language models to retrieve similar examples through embedding similarity search (Zhu et al., 2024b).

MT in LLMs inherits many of the challenges of traditional NMT, and one such is data related. Like NMT, MT in LLMs struggles with low-resource languages. While strong for English-centric translations, LLMs fall behind commercial MT systems for languages having limited training data. Moreover, performance of MT in LLMs varies based on linguistic similarity: Indo-European languages perform better than, e.g., Sino-Tibetan languages (Zhu et al., 2024b).

A second challenge is model related. Like unsupervised NMT models, LLMs do not explicitly learn word alignments, possibly resulting in incoherent translations. A challenge unique to MT in LLMs is its over-reliance on ICL. If given mismatched translation pairs in the prompt, LLMs fail to provide a translation. Additionally, the prompt design itself seems to be important; translation accuracy depends heavily on how prompts are formulated. At the same time, even unreasonable prompt designs sometimes produce valid translations, indicating a need for further research in this area (Zhu et al., 2024b).

A third challenge is that, like NMT, MT in LLMs displays optimization issues. LLMs use greedy or beam search, but these methods are not always optimal, at

times producing overly short, generic, or biased translations. Also, LLMs struggle with domain-specific terminology, while lacking the targeted adaptation possible for NMT models (Zhu et al., 2024b).

However, even if they are not optimized on multilingual data, LLMs show a surprisingly high performance in MT for high-resource language pairs, surpassing traditional encoder-decoder MT models (Vilar et al., 2023; Moslem et al., 2023). Moreover, the translation capabilities of LLMs are constantly evolving (Zhu et al., 2024b).

Notably, LLMs can learn to translate in a resource-efficient way, requiring only a small number of, for example, non-English monolingual resources to build a bilingual mapping between non-English and English (Zhu et al., 2024b). In contrast to classical NMT, LLMs learn from diverse training data that are a natural mix of many languages and combine them on a word, sentence, and document level (Brown et al., 2020).

Furthermore, He et al. (2023) argue that LLMs might be capable of the same translation process as human translators, gathering and analyzing such information as keywords, topics, and relevant ICL examples. However, Zhu et al. (2024b) state that how LLMs acquire their translation capability is still unclear.

3 Related Work

3.1 Prompting LLMs With ICL Examples

Prompting an LLM consists of rephrasing labeled test examples following clear task instructions and then feeding these examples directly to the LLM (Zhang et al., 2023a), which then identifies the implicit pattern shown in the demonstration examples and so learns from analogy. LLMs have shown great potential in leveraging ICL, i.e., reproducing a text generation pattern without the need for fine-tuning. Zhu et al. (2024a), and Chitale et al. (2024) claim that ICL is comparable to the implicit performance of gradient descent during inference. ICL is thus a form of Meta-Learning—developing a set of skills and pattern recognition abilities during training and then applying these abilities during inference to adapt to or recognize the given task (Brown et al., 2020).

However, unlike simpler monolingual downstream tasks, prompting for MT in LLMs likely requires more sophisticated parameters for selecting ICL examples (Tang et al., 2024). Additionally, Brown et al. (2020) argue that, as the scale of LLMs grows—from hundreds of millions of parameters to billions of parameters—their ICL abilities increase as well.

As Zhu et al. (2024b) and Chitale et al. (2024) show, prompting with ICL examples is more effective than prompting with instructions. As the number of ICL examples increases, the importance of explicit prompt instructions diminishes, and even perturbance of instructions affects MT output only minimally. Perturbing the ICL examples, on the other hand, has a large adverse effect on MT output, especially perturbations on the target side of the examples (Chitale et al., 2024).[1] A mismatch or misalignment of translation pairs can have drastic negative effects on translation quality (Chitale et al., 2024).

3.1.1 Prompt Design

The quality of the MT output in LLMs has been found to depend on the design of the ICL template. Zhu et al. (2024b) found that, overall, the simple template "$\langle X \rangle =$

[1] This finding and others, e.g. by Tan et al. (2024), suggest that the target side of ICL examples is more important for successful promoting than the source side.

$\langle Y \rangle$" achieves the highest Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) score compared to the commonly used [SRC] : $\langle X \rangle \rightarrow$ [TGT] : $\langle Y \rangle$.

For the experiments done by Zhang et al. (2023a), the simple template with an added specification of the source language performs best. Lu et al. (2024b) also note that specifying the source language improves the outcome. Zhu et al. (2024a), on the other hand, argue that specification of the source language can hurt the outcome in some language directions.

In addition, Peng et al. (2023b) emphasize the importance of including task-specific information in the prompt (e.g., instructions such as "You are an MT system"), arguing that inclusion of such information increases performance, particularly when translation is into low-resource languages. They also specify the domain in the prompt, finding that this consistently improves the LLM's output in terms of semantics. Zhu et al. (2024a) report more verbose and instruction-like prompts as the best prompts ("a good translation-source + target" and "version-target"). In contrast, Vilar et al. (2023) argue that prompt design only has a significant impact if the number of examples in the prompt is low.

Some researchers utilize advanced prompt designs that fully leverage the LLMs' reasoning potential. Puduppully et al. (2023) split the input sentence into chunks, translate each chunk independently, and then refine the translations using context. However, their focus is on translations between languages within the same language family.

Lu et al. (2024b) employ the "Chain of Dictionary" (CoD) method, which is inspired by CoT, to break down the reasoning process into steps. By eliciting reasoning in the LLM, they simultaneously elicit translation capability. Based on this method's good performance, we can infer that the ability to reason is closely tied with translation ability—that is, the ability to recognize connections between different languages. Like Peng et al. (2023b) and Zhu et al. (2024a), the authors utilize a more elaborate and verbose instruction, which apparently performs better for low-resource languages. They also argue that putting source and target language at the head of the reasoning chain is the most effective strategy. Since Lu et al. (2024b) basically only use one ICL example, albeit a chain of words in different languages, the instruction seems equally important here.

3.1.2 Prompt Language

On average, Zhang et al. (2023a) and Lin et al. (2022), have found that an English instruction template yields the best performance in MT, whereas a template in a language that does not constitute a major pretraining language of the LLM worsens the performance (Zhang et al., 2023a).

Moreover, Chitale et al. (2024) found that the language of the ICL source examples does not necessarily need to be the same as the primary source language,

and can indeed be successfully substituted with examples in a different language. Tan et al. (2024) only extract the target language side as ICL examples, arguing that this is sufficient for the LLM to learn the style of the source side and thus produce the desired output. However, they find that classic few-shot prompting including target and source side still proves as more effective.

Lu et al. (2024b) extend this strategy further by prompting the LLM with a chain of words in languages different from the primary source language, with chain words in high-resource languages performing better than those in low-resource languages, arguing that high-resource languages in particular trigger cross-lingual abilities in LLMs.

3.1.3 Number of Examples in the Prompt

The BLEU score has been found to improve as the number of ICL examples increases to 8 but declines thereafter (Zhu et al., 2024b). Hendy et al. (2023) argue that choosing 5 examples is optimal for good performance. Zhang et al. (2023a), Vilar et al. (2023), and Zhu et al. (2024a) also note that 5-shot prompts often outperform prompts having 10 or even more examples. Interestingly, Lu et al. (2024b) also report 5 examples in their reasoning chain as yielding the best performance. [2]

3.1.4 Prompt Examples' Similarity to Input

Semantic Similarity

The role of similarity in selecting ICL examples for MT remains a topic of debate. According to Zhu et al. (2024b) and Vilar et al. (2023), MT with semantically related ICL examples **does not outperform** MT with semantically unrelated examples. Similarly, Zhang et al. (2023a) argue that semantically related examples do not consistently yield better results. Supporting this, although Vilar et al. (2023) employed a k-Nearest Neighbours (kNN) approach to retrieve the most semantically similar examples from a sentence pool, they found no significant improvement over use of randomly selected examples. Zhang et al. (2023a), using multiple semantic similarity metrics—including cosine similarity of sentence embeddings, CaseSemScore-Src, and CaseSemScore-Tgt—also report no meaningful correlation between these similarity measures and quality of translation performance.

In contrast, Zhu et al. (2024a) report that selecting ICL examples based on fuzzy match and BM25 retrieval techniques outperforms random selection, although a

[2] Pool size for prompt retrieval is another possible parameter that could be investigated. Agrawal et al. (2023) argue that a larger pool of prompt examples does indeed improve the BLEU score, but the same can be achieved by providing multiple ICL examples.

possible explanation could be the specific method they employed. Adding nuance to this discussion, Kumar et al. (2023) suggest that similarity between the input source and the example target is more predictive of success than similarity between input and example source. Additionally, Lu et al. (2021) warn that randomly selected examples can lead to inconsistent performance and higher computational costs.

Lexical Similarity

Agrawal et al. (2023) found that lexical similarity of ICL examples to the input benefits translation performance. Having systematically compared the impact of semantic and lexical similarity of the ICL examples to the input, Zebaze et al. (2025) found that semantically similar ICL examples lead to a better translation output than lexically similar examples, especially for low-resource languages. In particular, source-to-target retrieval—i.e., matching the input source sentence to the target-side translation of ICL examples—shows good performance.

Beyond similarity, other related factors can also impact performance. For instance, quality of performance tends to decrease when duplicate examples are used, indicating that **diversity** among ICL examples is beneficial (Zhu et al., 2024b; Zebaze et al., 2025). Along these lines, Zhu et al. (2024a) propose refined selection strategies that incorporate both sentence-level and word-level information. At the sentence level, they introduce a margin-based similarity score, while at the word level, they leverage word embeddings to reduce noise and improve selection precision. Sia and Duh (2023) emphasize the importance of **coherency** between the ICL examples and the test sentence. Employing a moving window of prior gold translations immediately preceding the current test sentence, they found that coherency is a stronger predictor of translation quality than semantic similarity.

Other researchers have explored connections between semantic similarity and traditional translation methods. Moslem et al. (2023), for instance, liken the use of semantically related examples to fuzzy matching in TMs, and retrieve these examples from the dataset using embedding-based similarity. Exploring bilingual terminology extraction, they showed that integrating terms from a glossary further enhances performance—albeit occasionally at the expense of grammatical quality.

3.1.5 Sizes of Examples in the Prompt

ICL examples should be the length of a sentence; as the number of usages of word or document examples increases, translation performance degrades (Zhu et al., 2024b). Zhang et al. (2023a) found that longer examples help with translation performance, providing the LLM with more information about how inputs (the given text) relate to outputs (the expected response). Similarly, Sia and Duh

(2023) found that relatively short prompts (5–10 words) have a negative effect on translation quality. Longer prompts (15–20 words) have a positive effect in cases where they correspond with the median length of sentences in the source domain.

3.1.6 Quality of the Examples in the Prompt

Quality [3] of prompt examples does not always translate directly to improved MT performance. Zhang et al. (2023a) evaluated prompt examples using MTScore—an automatic Quality Estimation (QE) based on the COMET-QE model—and found that high example quality alone does not necessarily guarantee optimal translation results. Sia and Duh (2023) argue that ICL examples can steer LLMs toward either improved or degraded MT output, depending on the nature and design of the examples. Expanding on this idea, Jiao et al. (2023) explored the use of both low-quality and high-quality examples within a single prompt, framing this contrast as a form of constructive instruction to guide the LLM's behavior. In one case, they even explicitly annotated the type of error in the low-quality example to enhance the model's learning process.

3.1.7 LLM Settings

Lastly, the LLM itself can be tuned to better accommodate MT tasks. Peng et al. (2023b) argue that a lower temperature ensures that the LLM creates grammatically correct and deterministic output, which is ideal for MT tasks requiring a high degree of faithfulness to the source text and to a certain terminology.

3.1.8 Summary

The effectiveness of ICL-based MT depends heavily on prompt design. While ICL examples generally outperform simple instruction-based prompts, their effectiveness is sensitive to several factors. The number of examples, for instance, has an optimal range—typically around five—beyond which performance may decline. Prompt templates, the inclusion of task-specific instructions, and reasoning structures such as CoT methods can further enhance translation quality, particularly for low-resource languages.

Prompt language and semantic similarity of ICL examples also play important roles. Although semantically related examples might intuitively seem beneficial, findings show that their impact is inconsistent, and diversity among examples appears to matter more than strict similarity. Some studies suggest that coherence

[3] "Quality" in this context mainly refers to the noisiness of examples, caused by, e.g. spelling mistakes, grammatical mistakes, terminological mistakes.

between examples and the test sentence is more predictive of quality than semantic closeness alone. Yet, most of the studies emphasize that it is crucial that the ICL examples stem from the same domain, regardless of their semantic similarity to the input that needs to be translated or their quality (Agrawal et al., 2023; Sia and Duh, 2023). Treating a low-resource language as a domain can even prompt an LLM to generate satisfying translations of content in a low-resource language (Merx et al., 2024).

The size and quality of prompt examples are equally important. Sentence-length examples tend to outperform very short or overly long ones, and, while high-quality examples help, their presence alone does not guarantee better results. In some cases, mixing low- and high-quality examples can even provide the model constructive guidance.

Finally, model-level settings such as temperature can influence output stability and faithfulness. Lower temperature settings tend to produce more grammatically consistent and deterministic translations—an advantage for terminology-sensitive MT tasks.

3.2 Strategies for Retrieving ICL Examples

As we have seen, many researchers perform random selection of ICL examples, a strategy that has been shown to deliver satisfying results.

3.2.1 Maximizing an Evaluation Score

Some researchers maximize an evaluation metric such as BLEU to choose examples that improve the metric. Agrawal et al. (2023) maximize the BLEU score by selecting prompts based on the BLEU of the generated outputs relative to the references on a held-out development set.

3.2.2 Utilizing Unsupervised and Supervised Retrievers

Recent SOTA strategies retrieve examples that are similar to the source sentences in the test set, with the help of such **unsupervised retrievers** as, for example, **sparse retrievers** like BM25 (Zhu et al., 2024a; Agrawal et al., 2023; Tan et al., 2024; Zebaze et al., 2025). BM25 sorts training pairs— source pairs and their translations—based on their relevance to the query sentence using n-gram overlaps. Agrawal et al. (2023) use BM25, but with a caveat; they argue that BM25 might prioritize rare word matches and miss key terms from the source text. To maximize term coverage, they therefore re-rank the top 100 BM25 results using their proposed algorithm. Sia and Duh (2023) apply BM25 in combination with semantic similarity (Nearest Neighbours).

Other researchers implement **supervised retrievers**. Vilar et al. (2023) employed a kNN search to find suitable ICL examples. Merx et al. (2024) retrieved ICL examples based on semantic embeddings, but enriched them with examples retrieved based on lexical overlap (TF-IDF). Tang et al. (2024) created their own retriever that selects examples based on syntactic and lexical coverage alternately. Their algorithm converts a syntax tree into polynomials to combine syntactic and lexical information. Peng et al. (2023b) utilized the TopK method, which applies diverse ranking methods to extract the k most relevant examples from a set [4].

3.2.3 Combining Score- and Retriever-Based Methods

Kumar et al. (2023) present a combined score- and retriever-based method named CTQScorer, a regression model that selects ICL examples based simultaneously on multiple features, arguing that learned metrics are better at selecting examples than dense retrieval techniques. Their method combines a BM25 retriever with the

[4]This method could be compared to example-based MT, a subform of SMT

scoring function they developed based on the COMET score, an estimate of the translation quality the respective ICL example can provide.

A technique employed by Pourmostafa Roshan Sharami et al. (2024) combines an unsupervised retriever with domain-specific QE to select the ICEs most suitable for the task. From the retrieved examples, only the ones in the prompt that show a high QE score are included. Also sought by the authors were the best number and combination of ICEs to improve translation quality.

3.2.4 Sources for Prompts

KGs

To date, few researchers have attempted to integrate information from KGs into prompts. Zhang et al. (2023a) first extracted entities from the source sentence and then queried the MKG to retrieve bilingual entity pairs corresponding to the extracted entities. The triples introduced into the prompt were then used as term-translation pairs.

Dictionaries

Before the advent of LLMs, attempts were made to incorporate dictionaries into NMT. Arthur et al. (2016) augmented NMT with lexicons containing translations of low-frequency words. Zheng et al. (2021) improved kNN-NMT by augmenting it with a kNN datastore using only in-domain monolingual, target-language sentences.

Methods intended to augment LLMs with dictionaries generally tend to be straightforward. Merx et al. (2024), for example, retrieved knowledge from dictionaries and included it in their prompt to the LLM.

Ghazvininejad et al. (2023) extracted bilingual information on words and appended it to the prompt fed into the LLM. For a given source sentence, they first found dictionary entries for the words contained in the sentence. They then created a prompt that included both a request to translate the full sentence and a list of possible translations for specific words. They also experimented with few-shot prompting, feeding the LLM different sentences with dictionary entries before prompting with the sentence actually needed.

Lu et al. (2024a) developed Dictionary Insertion Prompting (DIP), which looks for the English counterparts of low-resource source words and inserts them into the prompts to the LLM, not only enhancing the translation into English but also the LLM's reasoning process in the low-resource language, e.g., for math tasks. The authors found that directly chaining the English dictionary words to the non-English source words performs better than simply feeding the dictionary information in bulk.

The same researchers (Lu et al., 2024b) prompted an LLM with a chain of translated words in different languages extracted from a dictionary. They linked word translations across multiple languages to enhance the LLM's multilingual translation capabilities. Like KGs, dictionaries provide easy access to ICL examples (Lu et al., 2024b).

4 Methodology

As described in the Introduction Chapter, the purpose of our research was to investigate the use of MKGs as a source of in-context examples employed to improve domain-specific MT using LLMs. Rather than relying solely on sentence-level parallel data, we explored how structured multilingual terminology from a KG can be leveraged to construct prompts that better employ LLMs' cross-lingual and reasoning capabilities. Our particular focus were prompting strategies inspired by CoT methods, which assume that auxiliary high-resource languages can help bridge semantic gaps between source and target languages, especially low-resource languages (Lu et al., 2024b). By grounding prompts in an MKG, we hypothesized that this approach would combine the strengths of structured knowledge representation and flexible LLM inference, offering a lightweight and adaptable solution for domain-adapted MT. In this chapter, we detail the methodology we employed to carry out our study. [1]

The primary method we employed in our study was based on the research of Lu et al. (2024b) described in the previous chapter. Lu et al. (2024b) speculate that few-shot ICL examples are not relevant to the translation performance, but instead conjecture that use of an 'auxiliary language' in the prompt can provide hints concerning meaning where no direct translation between the two languages exists.

Lu et al. (2024b) found that their method improved MT in ChatGPT for 67% (135 out of 200) of the languages they examined, with good improvement in most languages and excellent improvement in several. Moreover, this method succeeded in translating low-resource languages that ChatGPT failed to translate without dictionary knowledge. The authors compared their method with several others, including few-shot prompting, which their findings showed as lagging behind CoD, particularly in the translation of low-resource languages.

We tested the method proposed by Lu et al. (2024b), comparing it to the few-shot prompting baseline to determine whether high-resource languages truly activate an LLM's cross-lingual capabilities in a more effective way. Instead of a dictionary, we employed a KG as the source for multilingual prompts.

In both of the methods we tested, which are described below, we selected English as our source language and German as our target language. For our primary

[1] The code for this thesis can be found here: https://github.com/CarinaObster

method, we selected French, Spanish, and Italian as supporting languages for the CoD.

4.1 Dataset

4.1.1 KG

We employed EuroVoc, the European Union's (EU) multilingual thesaurus, as a KG resource. Created to process documents produced by EU institutions (Publications Office of the European Union, 2015), the EuroVoc KG covers topics related to both the EU and its member states, with a particular focus on the EU Parliament. Originally designed to manage and distribute information, EuroVoc currently allows users to search documents using a controlled vocabulary and semantic network, and its most recent edition (4.4) is available in 23 official EU languages.

EuroVoc uses semantic web technologies that adhere to the latest thesaurus standards and recommendations of the World Wide Web Consortium (W3C). It is based on the Simple Knowledge Organization System (SKOS) model, which is comprised of three fundamental elements: concepts, terms, and relations. A *concept* refers to an entity or idea that we aim to model or communicate about within a specific domain. A subject's concept can be viewed as the collective set of terms and labels used to refer to it in different languages or linguistic systems. Within EuroVoc, each concept is assigned a unique Uniform Resource Identifier (URI) to ensure consistent identification. A *term* is a linguistic expression, either a word or a phrase, that is used to refer to a specific concept within a particular language. In order to maintain compatibility with traditional thesauri, a single term is designated as the preferred label for each language.

The concepts in EuroVoc are connected through hierarchical and non-hierarchical (associative) relationships. In SKOS, the hierarchical relationships between concepts are represented using the properties skos:broader and skos:narrower, which correspond to the traditional thesaurus relationships 'broader term' (BT) and 'narrower term' (NT). For associative, non-hierarchical connections, SKOS uses the property skos:related, which corresponds to the conventional "related term" (RT) relationship in thesauri. Currently, EuroVoc comprises 6883 concepts, 4904 hierarchical relationships, and 6922 associative relationships.

Within the framework of our experiments, we extracted terms in different languages. The concepts as such and the relationships between them were not relevant to our study.

4.1.2 Corpus for the Baseline: EUR-Lex

We employed EUR-Lex[2], the official legal database of the EU maintained by the Publications Office of the European Union, as a corpus for extracting ICL examples for the baseline method. This EU-established corpus contains legal documents created by EU institutions and includes, among others, treaties, legal acts, and case-law. It contains over three million documents, ranging from the founding treaties of the 1950s to the present day. Each document has a CELEX number, a unique identifier, as well as an ELI number, a more recent standard that enables semantic web integration. Depending on the EU official languages that existed when the documents were adopted, EUR-Lex documents are available in up to 24 languages. The terminology used in EUR-Lex is aligned with that defined in EuroVoc. Additionally, many EUR-Lex documents are annotated with EuroVoc descriptors.

For the purposes of this research, we selected a subset of the EUR-Lex dataset: EUR-Lex 2/2016, a parallel corpus consisting of **English** and **German** texts. It constitutes one version of the parallel corpora derived from EUR-Lex compiled and released in 2016. We accessed the parallel corpus via SketchEngine, which is a platform that supports advanced text exploration and concordancing features. EUR-Lex 2/2016 encompasses around 19,000 documents and about 2 million sentence pairs. We did not split the dataset into training, validation, and testing datasets since we were not seeking to train the LLM but rather to test it directly by prompting with ICL examples.

We examined different subsets and selected terms from different EuroVoc domains that existing LLMs could have struggled to translate adequately. We identified the following terms as being suitable for our experiments:

- **EU-specific terminology** (Mori, 2018): Terms that are unique to the EU (e.g., EAGGF Guidance Section - *EAGFL-Ausrichtung*, Community transit - *gemeinschaftliches Versandverfahren*)
- **EU translation equivalents** (Pontrandolfo, 2019): Terms that are translated in a distinctive way in the EU glossary (e.g., vineyard - *Rebfläche*)
- **Context-dependent institutional equivalents** (Kjær, 2007): Terms that seem unambiguous on the surface but are translated in a more complex way in the EU glossary due to conceptual shifts (e.g. food consumption - *Nahrungsmittelverbrauch*, monetary support - *Währungsbeistand*)
- **Terms of art** (Šarčević, 1997): Terms that have a basic meaning but can have a slightly different meaning in a specific context (e.g. carcase - *Schlachtkörper*)

[2]https://eur-lex.europa.eu/content/help/eurlex-content/documents-in-eurlex.html?locale=en

Employing the SketchEngine website, we then searched the parallel corpus for suitable sentences containing these terms. SketchEngine searches by matching linguistically annotated tokens in an indexed corpus, retrieving occurrences based on exact linguistic criteria. We also cleaned the sentences retrieved for the dataset by removing white spaces, adding or removing punctuation, and contracting sentences where necessary. Additionally, we disregarded terms with clearly incorrect German translations in the KG, as well as terms for which the translations in the corpus and the KG differed.

4.2 Models

4.2.1 GPT-5

The multimodal LLM GPT-5 processes both image and text inputs to generate text outputs.[3] While its predecessor GPT-4 is claimed to have achieved top 10% scores with respect to professional and academic benchmarks, indicating advanced reasoning abilities, this ability has not yet been confirmed for GPT-5. Also unconfirmed is whether GPT-5 outperforms most SOTA and previous systems on the MMLU (Massive Multitask Language Understanding) benchmark and its translated variants. GPT-5's most significant improvement relative to its predecessor is its extended context window, which enables the model to remember and reason over a much longer text. It employs a Transformer-based design, pre-trained to predict subsequent tokens, facilitating coherent and contextually relevant text generation. Despite advancements, GPT-5 is not fully reliable and hence may exhibit "hallucinations" or inaccuracies, underscoring the need for cautious application. In translation, hallucinations often take the form of semantically plausible but non-existent information, lexical inventions, or deviations from the source text.

4.2.2 DeepSeek LLM: DeepSeek-V3 (7b chat)

DeepSeek-V3 is a Mixture-of-Experts (MoE) NNM. For each input, only a small subset of specialized sub-models known as "experts" are activated, allowing the model to scale to very large sizes while keeping computation efficient. It also uses Multi-Head Latent Attention (MLA), a memory-efficient attention mechanism that compresses key, value, and query representations into lower-dimensional latent vectors. In terms of benchmarks such as MMLU (Massive Multitask Language Understanding), GPQA (Graduate-level Google-proof QA) and MATH (Mathematics Assessment of Textual Heuristics), DeepSeek-V3 is the most powerful open-source

[3] openai.com/index/gpt-5-system-card (2025)

Table 1: Model Card for GPT-5 (OpenAI, 2025)

Model Name	GPT-5
Release Date	September 12, 2025
Organization	OpenAI
Architecture	Transformer-based multimodal model with routed sub-models (Fast and Reasoning modes)
Parameter Count	Undisclosed
Training Tokens	Undisclosed (Estimated >100T tokens)
Languages	Multilingual (Improved performance across high- and low-resource languages)
Context Length	Up to 256k tokens (Dynamic windowing)
Training Objective	Causal Language Modeling (Next-Token Prediction)
Fine-Tuning	Instruction tuning + RL from Human and AI Feedback
Notable Features	Enhanced multimodal reasoning (text, vision, structured data); improved factuality; system routing for efficiency; higher reliability and reduced hallucination rates
Use Cases	Advanced chat, coding, multimodal understanding (image + text), data analysis, creative writing, research assistance, tutoring
Licensing	Closed-source (Available via OpenAI API and ChatGPT Pro)
Performance Highlights	SOTA on MMLU, GSM8K, and SWE-bench Verified; ∼45% reduction in factual errors vs. GPT-4o; leading results on professional and academic benchmarks

base model, particularly for maths and coding. In several areas, its performance is comparable to that of closed-source models such as GPT-4 (DeepSeek-AI et al., 2024).

4.2.3 TowerLLM

Tower is an open, multilingual LLM that has been specifically optimised for translation workflows. Unlike general LLMs, Tower has been trained and fine-tuned specifically for translation and related tasks. Its authors first created TowerBase through the continued pre-training of LLaMA-2 using a dataset also containing parallel data. They then performed SFT on a dataset containing translation-related tasks (TowerBlocks) to create a model for following instructions tailored to the field of translation—TowerInstruct. Tower outperforms GPT-4, GPT-3.5, Google Translate, and DeepL on several human-evaluated translation tasks (Alves et al.,

Table 2: Model Card for DeepSeek-V3 (DeepSeek-AI, 2024)

Model Name	DeepSeek-V3
Release Date	December 27, 2024
Organization	DeepSeek-AI
Architecture	Mixture-of-Experts (MoE) Transformer
Parameter Count	671B total parameters (37B active per token)
Training Tokens	14.8 trillion tokens
Languages	Primarily English and Chinese (with emerging multilingual support)
Context Length	Up to 128k tokens
Training Objective	Causal Language Modeling + Multi-Token Prediction (MTP)
Fine-Tuning	Supervised Fine-Tuning (SFT) + RL + Reasoning Distillation
Notable Features	Multi-head Latent Attention (MLA); FP8 precision training; DualPipe parallelism for faster training and inference
Use Cases	Mathematical reasoning, code generation, QA, general NLP, and logic-intensive tasks
Licensing	Open-source
Performance Highlights	Leading open-source model in math, code, and reasoning tasks; rivals GPT-4o and Claude 3.5 on several benchmarks

2024).

4.3 Extracting ICL Examples from the KG

4.3.1 Baseline Method: Few-Shot Prompting

We employed simple cosine similarity as a baseline for extracting prompts. In prior work, embedding-based similarity has been adopted as a baseline method for selection of ICL examples. For instance, Adiga et al. (2024) computed cosine similarity explicitly as part of their metric; Chen et al. (2024) used retrieval-style ICL relying on similarity ranking; and Zebaze et al. (2025) demonstrated its utility in low-resource MT.

As exemplified in the paper by Conia et al. (2024b), we used Contriever, Facebook's unsupervised, dense IR model, which is trained using contrastive learning and does not rely on labelled data. It uses a bi-encoder architecture based on BERT, which independently encodes queries and documents. This architecture

Table 3: Model Card for TowerLLM (Unbabel, 2024)

Model Name	TowerLLM
Release Date	February 2024 (paper), June 2024 (public release)
Organization	Unbabel
Architecture	Transformer (Decoder-only, autoregressive)
Parameter Count	7B and 13B variants
Training Tokens	Not explicitly disclosed
Languages	Multilingual (Optimized for 10 major language pairs, capable of 200+)
Context Length	Not explicitly disclosed
Training Objective	Causal Language Modeling (Next-Token Prediction)
Fine-Tuning	Multi-phase fine-tuning: instruction tuning + translation fine-tuning
Notable Features	Optimized for translation workflows; designed to integrate with professional translation pipelines
Use Cases	MT, translation post-editing, QE, and multilingual chat
Licensing	Open-source
Performance Highlights	Outperforms GPT-4o, GPT-3.5, DeepL, and Google Translate in MT benchmarks for select language pairs

learns effective representations by pulling together similar text spans from the same document and pushing apart unrelated ones. The model has been found to outperform BM25 on most BEIR (Benchmarking IR) datasets in zero-shot settings and to display robust generalization across domains and languages, including multilingual and cross-lingual retrieval scenarios.

However, in employing Contriever, we encountered a problem. For some terms, either no prompt term was extracted, or fewer than the three we required were extracted. The reason was that candidate terms having a cosine similarity below 0.5 were filtered out, meaning that, sometimes, fewer than three terms passed the threshold, resulting in fewer matches being returned. To address this issue, we implemented a back-off method: an iterative wildcard search triggered if the KG search containing the exact term returned zero or only one prompt term. The input term prefix was iteratively shortened (e.g., vinification: vinif*, vini*, vin*) until the Application Programming Interface (API) returned more than one prompt term hit. The cosine similarity between the input and the prompt term hits was then calculated, and the top three results were chosen (see Table 4). Using this method preserved the semantics of the search term, which would not have been the case if we had used, for example, query expansion with synonyms.

Wildcard search	Sentence found
vini*	The annex must be supplemented by headings giving the classification of vine varieties in Greece.
vin*	Presses, crushers and other machinery, of a kind used in wine making, cider making, fruit juice preparation or the like.
vin*	Each specified region shall be precisely demarcated, as far as possible on the basis of the individual vineyard or vineyard plot.

Table 4: Prompt Sentences Retrieved Using Wildcard Search Queries

4.3.2 Proposed Method: CoD Prompting

Instead of a dictionary, as in Lu et al. (2024b), we use an MKG. Each unique term was queried against the Coreon API—the system we used to access the EuroVoc KG—, from which only the first retrieved concept was retained. Within this concept, the system extracted a single representative term per language—specifically for French, Italian, and Spanish—thus constructing a compact multilingual glossary.

For every sentence, the corresponding multilingual Coreon entries of the detected terms were assembled into per-sentence glossaries, which were then linearized into short "means-chains" (e.g., wine means *vin* means *vino* means *vino*), explicitly conveying cross-lingual term correspondences.

4.4 Constructing the Prompt

4.4.1 Baseline Method

For each sentence to be translated, the relevant ICL example sentences associated with the corresponding term were retrieved. They were then formatted into a short, readable block including the English ICL example sentence and its aligned German translation. In the end, this ICL example block was concatenated with an explicit translation instruction ("Translate the following text from English into German") (see Figure 4.4.1). This prompt design followed the Best Practices identified during our research: a short and simple instruction in English specifying source and target language and less than eight ICL examples. By using the EUR-Lex corpus, we were assured of attaining domain-consistent, sentence-length, and high-quality examples. Although extracting a higher number—between four and eight—of ICL examples as well as more diverse examples within the same domain might have been beneficial, the corpus we used in the context of our research was not sufficiently large to do so. Additionally, we set a low temperature for the model.

4.4.2 Proposed Method

Following the suggestion of Lu et al. (2024b)'s paper, we inserted the "means-chain" generated after the extraction stage into the user prompt, preceding the sentence to be translated. The final prompt presented to the model then consisted of three

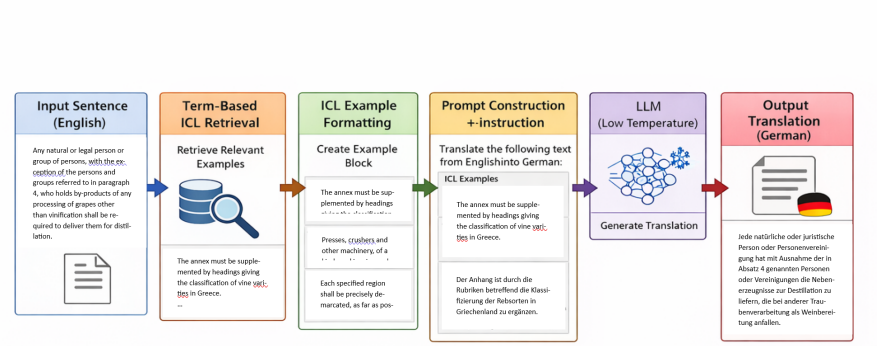

Figure 6: Pipeline for Baseline Method. Image generated with AI.

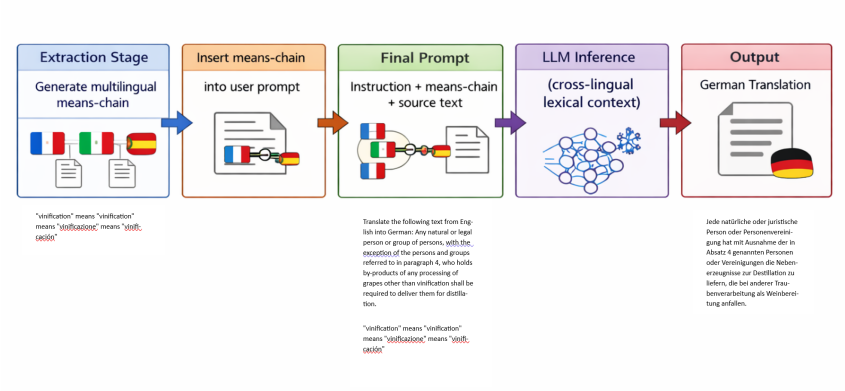

Figure 7: Pipeline for Proposed Method. Image generated with AI.

main components: an explicit translation instruction ("Translate the following text from English into German"), the multilingual "means-chain", and the source text itself. The resulting prompt provided the model with explicit lexical and cross-lingual context, encouraging it to incorporate the provided French, Italian, and Spanish equivalents when generating the German translation (see Figure 4.4.2).

This chapter has detailed the methodology which we employed in our research. The next chapter presents our results.

5 Results

To evaluate our model's output, we initially applied the standard automatic metrics—BLEU, Chrf, and Comet. Operating at segment level, with segments generally consisting of sentences, BLEU averages individual scores over a whole corpus (Papineni et al., 2002). More fine-grained than BLEU as it is based on character n-grams instead of word n-grams, chrF initially evaluates at segment level (Popović, 2015). In contrast, Comet is trained to predict human judgments, but also at sentence level (Rei et al., 2020).

As, for the most part, the translation of single terms was to be evaluated, the question arose as to whether the traditional metrics alone were sufficient for this task, since they operate on segment or sentence level. Therefore, we also applied the Multidimensional Quality Metrics (MQM) model (Lommel et al., 2013), shown in Eq. (1) below, on a subset of the results (100 sentences).[1]

$$\text{MQM} = 100 \left(1 - \frac{10\,\text{critical} + 5\,\text{major} + \text{minor}}{\text{tokens}}\right) \tag{1}$$

The main MQM category we employed was Accuracy, i.e., how far the target term accurately corresponded to the propositional content of the source term. If the criterion of *Accuracy* was adequately fulfilled, we then applied the *Terminology* category, i.e., how far a term conformed to normative domain or organizational terminology standards. We ignored other MQM categories, such as *Style*, *Linguistic Conventions*, and *Locale Conventions*, since they were not the focus of our research. Where the model employed a synonym present in the KG, we noted this in the MQM evaluation as "synonym".

Model	BLEU	ChrF	COMET
GPT-5	22.34	56.40	0.53

Table 5: Translation Performance GPT-5 Baseline Method

5.1 Baseline Method

5.1.1 GPT-5

Main Scores

The model achieved a BLEU score of 22.34, indicating a moderate level of n-gram overlap with the reference translations and so suggesting reasonable but improvable fluency and adequacy. Its ChrF score of 56.40 was relatively stronger, showing that the model captured character-level patterns well, a phenomenon often indicating improved handling of morphology and word forms. Finally, the COMET score of 0.53 pointed to acceptable overall translation quality from a semantic perspective, although there was clearly still room for improvement compared to higher performing systems.

MQM Score

The baseline method using GPT-5 produced an accurate output, with only a few serious issues mainly regarding use of wrong terminology, e.g., farm return - *Meldung der landwirtschaftlichen Betriebe* instead of *Betriebsbogen*, fishing permit - *Fischereigenehmigung* instead of *Fangerlaubnis*.

All in all, the translations met a high quality standard as shown in Tables 5 and 6.

5.1.2 Deepseek

Main Scores

Deepseek recorded a BLEU score of 11.63, suggesting limited overlap with the reference translations and indicating relatively weak fluency and accuracy compared to other systems. Its ChrF score of 45.83 indicates that it captured some character-level information, but less effectively than the first model, pointing to difficulties with word forms and consistency. The COMET score of 0.17 is quite low, reflecting

[1] In order to make a more accurate statement about the performance of the two methods in different domains, MQM would have to be applied to the entire result files, and the domains would have to be compared with each other. Unfortunately, this was not possible within the scope of a master's thesis.

MQM Error Category	Count
Terminology – Wrong Term: Minor	7
Terminology – Wrong Term: Major	12
Accuracy – Mistranslation: Minor	1
Accuracy – Mistranslation: Major	5
Accuracy – Untranslated: Major	1
Total	**26**
Score	**96.51**

Table 6: MQM Errors GPT-5 Baseline Method

Model	BLEU	ChrF	COMET
Deepseek V3	11.63	45.83	0.17

Table 7: Translation Performance Deepseek V3 Baseline Method

poor semantic alignment with the reference translations and overall low translation quality (see Table 7).

MQM Score

Deepseek produced some accuracy errors related to the baseline method (see Table 8), e.g., crop loss - *landwirtschaftlicher Verlust* instead of *Ernteausfall*, farm return - *Farmübermittlung* instead of *Betriebsbogen*, carcase - *Körper* instead of *Schlachtkörper*. Additionally, omissions were frequent, e.g., glasshouse cultivation - *Gewächshaus* instead of *Gewächshauskultur*. In addition, Deepseek made grave terminology mistakes, e.g., table wine - *Tischwein* instead of *Tafelwein*, fuel crop - *Treibstoffpflanze* instead of *Energiepflanze*.

5.1.3 TowerLLM

Main Scores

TowerLLM achieved a BLEU score of 18.62, indicating a moderate level of overlap with the reference translations and better performance than Deepseek, although still below GPT-5. Its ChrF score of 51.70 suggests a fair ability to model character-level patterns and word forms, reflecting reasonable handling of morphology. The COMET score of 0.38 points to moderate semantic adequacy, showing that the model captured some meaning correctly but left notable room for improvement in overall translation quality (see Table 9).

MQM Error Category	Count
Terminology – Wrong Term: Minor	16
Terminology – Wrong Term: Major	9
Terminology – Wrong Term: Critical	4
Accuracy – Mistranslation: Minor	2
Accuracy – Mistranslation: Major	13
Accuracy – Mistranslation: Critical	11
Accuracy – Omission: Major	1
Accuracy – Omission: Critical	4
Accuracy – Untranslated: Major	2
Total	62
Score	87.84

Table 8: MQM Errors Deepseek Baseline Method

Model	BLEU	ChrF	COMET
TowerLLM	18.62	51.70	0.38

Table 9: Translation Performance TowerLLM Baseline Method

MQM Score

TowerLLM also produced a significant number of accuracy errors (see Table 10), e.g., farm return - *Betriebsprüfungsunterlagen* instead of *landwirtschaftlicher Betriebsbogen*, share farming - *Gemeinschaftsvermietung* instead of *Teilpacht*, intra-Community trade - *Verkehr innerhalb der Gemeinschaft* instead of *Intra-EU-Handel*. Terminology errors were also quite common, e.g., replanting - *Umpflanzung* instead of *Wiederbepflanzung*, EAGGF Guidance Section - *EAGGF-Förderungssäule* instead of *EAGFL-Ausrichtung*, bonded wood - *Holzbinder* instead of *verdichtetes Holz*.

5.2 Proposed Method

5.2.1 GPT-5

As GPT-5 sometimes includes part or all of its reasoning process in the output, it becomes apparent that its reasoning process is not consistent across sentences. In some instances, it adheres to the process given in the prompt (*[English term] means [French term] means [Spanish term] means [Italian term]*), whereas in others, it apparently provides the German translation directly in isolation without considering other languages. At times, it also treats other words in a sentence as domain terms,

MQM Error Category	Count
Terminology – Wrong Term: Minor	17
Terminology – Wrong Term: Major	11
Accuracy – Mistranslation: Minor	1
Accuracy – Mistranslation: Major	12
Accuracy – Mistranslation: Critical	7
Accuracy – Omission: Minor	1
Accuracy – Untranslated: Major	1
Total	50
Score	92.50

Table 10: MQM Errors Tower LLM Baseline Method

Model	BLEU	ChrF	COMET
GPT-5 Proposed Method	23.21	57.76	0.53
GPT-5 Baseline Method	22.34	56.40	0.53

Table 11: Translation Performance GPT-5 Proposed Method

in addition to the explicitly given term used in the CoD. Obviously, some form of reasoning is occurring; for *power of attorney*, for example, GPT-5 adds the explanation that according to the CoD, *Mandat* can be used—the terms in the chain being *mandat*, *mandato*, and *mandato*. In this case, however, the reasoning is faulty, since these are false friends, and the correct German term would be *Vollmacht*. This behavior demonstrates that inclusion of CoD can actually hurt the model's outcome.

Main Scores

GPT-5 attained a BLEU score of 23.21, reflecting a solid level of n-gram overlap with the reference translations and slightly stronger fluency and accuracy than other models. Its ChrF score of 57.76 was relatively high, indicating good capture of character-level patterns and consistent handling of word forms and morphology. The COMET score of 0.53 suggests good semantic alignment with the references, placing this model and the proposed method among the stronger performers, although there is still room for further improvement (see Table 11).

MQM Score

A visible pattern is apparent across the produced translations, which are often literal and, at times, more colloquial, while the reference translations are the established German legal or administrative terms (see Table 12). GPT-5 did not

employ the officially correct terms for, e.g., fishing permit (*Fischereigenehmigung* instead of *Fangerlaubnis*) or Act of Accession (*Beitrittsakt* instead of *Beitrittsakte*). Sometimes, there are even inaccuracies and stylistic problems in the translation (e.g., fall in prices - *Preisreduzierung* instead of *Preisrückgang,* suggesting a deliberate reduction of prices; fishery product - *fischereiliches Erzeugnis* instead of *Fischereierzeugnis;* grubbing premium - *Ausreißprämie* instead of *Rodungsprämie*, a term that does not exist in German). In addition, GPT-5 produces translations with omissions, e.g., experimental farm - *Versuchsanstalt* instead of *landwirtschaftliche Versuchsanstalt*, or model farm - *Betrieb* instead of *Musterbetrieb*.

MQM Error Category	**Count**
Terminology – Wrong Term: Minor	14
Terminology – Wrong Term: Major	13
Accuracy – Mistranslation: Major	2
Accuracy – Mistranslation: Critical	2
Accuracy – Omission: Minor	1
Accuracy – Untranslated: Major	1
Total	33
Score Proposed Method	95.87
Score Baseline Method	96.51

Table 12: MQM Errors GPT-5 Proposed Method

Comparison

The baseline method excelled in translation of memorised EU glossary items, which it sometimes translated in a paraphrastic rather than precise way. The proposed method shows strength with respect to general idiomaticity and when nuance matters (see Table 13).

5.2.2 Deepseek

Deepseek did not seem to fully understand the required task with the base prompt (input + translations of terms in different languages). It apparently reasoned over how to proceed with the translation process, but reached the wrong conclusions. In several instances, the single terms from the multilingual chain were translated separately instead of being used to translate the German term. Sometimes, the model failed to use them at all, stating that they were not German terms and were thus to be omitted. Often, it provided explanations about the content of the sentence to be translated. An extension of the prompt pointing out that

Original	Baseline method	Proposed method
farm return	Meldung der landwirtschaftlichen Betriebe	**Betriebsformular**
investment aid	Investitionshilfe	**Investitionsbeihilfe**
common land	Gemeinweide	**Gemeinland**
farm income	Einkommen der Landwirte	**Einkommen der landwirtschaftlichen Betriebe**
replanting	Pflanzung	**Wiederanpflanzung**
self-sufficiency rate	Selbstversorgungsquote	**Selbstversorgungsgrad**
representative rate	Repräsentativsatz	**repräsentativer Wechselkurs**
rural exodus	ländliche Abwanderung	**Landflucht**
shellfish farming	**Muschelzucht**	Schalentierzucht
veterinary legislation	**Veterinärrecht**	Veterinärgesetzgebung
general aid scheme	**allgemeine Beihilfenregelung**	allgemeines Beihilfensystem
livestock product	**Erzeugnis aus Nutztieren**	Viehprodukt
carcase weight	**Schlachtgewicht**	Karkassengewicht
world consumption	**weltweiter Verbrauch**	Weltverbrauch
area sown	**Anbaufläche**	eingesäte Fläche
commitment of expenditure	**Mittelbindung**	Ausgabenverpflichtung
EAGGF Guidance Section	**EAGFL, Abteilung Ausrichtung**	Orientierungsabschnitt des EAGGF
pre-accession assistance	**Heranführungshilfe**	Vorbeitrittshilfe
sensitive product	**empfindliche Ware**	sensibles Produkt
grubbing premium	**Rodungsprämie**	Ausreißprämie

Table 13: Comparison of GPT-5 Baseline vs. Proposed Method

the multilingual chain given in the prompt should be used led to the model to use this information inaccurately, by associating the multilingual terms with the wrong source terms. Rearranging the prompt—adding the multilingual chain after the instruction—seemed to improve the translation result. Overall, however, the translation remained quite faulty.

At times, Deepseek seemed to make an attempt at using the multilingual chain for translation, claiming that it had completed the translation by considering the other terms given. Yet, at other times, the model directly substituted a term from the multilingual chain as the German term (e.g., action for annulment - *recours en annulation* instead of *Anfechtungsklage*). Considering the chain could have even damaged the model's outcome; at times, the multilingual terms seemed to confuse the model and lead to misspellings of German terms (e.g., dehydration - *Deshydratation* instead of *Dehydrierung* or *Wasserentzug*, maybe based on the French term *deshydratation*; usufruct - *Usufrucht* instead of *Nießbrauch*, maybe based on the terms *usufruit*, *usufrutto*, and *usufructo*). This behavior could indicate that the model is not able to independently distinguish between different languages, or only to a limited extent.

Model	BLEU	ChrF	COMET
Deepseek V3 Proposed Method	12.09	45.83	0.19
Deepseek V3 Baseline Method	11.63	45.83	0.17

Table 14: Translation Performance Deepseek V3 Proposed Method

Main Scores

Deepseek V3 achieved a BLEU score of 12.09, indicating limited n-gram overlap with the reference translations and relatively weak fluency and accuracy. Its ChrF score of 45.83 suggests modest capture of character-level patterns and points to challenges with word forms and consistency. The COMET score of 0.19 was low, reflecting poor semantic alignment with the reference translations and overall lower translation quality (see Table 14).

MQM Score

The term translations produced by Deepseek were greatly lacking in accuracy (see Table 15). Some terms were simply translated in a wrong way, e.g., livestock unit - *Einheit von Haustier* instead of *Vieheinheit,* clearance of accounts - *Abwicklung von Konten* instead of *Rechnungsabschluss.* Sometimes, the model inserted a term from the multilingual chain into the translation, e.g., crop loss - *perde de récolte* instead of *Ernteausfall.* The model even invented terms, e.g., glasshouse cultivation - *Glasberggaertierung* instead of *Unterglasanbau* or crop rotation - *Pflanzungsrund* instead of *Wechselwirtschaft.* In addition, there were purely terminological errors, e.g., export refund - *Exportgutschrift* instead of *Erstattung bei der Ausfuhr*, with *Gutschrift* not being used in the context of EU law, representative rate - *repräsentativer Zinssatz* instead of *repräsentativer Kurs*, with *Zinssatz* not meaning the same as “rate” in this context.

Comparison

As in the case of GPT-5, the baseline method performed better with established EU terminology that required memorised glossary entries, sometimes using calques or descriptive paraphrases instead of official terms. The proposed method, on the other hand, produced idiomatic German in complex EU contexts, while sometimes translating in an over-literal way or leaving terms untranslated (see Table 16).

5.2.3 TowerLLM

Like Deepseek, TowerLLM apparently did not always employ the correct reasoning process, e.g., it extracted additional terms from the sentence on its own and then

MQM Error Category	Count
Terminology – Wrong Term: Minor	14
Terminology – Wrong Term: Major	11
Terminology – Wrong Term: Critical	1
Accuracy – Mistranslation: Minor	2
Accuracy – Mistranslation: Major	7
Accuracy – Mistranslation: Critical	15
Accuracy – Omission: Major	1
Accuracy – Omission: Critical	4
Accuracy – Untranslated: Major	8
Accuracy – Untranslated: Critical	1
Total	64
Score Proposed Method	87.43
Score Baseline Method	87.84

Table 15: MQM Errors Deepseek Proposed Method

Original	Baseline method	Proposed method
intra-Community trade	inner-Communityische Handelspraxis	**innergemeinschaftlicher Handel**
common land	gemeinschaftliches Ackerland	**gemeinschaftliches Land**
farm income	Landwirtschaft	**landwirtschaftliche Einkunft**
single market	einheitlicher Wirtschaftsraum	**einheitlicher Markt**
world consumption	Weltversorgung	**Weltverbrauch**
self-sufficiency rate	Selbstversorgungszahl	**Selbstversorgungsgrad**
table wine	Tischwein	**Tafelwein**
surface water	Surfacewasser	**Oberflächenwasser**
plant health legislation	**Pflanzenschutzrecht**	Gesetz zur Pflanzengesundheit
shellfish farming	**Muschelzucht**	Muschelfarmung
aquatic environment	**aquatische Umwelt**	aquatischer Raum
livestock unit	**Viehseinheit**	Haustiereinheit
crops under glass	**Gewächshaus-Kultur**	Kultur unter Glas
export refund	**Exportrückerstattung**	Exportgutschrift
aid per hectare	**Hektarbeihilfe**	Hilfe pro Hektar
traceability	**Rückverfolgbarkeit**	Trägerschaft
commitment of expenditure	**Ausgabenbindung**	commitment of expenditure
sensitive product	**empfindliches Produkt**	sensibles Produkt
feedingstuffs	**Futtermittel**	feedingstuffs

Table 16: Comparison of Deepseek Baseline vs. Proposed Method

placed the German version first in the reasoning process.

Model	BLEU	ChrF	COMET
TowerLLM Proposed Method	16.0	51.20	0.36
TowerLLM Baseline Method	18.62	51.70	0.38

Table 17: Translation Performance TowerLLM Proposed Method

Main Scores

Nevertheless, TowerLLM performed better than Deepseek in terms of BLEU score (16.0) and ChrF score (51.2). In particular, the COMET score indicates greater fluency (0.36) (17).

MQM Score

TowerLLM produced a significant number of mistranslations (see Table 18), e.g., hill farming - *Hügelbau* instead of *Landwirtschaft in Berggebieten*; farm return - *Betriebsprüfungsbericht* instead of *Betriebsbogen*, with the first meaning a tax return. It also produced terminological ambiguities, e.g., uncultivated land - *unbebautes Land* instead of *Ödland*, with the first meaning undeveloped land and the second meaning wasteland; clearance of accounts - *Abschlussrechnung* instead of *Rechnungsabschluss*, with the first meaning a final invoice and the second meaning a regulatory process to audit and verify accounts. Yet, it managed to accurately translate some difficult terms with which GPT-5 and Deepseek had difficulty e.g., bonded wood - *agglomeriertes Holz* and fishing capacity - *Fangkapazität.*

MQM Error Category	Count
Terminology – Wrong Term: Minor	13
Terminology – Wrong Term: Major	11
Accuracy – Mistranslation: Minor	2
Accuracy – Mistranslation: Major	16
Accuracy – Mistranslation: Critical	2
Accuracy – Untranslated: Major	2
Accuracy – Untranslated: Critical	1
Total	47
Score Proposed Method	93.03
Score Baseline Method	92.50

Table 18: MQM Errors Tower LLM Proposed Method

Comparison

The difference between the results from the baseline and proposed method for TowerLLM was not large. The baseline method was, again, stronger with respect to memorised glossary items but sometimes fell into calques, colloquialisms, or misleading simplifications. The proposed method excelled in technical accuracy and EU and legal standard terminology, but sometimes substituted with less precise synonyms (see Table 19).

Original	Baseline method	Proposed method
crops under glass	Glasgemüse	**Anbau unter Glas**
financial compensation	finanzielle Entschädigung	**finanzielle Ausgleichszahlung**
export refund	Exportrückerstattung	**Exporterstattung**
representative rate	Referenzsatz	**repräsentativer Wechselkurs**
bonded wood	Holzbinder	**agglomeriertes Holz**
animal feedingstuffs	Futter	**Futtermittel**
fuel crop	Brennholzpflanze	**Energiepflanze**
rural exodus	ländlicher Exodus	**Landflucht**
single-crop farming	Einfachkultur	**Monokultur**
water management	Wassermanagement	**Wasserbewirtschaftung**
aquatic environment	**aquatische Umwelt**	aquatische Umgebung
permanent crop	**dauernde Kultur**	dauerhafte Kultur
fall in prices	**Preisrückgang**	Preissenkung
meat product	**Fleischerzeugnis**	Fleischprodukt
aid per hectare	**Beihilfe pro Hektar**	Flächenergänzungsbeihilfe
surface water	**Oberflächenwasser**	Fließgewässer
fishing vessel	**Fischereischiff**	Fangboot

Table 19: Comparison of TowerLLM Baseline vs. Proposed Method

5.3 Comparison of the Different Models

Overall, clear performance differences could be observed across the evaluated models and methods. **GPT-5** consistently achieved the strongest results among all systems. In the baseline setting, it outperformed the other models across BLEU, ChrF, and COMET, indicating better fluency, lexical choice, and semantic adequacy. When the proposed method was applied, GPT-5 showed further improvements in BLEU and ChrF, confirming that it benefits the most of any of the models from the enhanced approach while maintaining strong semantic consistency, as reflected by its stable COMET score (see Table 20).

TowerLLM demonstrated mid-range performance. Its baseline scores were notably higher than those of Deepseek V3, particularly in BLEU and COMET,

Model	Method	BLEU	ChrF	COMET
GPT-5	Baseline	22.34	56.40	0.53
Deepseek V3	Baseline	11.63	45.83	0.17
TowerLLM	Baseline	18.62	51.70	0.38
GPT-5	Proposed	23.21	57.76	0.53
Deepseek V3	Proposed	12.09	45.83	0.19
TowerLLM	Proposed	16.00	51.20	0.36

Table 20: Overall Translation Performance for Baseline and Proposed Methods

suggesting better sentence-level quality and meaning preservation. However, TowerLLM still lagged behind GPT-5 by a clear margin, indicating limitations in both surface form accuracy and semantic alignment.

Deepseek V3 performed weakest overall. Both baseline and proposed configurations yielded relatively low BLEU and COMET scores, reflecting difficulties in producing fluent translations and preserving meaning. While the proposed method led to small improvements in BLEU and COMET, the gains were modest and did not close the gap with TowerLLM or GPT-5. This suggests that Deepseek V3—possibly due to the small size of the 7B version—is less responsive to the proposed enhancements and may require more substantial architectural or training changes to improve translation quality.

In summary, GPT-5 was the most robust and high-performing model across all metrics, with TowerLLM occupying a middle position and Deepseek V3 trailing behind, with only limited gains from the proposed method. Thus, the proposed method primarily improved surface-level translation quality (BLEU), offering modest benefits for ChrF and having minimal impact on semantic alignment as measured by COMET. The proposed method seemed primarily to improve wording and phrasing but did not necessarily produce translations that fully captured the intended meaning.

6 Discussion

The baseline and proposed methods each have strengths and weaknesses. Overall, the results suggest that MKG-derived prompts can marginally improve domain-specific MT, but their effectiveness is model- as well as method-dependent.

As the **baseline method** adheres, in general, to EU glossaries (e.g., *Fleischerzeugnis* vs. *Fleischprodukt*), its translations tend to be safe, memorised choices often incorporating codified glossary entries (e.g., *Muschelzucht* vs. *Schalentierzucht* or *Schalentierhaltung*). For highly **formalised** EU terminology, this tendency therefore more frequently produces accurate translations. On the other hand, the baseline method sometimes produces overly **descriptive** or **paraphrastic** translations (e.g., *Meldung der landwirtschaftlichen Betriebe* vs. *Betriebsformular*), often adopting a bureaucratic and wordy style when a more compact, idiomatic form exists and would be more appropriate. Occasionally, it also yields **false friends** or **anglicisms** (e.g., *Surfacewasser* instead of *Oberflächenwasser*).

In contrast, since the **proposed method** prefers compact, established words (e.g., *Landflucht* vs. *ländliche Abwanderung* or *ländlicher Exodus*), it produces more **idiomatic**, **natural-sounding** translations. It is also better than the baseline method at capturing nuance, handling prefixes and subtle distinctions well (e.g., *Wiederanpflanzung* vs. *Pflanzung* or *Umpflanzung*). In some domains, its translations are therefore more accurate (e.g., *Energiepflanze* vs. *Treibstoffpflanze* or *Brennholzpflanze).* In addition, it avoids calques and anglicisms. On the other hand, the proposed method also invents new terms where a fixed EU term exists (e.g., *Ausreißprämie* or *Absterbpauschale* vs. *Rodungsprämie*), and it occasionally leaves terms **untranslated** or **mistranslated** (e.g., *Commitment of expenditure* or *Verpflichtung von Ausgaben* vs. *Mittelbindung*).

Overall, the results show clear performance differences among the models. GPT-5 consistently achieves the strongest results across all metrics—particularly BLEU and ChrF—outperforming the other systems in both baseline and proposed settings and benefiting the most from the proposed method by incorporating it into its reasoning process. TowerLLM demonstrates moderate performance, surpassing Deepseek V3 in all baseline metrics but still trails GPT-5 by a noticeable margin, indicating room for improvement in fluency and semantic accuracy. With low BLEU and COMET scores, Deepseek V3 performs weakest overall and displays only modest gains from the proposed method, suggesting limited responsiveness to the enhancements and limited understanding of the prompt. In summary, GPT-5

emerges as the most robust model, with TowerLLM occupying a middle position and Deepseek V3 lagging behind in translation quality. As a quick further experiment has shown, both the baseline and the proposed method only outperform prompting the LLMs without any ICL or CoD examples by a small margin.

7 Conclusion

The goal of this research was to investigate whether MKGs can be used to improve domain-specific MT in LLMs via prompting and to compare the performance of a CoD-inspired method with that of a traditional few-shot baseline. The results suggest that KG-derived prompts can indeed improve translation quality, particularly for domain-specific terminology, but not significantly. Also, their effectiveness depends on several interacting factors. Importantly, the performance of the two prompting strategies appears to vary across subdomains. While the baseline few-shot method can benefit from sentence-level contextual information, the CoD approach relies primarily on term-level cues extracted from the KG. This difference may explain why the proposed method's performance is more robust for highly standardized terminology (e.g., EU-specific legal terms) but less consistent when broader sentential context is required. These findings align with prior work suggesting that LLM-based MT is particularly sensitive to domain coherence and terminological consistency (Agrawal et al., 2023; Sia and Duh, 2023).

A potential improvement to the proposed method concerns **prompt design**. As the current implementation largely presents multilingual term chains with limited explicit guidance, extending the prompt with a more extensive instruction—such as encouraging the model to use the terms comprising the multilingual chain as semantic pointers—may enable the LLM to better integrate the provided knowledge. Prior research on prompting indicates that even minimal task-specific instructions can improve performance when the number of ICL examples is limited, as is the case for CoD-style prompts (Peng et al., 2023b; Vilar et al., 2023).

Another factor that likely influences performance is **domain homogeneity**. The dataset used in this research draws from multiple EuroVoc subdomains, which may introduce conceptual and stylistic variation. Results indicate that both methods could benefit from restricting prompts and test sentences to a single, well-defined domain. Similarly, feeding the model larger and more coherent text chunks—rather than isolated sentences—could allow the LLM to better exploit discourse-level cues and align KG-derived terminology with its contextual usage.

A notable limitation of the current approach is that the KG is used only as a source of multilingual labels, while explicit **relationships between concepts** are ignored. Incorporating hierarchical (e.g., broader/narrower) or associative relations from EuroVoc could provide additional semantic context, such as a term's subdomain or conceptual proximity to related terms. This may enable a form of

lightweight multi-hop reasoning, where the model can infer meaning not only from direct translations but also from the conceptual network's structure. Given recent findings on KG-enhanced reasoning in LLMs, this represents a promising direction for future work (Pan et al., 2024; Ji et al., 2024).

Finally, the results highlight a broader trade-off between **contextual richness and scalability**. While KG-based prompting is comparatively easy to update and computationally inexpensive compared to retrieval-based NMT systems, its effectiveness depends heavily on careful prompt construction and domain alignment. This suggests that KG prompting is best viewed not as a replacement for traditional MT approaches, but as a complementary strategy—one particularly well-suited for terminology-sensitive, domain-specific translation scenarios.

7.1 Limitations

This study has several limitations that should be taken into account when interpreting the results. First, the evaluation was restricted to a relatively small and carefully selected set of sentences, focusing on specific types of EU-related terminology. While this permitted a targeted analysis of domain-specific translation behavior, it limited the generalizability of the findings to other domains or text types. Second, the KG was used solely as a source of multilingual labels, without exploiting its relational structure. As a result, potentially useful semantic information encoded in hierarchical or associative relations was not made available to the model.

7.2 Future Work

Future research could extend this work in several directions. A natural next step would be to incorporate explicit KG relations, i.e., broader, narrower, or related concepts, into the prompt, allowing the model to access a richer semantic context and potentially perform limited multi-hop reasoning. Another promising direction would be the development of hybrid prompting strategies that combine KG-derived term chains with a small number of sentence-level ICL examples, thereby balancing terminological precision with contextual fluency. Additionally, restricting prompts and test data to a single, well-defined subdomain, or automatically selecting KG concepts based on domain similarity, may further improve performance. Finally, future studies should include human evaluation by domain experts to better assess translation quality in terms of adequacy, consistency, and compliance with domain-specific conventions.

8 List of Acronyms

Acronym	Meaning
API	Application Programming Interface
BLEU	Bilingual Evaluation Understudy
BPE	Byte-Pair Encoding
CAT	Computer-Aided Translation
CoT	Chain of Thought
CoD	Chain of Dictionary
DIP	Dictionary Insertion Prompting
EL	Entity Linking
EU	European Union
ICL	In-Context Learning
IR	Information Retrieval
KG	Knowledge Graph
kNN	k-Nearest Neighbor
LLM	Large Language Model
LPG	Labeled Property Graph
MLA	Multi-Head Latent Attention
MMLU	Massive Multitask Language Understanding
MoE	Mixture of Experts
MKG	Multilingual Knowledge Graph
MQM	Multidimensional Quality Metrics
MT	Machine Translation
NLP	Natural Language Processing
NNM	Neural Network Model
NMT	Neural Machine Translation
POS	Part of Speech
PLM	Pretrained Language Model
QA	Question Answering
QE	Quality Estimation
RAG	Retrieval-Augmented Generation
RBMT	Rule-Based Machine Translation
RDF	Resource Description Framework
RL	Reinforcement Learning
RNN	Recurrent Neural Network
SKOS	Simple Knowledge Organization System
SMT	Statistical Machine Translation
SOTA	State of the Art
TF-IDF	Term Frequency–Inverse Document Frequency
SFT	Supervised Fine-Tuning
TM	Translation Memory
URI	Uniform Resource Identifier

Table 21: List of Acronyms

Abstract

Die vorliegende Masterarbeit untersucht, ob Multilinguale Wissensgraphen (MKGs) als Quelle für In-Context-Learning-Beispiele genutzt werden können, um die domänenspezifische Maschinelle Übersetzung (MT) mit Large Language Models (LLMs) zu verbessern. Anstatt ausschließlich auf satzparallele Daten zurückzugreifen, wird erforscht, inwiefern strukturierte mehrsprachige Terminologie aus einem Wissensgraphen zur Konstruktion von Prompts eingesetzt werden kann, die die crosslingualen Fähigkeiten und das Reasoning-Potenzial von LLMs gezielter aktivieren.

Als Datenbasis dient EuroVoc, der mehrsprachige Thesaurus der Europäischen Union, der als Wissensgraph eingebunden wird. Evaluiert werden drei LLMs – GPT-5, DeepSeek-V3 und TowerLLM – anhand von EU-spezifischer Rechtsterminologie aus dem EUR-Lex-Korpus. Es werden zwei Methoden miteinander verglichen: eine Few-Shot-Baseline, bei der einschlägige Satzpaare als In-Context-Beispiele verwendet werden, sowie eine auf dem Chain-of-Dictionary-Ansatz (CoD) basierende Methode, bei der mehrsprachige Terminologieketten aus dem Wissensgraphen in den Prompt eingebettet werden.

Die Ergebnisse zeigen, dass KG-basierte Prompts die Übersetzungsqualität für domänenspezifische Terminologie partiell verbessern können, wobei die Wirksamkeit stark modellabhängig ist. GPT-5 erzielt durchgehend die besten Ergebnisse und profitiert am stärksten von der vorgeschlagenen Methode. TowerLLM belegt eine mittlere Position, während DeepSeek-V3 die schwächsten Resultate aufweist und kaum von der CoD-Methode profitiert. Die Baseline-Methode erweist sich als stärker bei fest etablierten EU-Glossareinträgen, während die CoD-Methode bei idiomatischer Natürlichkeit und terminologischer Nuanciertheit Vorteile zeigt. Die Arbeit schließt mit einer Diskussion der Limitationen und Ansätzen für zukünftige Forschung, darunter die Einbeziehung relationaler KG-Strukturen und hybride Prompting-Strategien.

Bibliography

Rishabh Adiga, Lakshmi Subramanian, and Varun Chandrasekaran (2024). Designing informative metrics for few-shot example selection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10127–10135, Bangkok, Thailand. Association for Computational Linguistics.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad (2023). In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Duarte M. Alves, José P. Pombal, Nuno M. Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André Martins (2024). Tower: An open multilingual large language model for translation-related tasks. *arXiv*, abs/2402.17733.

Philip Arthur, Graham Neubig, and Satoshi Nakamura (2016). Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio (2014). Neural machine translation by jointly learning to align and translate. volume abs/1409.0473.

Ryan Brate, Minh-Hoang Dang, Fabian Hoppe, Yuan He, Albert Meroño-Peñuela, and Vijay Sadashivaiah (2022). Improving language model predictions via prompts enriched with knowledge graphs. In *Deep Learning for Knowledge Graphs Workshop (DL4KG) at ISWC 2022*, Hangzhou, China. Hal-03991267.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). Language models are few-shot learners. In *Advances*

in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Huiyao Chen, Yu Zhao, Zulong Chen, Mengjia Wang, Liangyue Li, Meishan Zhang, and Min Zhang (2024). Retrieval-style in-context learning for few-shot hierarchical text classification. *Transactions of the Association for Computational Linguistics*, 12:1214–1231.

Pranjal Chitale, Jay Gala, and Raj Dabre (2024). An empirical study of in-context learning in LLMs for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7384–7406, Bangkok, Thailand. Association for Computational Linguistics.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, and Yunyao Li (2024a). Enhancing machine translation experiences with multilingual knowledge graphs. In *AAAI Conference on Artificial Intelligence*.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li (2024b). Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, et al. (2024). DeepSeek-V3 technical report. *arXiv*, abs/2412.19437.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer (2023). Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv*, abs/2302.07856.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang (2023). Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Amr Hendy, Mohamed Gomaa Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla (2023). How good are GPT models at machine translation? a comprehensive evaluation. *arXiv*, abs/2302.09210.

Nourhan Ibrahim, Samar AboulEla, Ahmed F. Ibrahim, and Rasha Kashef (2024). A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): Models, evaluation metrics, benchmarks, and challenges. *Discover Artificial Intelligence*, 4:76.

Yixin Ji, Kaixin Wu, Juntao Li, Wei Chen, Mingjie Zhong, Xu Jia, and Min Zhang (2024). Retrieval and reasoning on KGs: Integrate knowledge graphs into large language models for complex question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7598–7610, Miami, Florida, USA. Association for Computational Linguistics.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu (2023). ParroT: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020, Singapore. Association for Computational Linguistics.

Lucie-Aimée Kaffee, Russa Biswas, Catharina Maria Keet, Edlira Vakaj, and Gerard de Melo (2023). Multilingual knowledge graphs and low-resource languages: A review. *Transactions on Graph Data and Knowledge (TGDK)*, 1:10:1–10:19.

Anne Lise Kjær (2007). Legal translation in the European Union: A research field in need of a theoretical framework. In K. Kredens and S. Goźdź-Roszkowski, editors, *Language and the Law: International Outlooks*, pages 69–95. Peter Lang, Frankfurt am Main.

Philipp Koehn (2010). Statistical machine translation. In *Machine Translation Summit*.

Philipp Koehn (2017). Neural machine translation.

Philipp Koehn (2020). *Neural Machine Translation*. Cambridge University Press.

Aswanth Kumar, Ratish Puduppully, Raj Dabre, and Anoop Kunchukuttan (2023). CTQScorer: Combining multiple features for in-context example selection for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7736–7752, Singapore. Association for Computational Linguistics.

Pranjal Kumar (2024). Large language models (LLMs): Survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57:260.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing (2024). Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth

Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li (2022). Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit (2013). Multidimensional quality metrics: A flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Hongyuan Lu, Zixuan Li, and Wai Lam (2024a). Dictionary insertion prompting for multilingual reasoning on multilingual large language models. *arXiv*, abs/2411.01141.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei (2024b). Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp (2021). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Annual Meeting of the Association for Computational Linguistics*.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang (2024). A paradigm shift: The future of machine translation lies with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova (2024). Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Laura Mori, editor (2018). *Eurolectology: EU Varieties of the National Languages*, volume 146 of *Benjamins Translation Library*. John Benjamins Publishing Company, Amsterdam.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way (2023). Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Diego Moussallem, Axel-Cyrille Ngonga Ngomo, Paul Buitelaar, and Mihael Arcan (2019). Utilizing knowledge graphs for neural machine translation augmentation. *Proceedings of the 10th International Conference on Knowledge Capture*.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36:3580–3599.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne (2023a). Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, pages 1–32.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao (2023b). Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Gianluca Pontrandolfo (2019). Eurolects and legal translation: An empirical approach. In Fernando Prieto Ramos, editor, *Legal Translation and Court Interpreting: Ethical Values, Quality, Competence Training*, pages 65–89. Peter Lang, Bern.

Maja Popović (2015). chrF: Character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*.

Javad Pourmostafa Roshan Sharami, Dimitar Shterionov, and Pieter Spronck (2024). Guiding in-context learning of LLMs through quality estimation for machine translation. In *Conference of the Association for Machine Translation in the Americas*.

Publications Office of the European Union (2015). *EuroVoc Thesaurus. Volume 1, Alphabetical Version. Part A, A…I*. Publications Office of the European Union, Luxembourg.

Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw, and Nancy Chen (2023). DecoMT: Decomposed prompting for machine translation between related languages using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4586–4602, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Susan Šarčević (1997). *New Approach to Legal Translation*. Kluwer Law International, The Hague.

Rico Sennrich, Barry Haddow, and Alexandra Birch (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Suzanna Sia and Kevin Duh (2023). In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 173–185, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Jonathan Slocum (1985). A survey of machine translation: Its history, current status and future prospects. *Computational Linguistics*, 11(1):1–17.

Weiting Tan, Haoran Xu, Lingfeng Shen, Shuyue Stella Li, Kenton Murray, Philipp Koehn, Benjamin Van Durme, and Yunmo Chen (2024). Narrowing the gap between zero- and few-shot machine translation by matching styles. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 490–502, Mexico City, Mexico. Association for Computational Linguistics.

Chenming Tang, Zhixiang Wang, and Yunfang Wu (2024). SCOI: Syntax-augmented coverage-based in-context example selection for machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9956–9971, Miami, Florida, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster (2023). Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu (2023). Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le (2021). Finetuned language models are zero-shot learners. *arXiv*, abs/2109.01652.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden (2025). In-context example selection via similarity search improves low-resource machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch (2023a). Prompting large language model for machine translation: A case study. *arXiv*, abs/2301.07069.

Min Zhang, Limin Liu, Yanqing Zhao, Xiaosong Qiao, Su Chang, Xiaofeng Zhao, Junhao Zhu, Ming Zhu, Song Peng, Yinglu Li, Yilun Liu, Wenbing Ma, Mengyao Piao, Shimin Tao, Hao Yang, and Yanfei Jiang (2023b). Leveraging multilingual knowledge graph to boost domain-specific entity translation of ChatGPT. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 77–87, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Yang Zhao, Lu Xiang, Junnan Zhu, Jiajun Zhang, Yu Zhou, and Chengqing Zong (2020a). Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4495–4505, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yang Zhao, Jiajun Zhang, Yu Zhou, and Chengqing Zong (2020b). Knowledge graphs enhanced neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, pages 4039–4045. International Joint Conferences on Artificial Intelligence Organization.

Xin Zheng, Zhirui Zhang, Shujian Huang, Boxing Chen, Jun Xie, Weihua Luo, and Jiajun Chen (2021). Non-parametric unsupervised domain adaptation for

neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4234–4241, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shaolin Zhu, Menglong Cui, and Deyi Xiong (2024a). Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629, Torino, Italia. ELRA and ICCL.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li (2024b). Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Declaration of Authorship:

I declare that this Master Thesis has been written by myself. I have not used any other than the listed sources, nor have I received any unauthorized help.

I hereby certify that I have not submitted this Master Thesis in any form (to a reviewer for assessment) either in Austria or abroad.

Furthermore, I assure that the (printed and electronic) copies I have submitted are identical.

Date: 13.05.2026 Signature:

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of University of Vienna's products or services.