Journal of Digital Terminology and Lexicography - ISSN:3103-3601
Vol. 1, No. 2 (2025) DOI: 10.25430/pupj.jdtl.1763379938

An Evaluation of Terminology-Augmented Generation (TAG) and Various Terminology Formats for the Translation Use Case

Anna Lackner[1†], Alena Vega-Wilson[1†], Christian Lang[1*†]

[1]Kaleidoscope GmbH, Landstraße 99-101, Vienna, 1030, Austria.

*Corresponding author(s). E-mail(s): christian.lang@kaleidoscope.at
Contributing authors: anna.lackner@kaleidoscope.at; alena.vega-wilson@eurocom.at
†These authors contributed equally to this work.

Abstract

In this study, we demonstrate the effectiveness of Terminology-Augmented Generation (TAG) for Large Language Model (LLM)-based Machine Translation (MT) and analyze the impact of terminology formats for that use case. By conducting empirical evaluations using OpenAI's GPT-4o, GPT-4o-mini, and the open-weights Gemma3:12b and Mistral 7B, Mistral Nemo, and Mistral Large (2411) models, we systematically explore various established terminology formats (including TBXv3) and compare the results to alternative structured formats and their impact on generation quality. Our findings, on both a preexisting test dataset and a dataset created from real-world customer documents, show that TAG with capable LLMs delivers results on-par or better than a fine-tuned NMT baseline, and that specific formatting strategies can improve model accuracy and recall of in-context knowledge, albeit not to the scale we originally expected. Our findings inform the design of terminology integration strategies for LLM-based MT, improving term adherence, domain adequacy, and translation consistency in specialized communication.

Keywords: NLP, LLM, Retrieval-Augmented Generation, Terminology-Augmented Generation, Machine Translation

1 Introduction

With the advent of Large Language Models (LLMs) such as the GPT-series by OpenAI, in-context learning has emerged as a new method for instilling knowledge into Artificial Intelligence (AI) systems without the need for fine-tuning or retraining, allowing for greater accessibility and real-time application of up-to-date and proprietary knowledge. One of the most common approaches of providing in-context knowledge to LLMs is the concept of Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). However, general RAG frameworks are designed for unstructured knowledge retrieval in documents rather than structured terminology information, typically employing a form of dense vector-based semantic search (Gupta, Ranjan, & Singh, 2024; Lewis et al., 2020). In first explorations of such RAG frameworks, we found that they are not well suited for terminology retrieval for a variety of reasons, some of which are:

1. The retrieval process is comparatively slow;
2. Retrieval is generally quite fuzzy, leading to noisy data;
3. Arbitrary chunking of data may lead to critical information loss;
4. Retrieval methods are often limited to top-k hits, potentially leading to silence in the retrieved data;
5. Typical terminology formats (XML) are not well suited for vector-based semantic search.

We aim to explore two major components for efficient Terminology-Augmented Generation (TAG). Firstly, we explore the retrieval for TAG as a specialized extension of RAG, using readily available terminology APIs in Kalcium Quickterm. We explore the impact of TAG in terms of speed, reliability, and general feasibility for Machine Translation (MT), as a downstream task with various LLMs. Our second focus will be on the question of how to format the retrieved terminological entries when providing them as in-context knowledge to LLMs. Although LLMs demonstrate remarkable abilities to parse XML – the typical terminology exchange format that is also standardized in the TBX XML specification (ISO 30042, 2019) – we examine if the verbose nature of the XML structure is detrimental for providing terminology to LLMs, and if it is, find viable alternatives for providing structured terminological knowledge to LLMs at run-time.

This work builds on our previous paper (Lackner, Vega-Wilson, & Lang, 2025), published as part of the Multilingual Digital Terminology Today 2025 conference, in which we introduced first results of the performance of TAG for translation.

2 Related Work

RAG represents a significant architectural advancement in LLM systems, merging neural generation capabilities with dynamic access to external knowledge repositories. Initially developed to address static model knowledge constraints, RAG enables systems to anchor outputs in contextually retrieved textual passages. However, this approach exhibits substantial limitations: retrieved content frequently proves noisy or contextually irrelevant, lacking optimization for terminological precision and multilingual consistency. Such deficiencies become particularly acute in specialized domains — law, medicine, scientific research — where terminological accuracy and traceability constitute critical requirements.

Additionally, generic RAG solutions, mostly based on vector similarity, do not offer methods to filter within the data or retrieve and use the exact terms from a termbase. Vector-based RAG also slows down generation due to the lengthy retrieval process, and terminology is not available in real-time, because these vector datasets need to be created and stored separately from the termbase. The actual retrieval process is difficult to control, and there is a risk of losing essential information or the connection between permitted and prohibited designations, because of the arbitrary "chunking" of data employed in typical RAG frameworks (Fleischmann, 2025).

That said, TAG can be seen as a specialized form of RAG (Lewis et al., 2020), as both retrieve information from an external data source and provide it to an LLM as in-context knowledge.

In contradistinction to RAG's reliance on large-scale unstructured text retrieval via vector-space similarity, the TAG paradigm leverages structured, expert-curated knowledge resources. Systems using TAG interface directly with multilingual termbases, domain-specific ontologies, and controlled glossaries, accessing formally defined concepts, interlingual equivalents, and controlled relationships through deterministic API queries rather than similarity-based document fragments.

The architectural distinctions are fundamental. Where RAG retrieves loosely relevant paragraphs yielding opaque sources, TAG employs precise data models of termbases offering transparent, machine-readable formats with filtered, curated entries. This structural divergence enables terminologically faithful generation, that remains explainable and verifiable. TAG's operational framework incorporates several essential components: a terminology access layer supporting structured queries; filtering and reasoning modules aligning retrieved data with input contexts; generation modules conditioning outputs through prompt engineering or adapter layers; and, crucially, human-in-the-loop workflows enabling expert validation during content generation.

Compared to classic RAG, TAG is efficient in processing speed and token consumption. Real-time access to terminology enables LLMs to access current and specific information from termbases. TAG also aims to be simple to implement and easily understandable to users (Fleischmann, 2025). Classical terminological methods provide TAG with distinct advantages: deterministic access through exact search and controlled retrieval mechanisms; flexible data models compatible with LLM requirements as they are centered on linguistic data; and real-time API access circumventing slower embedded content retrieval. Rather than mimicking RAG's architecture, TAG should be conceptualized as a complementary paradigm capitalizing on terminological infrastructure strengths, positioning it to support high-precision, domain-sensitive multilingual generation, where conceptual clarity and expert-defined usage patterns remain paramount.

Table 1 Comparison of prior methods by Dinu et al. (2019) (baseline) to two non-augmented LLMs (GPT-4o-Mini, GPT-4o) and a terminology augmented LLM (GPT-4o) for en-de segment-based translation on the IATE annotated WMT 17 test set. ↑ and ↓ represent significantly better and worse scores than the baseline system with p-value < 0.05.

Model	Term %	BLEU	COMET	Time(s)
Baseline	94.5	26.0	-	0.20
GPT-4o	87.2 (-7.3) ↓	**35.7 (+9.7)** ↑	**0.880**	2.33
GPT-4o-Mini	87.4 (-7.1) ↓	33.7 (+7.7) ↑	0.876	1.95
GPT-4o with TAG	**96.37 (+1.87)** ↑	35.5 (+9.5) ↑	0.877	2.04

TAG supports terminological applications across multiple downstream tasks, and can be "particularly beneficial for content generation tasks that require adherence to domain-specific or corporate language norms, such as specialized translation or technical communication, where consistent use of (validated) terminology is essential" (Heinisch, 2025). Di Nunzio (2025) describes TAG as a new paradigm for integrating curated terminology resources into generative AI workflows. Therefore, we believe that TAG also increases the value of terminology in both commercial and non-commercial organizations, as it can be used "much more versatilely in tasks such as machine translation, terminology checking, or even creation of new content" (Lang & den Nest, 2025).

In our previous work (see Lackner et al., 2025) we evaluated the performance of TAG for the MT task on the WMT 2017 dataset. We followed the experimental setup of Dinu, Mathur, Federico, and Al-Onaizan (2019) for the WMT 2017 English-German news translation task,[1] which they also used for their evaluation of MT models fine-tuned on using terminology, that is provided at run time (Dinu et al., 2019). We used their best performing NMT model as a baseline for comparison to LLMs with TAG.

We evaluated the translation results in line with Dinu et al. (2019), by using BLEU (Bilingual Evaluation Understudy) (Papineni, Roukos, Ward, & Zhu, 2002) and calculating the average generation time for the translation. Additionally we also used COMET (Crosslingual Optimized Metric for Evaluation of Translation) (Rei, Stewart, Farinha, & Lavie, 2020). An overview of the results from the WMT17 dataset can be found in Table 1, and for a detailed analysis see Lackner et al. (2025).

The results of this first TAG evaluation showed that an LLM with TAG outperformed the NMT baseline. However, during the evaluation we encountered several issues of terminological nature in the dataset (Lackner et al., 2025). Among other things, the terminology extracted from IATE contained primarily common nouns (e.g. *month*, *eggs* or *tobacco* and proper nouns like country names (e.g. *Syria*), which would generally not be considered as terminology, and would often be correctly translated by both NMT models and LLMs trained on common domain texts. Moreover, the terminology in this provided glossary was partially not in the infinite singular form (e.g. *eggs*, *victories*), contained verbs (e.g. *covering*, *to arrest*), and sometimes articles (e.g. *Die Republikaner*). Furthermore, since the terminology was only provided as a glossary, no additional terminological information besides the term itself was available, leading to issues with the disambiguation of terms (Lackner et al., 2025).

These underlying terminological issues highlight the need for an evaluation of TAG for translation on a dataset containing real-world terminology.[2] In this work, we therefore evaluate TAG for translation on a custom dataset based on a customer's data and terminology, to evaluate the full potential of TAG that goes beyond simple glossaries, and provides LLMs with relevant terminological information.

3 Methodology

For the de-en language pair six different popular language models were used for the translation with TAG: OpenAI's closed source GPT-4o and GPT-4o-Mini (Model snapshots GPT-4o 2024-11-20 and GPT-4o Mini 2024-07-18), as well as the instruction-tuned open-weights Gemma3:12b (4bit int quantized),[3] Mistral 7B,[4]

[1] https://www.statmt.org/wmt17/translation-task.html
[2] Terminology resources that are used in production at companies, and not purpose-made for the evaluation of TAG. For example, definitions might not always be present and metadata may only be present if it was deemed relevant for a specific use case.
[3] https://ollama.com/library/gemma3:12b
[4] https://mistral.ai/news/announcing-mistral-7b

Mistral NeMo[5] and Mistral Large 2[6] models. For the de-cs and de-it language pairs, we evaluated only the OpenAI models and Mistral 7B. This restriction was due to the considerable time required for running the open-weights models on local hardware, manual output review, and the limited research hours available within our project budget. Nevertheless, the chosen models provide representative coverage of current high and low-performing systems, and give a good general overview of TAG effectiveness in the evaluated language pairs.

We decided to use the GPT-4o series models, as they represented the state-of-the-art models at OpenAI.[7] Further, we decided to use the Mistral models, as popular European alternative to US AI models, and added the Mistral NeMo and Mistral Large 2 model specifically for their touted multilingual training, as Mistral 7B proved to perform very poorly in the translation task. Finally, Gemma3:12b was used as a popular and readily available model, that ran at an acceptable speed on our local hardware, and we had tested previously with good results.

3.1 LLM Setup

For all models we set three major parameters controlling the variance of the generated output to a fixed value: temperature is set to 0.2, top-p is kept at the default value of 1, and the seed is set to 42. While this still allows for some variance, we found that setting the temperature lower tends to reduce the perceived and measured quality of the output.

Since most of our previous testing was done using OpenAI models, our prompting techniques are possibly favoring OpenAI trained models. This evaluation is therefore not to be interpreted as a comparison between different models, but rather as an exploration of various effects of different prompting formats for TAG with different LLM backends.

We set up a system prompt that is shared across all models. For the translation with non-augmented models, we translate both datasets with the simple system prompt:

"Translate the text provided by the user from and into the language specified by the user. Only return the translation."

Followed by the user prompt:

"Translate from {source language} to {target language}: {text}"

For the models with TAG we use a more complex system prompt, which can be found in Appendix A.1.

For terminology augmentation we explore a variety of possible formats, in which the terminological information is provided to the LLMs. The terminology formats range from TBX DCA and TBX DCT (version 3) (LTAC Global, 2021), which are native outputs of terminology systems to other structured outputs: JSON (see Listing 1), Markdown (see Listing 2), and YAML (see Listing 3). The Markdown format is used as the default terminology format unless stated otherwise.

To access the models, we employ the open-source AI interface Open WebUI[8] (Baek, Hussain, Liu, Vincent, & Kim, 2025) in combination with open-source ollama framework,[9] which allows us to access both open-source and proprietary models via one single OpenAI-conformant REST-API interface. Open WebUI also allowed us to set up custom pipelines and filters, which enabled us to run TAG with every API-request sent to TAG-enabled models, while also allowing us to inspect results and demo the effect in a user-friendly web ui. We implemented TAG in Python to make the testing easily reproducible with other OpenAI compatible endpoints. The TAG code itself relies on the Kalcium REST-API,[10] which provides ready-made endpoints for advanced term recognition, using various established term recognition methods such as *fuzzy matching* and *stemming*. We use the `/kalcrest/retrieval/content-of-entries-by-langid` endpoint for all of our tests, using analysis profiles with a similarity rate set to 70%, stemming enabled and parse the returned JSON as required for all our tests. We make the code and test sets[11] available in our Github repository.[12]

[5] https://mistral.ai/news/mistral-nemo
[6] https://mistral.ai/news/mistral-large-2407 and https://docs.mistral.ai/models/mistral-large-2-1-24-11
[7] https://openai.com/index/hello-gpt-4o/
[8] https://github.com/open-webui/open-webui
[9] https://github.com/ollama/ollama
[10] https://demo.kaleidoscope.at/kalcrest/swagger/index.html
[11] The release of the customer dataset and termbase is still pending approval.
[12] https://github.com/clang88/tag-mt-formats-eval

Listing 1 Example of JSON as terminology format

```
[
  {
    "definition": "Das massive, die eigentliche Gestalt ausmachende Teil eines Möbels.
    Der Korpus ...",
    "en-gb_term_1": "cabinet",
    "en-gb_term_1_usage": "Approved",
    "de-at_term_1": "Korpus",
    "de-at_term_1_usage": "Approved",
    "de-at_term_1_usage note": "Plural: Korpusse, NICHT Korpen!",
    "de-at_term_2": "Korpusse",
    "de-at_term_2_usage": "Approved",
    "de-at_term_3": "Korpen",
    "de-at_term_3_usage": "Forbidden"
  }
]
```

Listing 2 Example of Markdown as terminology format

```
## Concept 322
* Definition: Das massive, die eigentliche Gestalt ausmachende Teil eines Möbels. Der
    Korpus ...
### cabinet
#### Possible translations:
1. Korpus
  Usage: Approved
  Usage: Plural: Korpusse, NICHT Korpen!
2. Korpusse
  Usage: Approved
```

Listing 3 Example of YAML as terminology format

```
concept 322: \\
  - Definition: Das massive, die eigentliche Gestalt ausmachende Teil eines Möbels. Der
     Korpus ...
  - source_term: cabinet
    - target_term 1: Korpus
      - Usage: Approved
      - Usage note: Plural: Korpusse, NICHT Korpen!
    - target_term 2: Korpusse
      - Usage: Approved
```

3.2 Dataset and Evaluation Setup

We use a small custom dataset based on a customer's texts and terminology with three language directions: German-English, German-Czech, and German-Italian. While the source language is of the Austrian German variety (de-AT), for simplicity, we simply refer to it as "German" or "de". Likewise, while our test set includes a small amount of US English texts (en-US), the majority of the dataset targets the British variety (en-GB), with only approx. 17.5% (3,500 words vs. 20,000+ words) of the test set translated into en-US. As the current analysis does not aim to investigate the effects of different language varieties in this evaluation, all language pairs are referred to at the level of individual languages (ISO 639, 2023), and, thus, identified by their two-letter language identifier "de" (German), "en"(English), "it" (Italian) and "cs" (Czech), rather than by language variety-specific identifiers such as "de-AT" or "en-US".

We made an effort to segment-align any equivalent segments in our custom dataset to create a segment-based test set. To also evaluate the long-context performance of translation with TAG, we additionally split the custom dataset into paragraphs of a maximum of 1,000 characters, which we used for the paragraph-based translation (see Section 4.4).

Our experimental setup is in line with the setup of Dinu et al. (2019) and our previous work (Lackner et al., 2025). The translation results were evaluated using BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) and COMET (Crosslingual Optimized Metric for Evaluation of Translation) (Rei et al., 2020). BLEU scores were computed using SacreBLEU to obtain comparable and reproducible results (Post, 2018). We performed paired bootstrap with the default parameters and a p-value < 0.05 for BLEU and COMET to evaluate the statistical significance of the results (Koehn, 2004). Additionally, we also computed the average time of the models to generate the translation of one segment or paragraph. However, LLM latency is not the primary focus of our research, as various factors including model size, input length, API calls for models and TAG, and hardware specifications all affect the average translation time. Nonetheless, we included the average translation time for each model to align with the research by Dinu et al. (2019) and for the sake of completeness, but we will not explore latency in detail and we leave this open for future work. Our priority is to demonstrate whether the TAG process significantly impacts the average LLM translation time.

We created a termbase using existing terminology resources curated by the customer and our team. We extracted the terminology from the human reference translation that was present in this termbase using fuzzy search, and then manually validated it to use it as the reference terminology. Therefore, the terminology adherence score is based on the terminology present in the human reference translation, and not on the terminology present in the source text.

To detect whether the correct terminology was used in the LLM translation, we apply a fuzzy matching strategy similar to Exel, Buschbeck, Brandt, and Doneva (2020). Specifically, as mentioned above, we stem words using the stemming engine in Kalcium Quickterm, and perform a fuzzy search with a similarity rate of 70%. We acknowledge that this approach does not yield perfect results, and can lead to false positives or negatives, e.g. for discontinuous terms or morphological variants that fail to be stemmed correctly. For this reason we also manually sample the results to detect any irregularities stemming from erroneous term recognition. To evaluate this, we calculate the percentage of correctly translated terms, which we refer to as terminology adherence.

4 Results

We used the default terminology format Markdown (see Section 3.1) and evaluated on a segment level for the comparison of non-augmented and terminology-augmented LLMs unless stated otherwise. The termbase contained additional terminological information besides the term itself, and, if available, the corresponding definition, usage status (e.g. *approved*, *forbidden*) and the usage note (usage recommendation) of the retrieved concept were passed along with the term to the LLM as terminological context. The terminology adherence and the other scores are computed as mentioned in Section 3.2.

4.1 Results German-English Translation

For the de-en translation, six different models with TAG, namely Gemma3:12B, GPT-4o-Mini, GPT-4o, Mistral 7B, Mistral NeMo, and Mistral Large 2, were used. GPT-4o-Mini without TAG was used as a baseline, but we also provide results for GPT-4o without TAG for comparison.

The results indicate that the terminology-augmented LLMs with TAG outperform the non-augmented LLMs in terms of the correct translation of terminology, as a significant increase in the terminology adherence score is visible for all models with TAG (see Table 2). While both non-augmented LLMs (GPT-4o-Mini and GPT-4o) were outperformed by the models with TAG, the bigger model GPT-4o outperformed the smaller GPT-4o-Mini model across all scores, including the terminology adherence. However, this increase in the terminology adherence is not as stark compared to the models with TAG, where also the smaller GPT-4o-Mini model with TAG outperforms the non-augmented bigger GPT-4o across all scores, and in particular the term adherence by far (+20.16).

The SacreBLEU and COMET scores improve only slightly for Gemma3, GPT-4o-Mini, and GPT-4o with TAG, however the paired bootstrap test indicates that the difference between the SacreBLEU and COMET

Table 2 Comparison of two non-augmented LLMs (GPT-4o-Mini, GPT-4o), and different terminology augmented LLMs (Gemma3:12B, GPT-4o-Mini, GPT-4o, Mistral NeMo, Mistral Large, Mistral 7B) for de-en segment-based translation on the BLUM custom dataset. ↑ and ↓ represent significantly better and worse scores than the baseline system with p-value < 0.05.

Model	Term %	SacreBLEU	COMET	Time(s)
GPT-4o-Mini (without TAG)	67.48	32.09	0.8514	1.88
GPT-4o (without TAG)	70.02 (+2.54) ↑	33.24 (+1.15)	0.8574 (+0.0060) ↑	2.23
GPT-4o-Mini with TAG	90.18 (+22.27) ↑	33.50 (+1.41) ↑	**0.8597 (+0.0083)** ↑	2.38
GPT-4o with TAG	**90.70 (+23.22)** ↑	33.44 (+1.35) ↑	0.8596 (+0.0082) ↑	2.65
Gemma3:12B with TAG	87.72 (+20.24) ↑	**36.64 (+4.55)** ↑	0.8574 (+0.0060) ↑	7.16
Mistral 7B with TAG	-	-	-	-
Mistral NeMo with TAG	86.36 (+18.88) ↑	29.17 (-2.92) ↓	0.8212 (-0.0302) ↓	**1.60**
Mistral Large 2 with TAG	87.72 (+20.24) ↑	31.74 (-0.35)	0.8406 (-0.0108) ↓	1.82

scores for these models with TAG, and the non-augmented baseline LLM, are statistically significant. For Mistral NeMo and Mistral Large 2 with TAG, the results for SacreBLEU and COMET were either significantly worse, or slightly but not significantly worse than the baseline. It is important to note that the output of Mistral 7B was not usable, and therefore could not be evaluated. A possible explanation for this is that we have to assume that Mistral 7B was trained on English data only, since neither the official product description (Mistral AI Team, 2023) of the model nor its introduction paper (Jiang et al., 2023) explicitly state that the model was pretrained on multilingual data.

The big increase in the terminology adherence score, but only a marginal increase or even decrease in SacreBLEU and COMET scores for all models with TAG is striking. As aforementioned, we performed a paired bootstrap test, which indicated that almost all results are significant (except for the SacreBLEU score of Mistral Large 2) compared to the baseline. A possible explanation for the marginal increase in SacreBLEU and COMET scores for some of the models with TAG is that not all of the sentences in the dataset contain terminology, and the terminology only makes up a small part of the dataset. This explanation is strengthened by the observations made by Dinu et al. (2019), who state that little variation concerning the BLEU scores can be expected in test sets focusing on terminology, as the terminology only makes up a small part of the dataset. This is true for both of our test sets, and explains the marginal difference in SacreBLEU and COMET scores.

The analysis of the terminology errors made by the augmented LLMs (see Table 3 for some of the most common errors) showed that the majority occurred due to missing or incorrect terminology in the source text, that was (correctly) used in the human reference translation. These errors stem from our evaluation method, and the inconsistent use of terminology in the human texts. As already mentioned in Section 3.2, the terminology used as a reference to calculate the terminology adherence is based on the human reference translation. Therefore, if terms were not or incorrectly used in the source text, but correctly in the human reference translation, it is counted as an error if the term is also not present in the LLM translated text.

Other errors can be traced back to inconsistencies or missing information in the termbase. Many entries in the termbase have multiple terms with an admitted usage status, but lack additional information like a usage note, which is necessary for disambiguating when to use which admitted term. This led to multiple errors in the correct translation of terms, since the augmented LLMs also used correct admitted terms, that were not in the human reference translation, and were therefore counted as an incorrect translation of the term. Another type of error common to all LLMs occurred when having to translate compound words, where only part of it is in the termbase, for example the term *baguette pull-out*, which is the compound word *Baguetteauszug* in German, where only *pull-out* (*Auszug*) is in the termbase. This is most likely a limitation of the retrieval method, and could potentially be mitigated by adjusting the search parameters, so that also compound words, where only part of the word is in the termbase, are found and retrieved.

4.2 Results German-Czech Translation

For the de-cs translation three different models with TAG, namely GPT-4o-Mini, GPT-4o, and Mistral 7B were used, and the non-augmented GPT-4o-Mini was used as a baseline for comparison. The results show

Table 3 Some of the most common errors across all models from the de-en translation. The predictions here are examples from GPT-4o, however these errors are also occurred with the other models.

#	Expected Term	Source	Prediction	Comment
1_27	pull-out	In Frankreich hingegen hat die liegende Lagerung der Weine hohe Priorität, am besten in einem Weinschrank (armoire à vin) direkt neben dem praktischen *Baguetteauszug*.	In France, on the other hand, the horizontal storage of wines has high priority, preferably in a tall cabinet (armoire à vin) right next to the practical *baguette drawer*.	The compound word *Baguetteauszug* (*baguette pull-out*) is not in the termbase, but only the term *Auszug* (*pull-out*)
1_47	space requirement	Das ergibt unterschiedliche Ansprüche an den *Stauraum* in der Küche.	This results in different demands on the *storage space* in the kitchen.	The term *Platzbedarf* (*space requirements*) is not in the source text, but the term *Stauraum* (*storage space*)
1_50	mounting plate	Diese kann mit Hilfe von verdreht verbauten Scharnieren auch elegant hinter einer nach innen öffnenden Tür versteckt werden, die sich in die Küchenfront perfekt integriert.	This can also be elegantly hidden behind an inward-opening door, which is perfectly integrated into the kitchen front, with the help of cleverly installed hinges.	The term *mounting plate* is only in the reference translation and not the source text
2_38	height, living room	Durch die flexiblen Maße der Schranklösung konnte der Tischler die passende *Größe* wählen und den Plattenschrank so gestalten, dass er in unmittelbarer Nähe des Plattenspielers stehen kann und sich harmonisch in den *Wohnraum* einfügt.	Due to the flexible dimensions of the cabinet solutions, the cabinet maker was able to choose the appropriate *size* and design the record cabinet so that it can stand in close proximity to the turntable and harmoniously blend into the *living space*.	The terms *Höhe* (*height*) and *Wohnzimmer* (*living room*) are not in the source text but synonyms (*Größe*, *Wohnraum*) that are not in the termbase
3_14	step	„Luisa verwendet die *Sockellösung* gern, um sich die Schuhe anzuziehen.	"Luisa likes to use the *plinth solution* to put on her shoes."	The term *Sockellösung* (*plinth solution*) and not *step* was used in the source text

Table 4 Comparison of a non-augmented LLM (GPT-4o-Mini) and terminology augmented LMs (GPT-4o-Mini, GPT-4o, Mistral 7B) for de-cs segment-based translation on the custom dataset. ↑ and ↓ represent significantly better and worse scores than the baseline system with p-value < 0.05.

Model	Term %	SacreBLEU	COMET	Time(s)
GPT-4o-Mini	63.33	30.87	0.9001	**2.00**
GPT-4o-Mini with TAG	90.64 (+27.31) ↑	31.67 (+0.8) ↑	0.9017 (+0.0016)	2.70
GPT-4o with TAG	**95.42 (+32.09)** ↑	**34.30 (+3.43)** ↑	**0.9117 (+0.0116)** ↑	3.18
Mistral 7B with TAG	-	-	-	-

that using terminology-augmented LLMs is also beneficial for the translation of de-cs, as both GPT-4o-Mini and GPT-4o with TAG outperform the non-augmented LLM significantly in the term adherence score (see Table 4). Moreover, these two models with TAG also significantly outperformed the non-augmented model in the SacreBLEU, and GPT-4o with TAG also outperformed the baseline regarding the COMET score.

Compared to the de-en translation, there is a greater variance between the models with TAG performance regarding the term adherence. Although both models yielded similar results for the de-en translation, for the term adherence with a difference of only 0.52%, for the de-cs translation the larger GPT-4o outperforms the smaller GPT-4o-Mini by almost 5%, and generally scores better results for SacreBLEU and COMET too. Similar to the de-en translation, the Mistral 7B model did not generate any usable translations, which is why its output could not be evaluated.

The analysis of the terminology errors showed that similarly to the de-en translation most of the errors made by the terminology-augmented LLMs occurred due to terminological mistakes in the human source

Table 5 Some of the most common errors across all models from the de-cs translation. The predictions here are examples from GPT-4o, however these errors are also occurred with the other models.

#	Expected Term	Source	Prediction	Comment
1_58	úložný prostor	Wer die Küche also lieber nur als Ablagefläche verwendet, kann die Beweise dafür schnell hinter dem Pocketsystem REVEGO verschwinden lassen.	Kdo tedy raději používá kuchyň pouze jako odkládací plochu, může důkazy rychle schovat za systémem kapes REVEGO.	The term *Stauraum* (*úložný prostor*) is not in the source text
1_86	stojánek na kořenky	Es gibt besondere Möbel wie den Abtropfschrank in Finnland, einen speziellen *Gewürzbehälter* (Masala Dabba) in Indien oder einen eigenen Weinschrank (armoire à vin) in Frankreich.	Ve Finsku existují speciální kusy nábytku, jako je odkapávací skříň, v Indii speciální *kořenková nádoba* (Masala Dabba) nebo ve Francii vlastní vinná skříň (armoire à vin).	The term *Gewürzhalter* (*stojánek na kořenky*) is not in the source text, but a synonym (*Gewürzbehälter*) which is not in the termbase
2_52	box systémy	Mit allen *Box-* und Führungs*systemen* von Blum umsetzbar	Se všemi *boxovými* a vodicími *systémy* od Blum realizovatelné	The term (*box systémy*) was used as an adjective (*boxovými systémy*)
2_57	výsuv	Selbst schwer beladen einfach zu bedienen	I těžce naložený snadno ovladatelný	The term *Auszug* (**výsuv**) is not in the source text
4_4	šroubovák	„Bereits als junger Bub, habe ich an allem herumgeschraubt und wollte Mamas Möbel optimieren“, erzählt der 43-Jährige.	„Už jako malý kluk jsem na všem kutal a chtěl vylepšovat mámin nábytek,“ vypráví třiačtyřicetiletý.	The term *Schraubenzieher* (*šroubovák*) is not in the source text

and reference translation (see Table 5 for some of the most common errors). Again, several terms were missing or used incorrectly in the source text, but present in the reference translation.

The same issues concerning compound words that occurred for the de-en translation are also present for the de-cs translation, which emphasizes the need to further investigate and test the search parameters used for the retrieval.

In contrast to the de-en translation results, another type of error was found where Czech terms were not used as is, but derived or slightly changed by the LLMs. For example, for the compound Czech noun *box systémy* (*box system*), the first part of the compound term (*box*) was turned into an adjective (*boxovými*), resulting in *boxovými systémy*.

In some cases, words from the German source text were not translated into Czech, or both the German source word as well as the Czech translation was provided in the output. This happened with both GPT-4o-Mini and GPT-4o, and for terms and non-terms in the source text (see Table 6).

It is important to note that even though the SacreBLEU and COMET scores are relatively high for the de-cs translation with the LLMs, and most of the terminology was correctly used by the terminology-augmented models, our language specialists that reviewed the errors mentioned that translation quality was poor, contained grammatical as well as semantic errors, and did not make sense in some cases.

4.3 Results German-Italian Translation

For the de-it translation the same three terminology-augmented LLMs as with the de-cs translation were used: GPT-4o-Mini, GPT-4o, and Mistral 7B, and GPT-4o-Mini without TAG was used as a baseline.

The baseline model GPT-4o-Mini scored the lowest results across all models and languages for the terminology adherence with slightly less than 50% correctly translated terms. Using TAG with GPT-4o-Mini and GPT-4o increased this result greatly, and scored up to 90% correctly translated terms (see Table 7). Although there is some variance between the term adherence performance of the two models, it is not as stark as for the de-cs translation with around 3% improvement when using GPT-4o over GPT-4o-Mini. Furthermore, the terminology-augmented models significantly outperformed the non-augmented model in the SacreBLEU and COMET scores.

Table 6 Examples of German words (terms and non-terms) that were not or only partly translated into Czech

#	Model	Source	Prediction	Reference
12_96	GPT-4o	*Tablar, Küche, Badezimmer, Home-Office, Wohnzimmer, Oberschrank, BLUMOTION, TIP-ON*	*Tablar*: policový výsuv *Küche*: kuchyně *Badezimmer*: koupelna *Home-Office*: domácí kancelář *Wohnzimmer*: obývací pokoj *Oberschrank*: horní skříňka *BLUMOTION*: BLUMOTION TIP-ON: TIP-ON	koupelna, domácí kancelář, obývací pokoj, horní skříňka, BLUMOTION, TIP-ON
6_8	GPT-4o-Mini	*Oft* ist das Arbeitsumfeld zu Hause viel entspannter als im Büro.	*Oft* je pracovní prostředí doma mnohem uvolněnější než v kanceláři.	Pracovní prostředí doma je často mnohem uvolněnější než v kanceláři.
6_90	GPT-4o-Mini	*AVENTOS-Klappen* bieten uneingeschränkten Zugriff und zusätzlichen Stauraum	*AVENTOS-Klappen* nabízejí neomezený přístup a dodatečný úložný prostor.	Výklopy AVENTOS nabízejí neomezený přístup a další úložný prostor

Table 7 Comparison of a non-augmented LLM (GPT-4o-Mini) and terminology augmented LMs (GPT-4o-Mini, GPT-4o, Mistral 7B) for de-it segment-based translation on the custom dataset. ↑ and ↓ represent significantly better and worse scores than the baseline system with p-value < 0.05.

Model	Term %	SacreBLEU	COMET	Time(s)
GPT-4o-Mini	49.38	34.05	0.8542	**1.94**
GPT-4o with TAG	**90.70 (+41.32)** ↑	**39.06 (+5.01)** ↑	**0.8739 (+0.0197)** ↑	2.64
GPT-4o-Mini with TAG	87.66 (+38.28) ↑	37.31 (+3.26) ↑	0.8651 (+0.0109) ↑	2.36
Mistral 7B	-	-	-	-

Most of the terminology errors that occurred with the models for the de-it translation (see Table 8 for some of the most common errors) stem from the same terminological errors in the human source text, and reference translation that have already been identified for the de-en and de-cs translation.

Several errors can be traced back to additional mistakes in the human reference translation. In the termbase, a distinction is made between the terms *pressione* (*touch*) and *tocco leggero* (*light touch*). However, this distinction is not correctly used in the human reference translation: whenever the German term for *light touch* (*leichtes Antippen*) is used in the source text, it is translated as *touch* (*pressione*). As a result, when the terminology-augmented models used the correct translation for *light touch*, in their translation it was counted as an error since the term for *touch* and not *light touch* was in the human reference translation. Additionally, the difficulties with compound words are also found with the de-it translation, and similar to the de-cs translation, in some cases single source words were not translated into Italian, or both the German and Italian words were provided in the generated output. In particular, the smaller GPT-4o-Mini model both non-augmented and terminology-augmented, did not translate several words (terms and non-terms) into Italian (see Table 9).

4.4 Results Terminology Formats (de-en Translation)

For the evaluation of the impact of the terminology format on the performance of TAG for translation, the following five different formats were used: JSON, Markdown, TBX (DCA), TBX (DCT), and YAML (see Section 3.1 for an overview of the terminology formats). These formats were evaluated with the two best performing models concerning the term adherence for de-en translation: GPT-4o (see Table 10 and Table 11) and GPT-4o-Mini (see Table 12). Previously, the system prompt in Appendix A.1 was used for the translation with TAG. Since later tests indicated that the description of the format in the system prompt does not have a significant impact on the performance of the model with TAG, we decided to adjust the system prompt, and leave out the terminology format description (see Appendix A.2).

For segment-based translation, the results with GPT-4o show marginal changes across all scores for the different terminology formats. While GPT-4o with TAG significantly outperformed the non-augmented

Table 8 Some of the most common errors across all models from the de-it translation. The predictions here are examples from GPT-4o, however these errors are also occurred with the other models.

#	Expected Term	Source	Prediction	Comment
1_17	lavastoviglie	Hand oder Maschine	Mano o macchina	The term *Geschirrspüler* (*lavastoviglie*) is not in the source text
1_87	portaspezie	Es gibt besondere Möbel wie den Abtropfschrank in Finnland, einen speziellen *Gewürzbehälter* (Masala Dabba) in Indien oder einen eigenen Weinschrank (armoire à vin) in Frankreich.	Ci sono mobili particolari come il mobile scolapiatti in Finlandia, un *contenitore speciale per spezie* (Masala Dabba) in India o un proprio mobile a colonna per vini (armoire à vin) in Francia.	The term *Gewürzhalter* (*portaspezie*) is not in the source text, but a synonym (*Gewürzbehälter*) which is not in the termbase
3_12	zoccolo	Luisa steht auf, gibt der Schublade einen kleinen Stoß und sie verschwindet im *Sockel*.	Luisa si alza, dà una piccola spinta al cassetto e questo scompare nel *basamento*.	The term *Sockel* (*zoccolo*) is only as part of compound terms in the termbase (e.g. *Sockellösung*)
3_24	pedana	Ein kleines blondes Mädchen sitzt auf einer ausgezogenen *Tritthilfe* SPACE STEP und zieht einen blauen Gummistiefel an	Una piccola bambina bionda è seduta su una *SPACE STEP* estratta e si infila uno stivale di gomma blu.	The term *Trittfläche* (*pedana*) is not in the source text, but a synonym (*Tritthilfe*) which is not in the termbase
3_74	pressione	Ein *leichtes Antippen* genügt, beide Hände bleiben frei und tiefes Bücken wird überflüssig.	Un *tocco leggero* è sufficiente, entrambe le mani rimangono libere e non è necessario chinarsi profondamente.	*Leichtes Antippen* should have been translated as *tocco leggero* and not *pressione* (*Antippen*)

Table 9 Examples of German words (terms and non-terms) that were not or only partly translated into Italian

#	Model	Source	Prediction	Reference
6_8	GPT-4o-Mini	In Frankreich hingegen hat die *liegende* Lagerung der Weine hohe Priorität, am besten in einem Weinschrank (armoire à vin) direkt neben dem praktischen *Baguetteauszug*.	In Francia, invece, la *liegende* Traverse dei vini ha alta priorità, meglio in un mobile a colonna direttamente accanto al pratico *Baguetteauszug*.	In Francia invece è prioritario poter riporre le bottiglie di vino in posizione orizzontale, possibilmente in una credenza dedicata (armoire à vin) direttamente accanto al pratico cassetto per le baguette.
6_90	GPT-4o-Mini	Ein kleines blondes Mädchen sitzt auf einer ausgezogenen *Tritthilfe* SPACE STEP und zieht einen blauen Gummistiefel an	Una piccola ragazza bionda è seduta su una *Tritthilfe* SPACE STEP estratta e indossa un stivale di gomma blu.	Una bambina bionda si siede su una pedana estratta SPACE STEP e indossa uno stivale di gomma blu
14_55	GPT-4o-Mini	ORGA-LINE-*Inneneinteilung* kombiniert Schalen mit Rahmen und ist dadurch besonders flexibel einsetzbar.	L'*inneneinteil*azione ORGA-LINE combina suddivisioni interne con vaschette ed è quindi particolarmente flessibile nell'uso.	La suddivisione interna ORGA-LINE combina vaschette e telai e può quindi essere utilizzata in modo particolarmente flessibile.

GPT-4o, the terminology format itself does not seem to have a great impact on the terminology adherence with a difference of slightly less than 4% between the best and worst performing terminology format. During the analysis of the terminology errors, it became evident that most of the errors were the same for all five different terminology formats. Moreover, several errors made by GPT-4o with the two worst performing terminology formats - Markdown and YAML - were errors where another admitted term of the concept was used, for example the acronym of a term. Interestingly, while the model varied between the

Table 10 Comparison of a non-augmented LLM (GPT-4o) and a terminology augmented LLM (GPT-4o) with five different terminology formats (JSON, Markdown, TBX (DCA), TBX (DCT), YAML) for de-en segment-based translation on the custom dataset. ↑ and ↓ represent significantly better and worse scores than the baseline system with p-value < 0.05.

Terminology Format	Term %	SacreBLEU	COMET	Time(s)
Without TAG	70.02	33.24	0.8574	**2.23**
JSON	90.88 (+20.86) ↑	**34.52 (+1.28)** ↑	0.8625 (+0.0051) ↑	2.83
Markdown	87.45 (+17.43) ↑	34.03 (+0.79) ↑	0.8586 (+0.0012)	2.99
TBX (DCA)	**91.31 (+21.29** ↑	34.45 (+1.21) ↑	0.8626 (+0.0052) ↑	3.20
TBX (DCT)	91.00 (+20.98) ↑	34.30 (+1.06) ↑	**0.8627 (+0.0053)** ↑	3.06
YAML	89.78 (+19.76) ↑	34.27 (+1.03) ↑	0.8616 (+0.0042) ↑	3.49

Table 11 Comparison of a non-augmented LLM (GPT-4o) and a terminology augmented LLM (GPT-4o) with five different terminology formats (JSON, Markdown, TBX (DCA), TBX (DCT), YAML) for de-en paragraph-based translation (each paragraph is up to 1,000 characters long) on the custom dataset. ↑ and ↓ represent significantly better and worse scores than the baseline system with p-value < 0.05.

Terminology Format	Term %	SacreBLEU	COMET	Time(s)
Without TAG	74.11	33.19	0.8643	4.96
JSON	91.46 (+17.35) ↑	35.49 (+2.30) ↑	0.8714 (+0.0071) ↑	4.79
Markdown	92.32 (+18.21) ↑	35.33 (+2.14) ↑	0.8712 (+0.0069) ↑	**4.41**
TBX (DCA)	90.65 (+16.54) ↑	34.98 (+1.79) ↑	0.8717 (+0.0074) ↑	7.38
TBX (DCT)	90.57 (+16.46) ↑	35.37 (+2.18) ↑	**0.8728 (+0.0085)** ↑	6.80
YAML	**92.40 (+18.29)** ↑	**35.50 (+2.31)** ↑	0.8716 (+0.0073) ↑	4.42

Table 12 Comparison of a non-augmented LLM (GPT-4o-Mini) and a terminology augmented LLM (GPT-4o-Mini) with five different terminology formats JSON, Markdown, TBX (DCA), TBX (DCT), and YAML for de-en paragraph-based translation (each paragraph is up to 1,000 characters long) on the custom dataset. ↑ and ↓ represent significantly better and worse scores than the baseline system with p-value < 0.05.

Terminology Format	Term %	SacreBLEU	COMET	Time(s)
Without TAG	68.18	32.14	0.8557	6.09
JSON	87.53 (+19.35) ↑	**34.28 (+2.14)** ↑	0.8655 (+0.0098) ↑	**6.05**
Markdown	86.41 (+18.23) ↑	33.68 (+1.54) _	0.8630 (+0.0073) ↑	**6.05**
TBX (DCA)	83.59 (+15.41) ↑	34.08 (+1.94) ↑	0.8655 (+0.0098) ↑	6.48
TBX (DCT)	83.71 (+15.53) ↑	34.02 (+1.88) ↑	0.8652 (+0.0095) ↑	6.74
YAML	**88.08 (+19.9)** ↑	34.01 (+1.87) ↑	**0.8667 (+0.0110)** ↑	6.24

admitted terms of a concept when using YAML or Markdown as a terminology format, it almost exclusively used only the first admitted term of a concept with the two TBX formats.

To evaluate the effectiveness of TAG for long-context translation, we also performed a paragraph-based translation with the five different terminology formats and GPT-4o. The dataset remained the same, but instead of splitting it into sentences or segments, it was split into paragraphs with a length of up to 1,000 characters. The results for the paragraph-based translation indicate slight increases in the SacreBLEU and COMET scores across all five terminology formats, compared to the segment-based translation.

Interestingly, while terminology adherence increased for Markdown, YAML, and JSON formats, a slight decrease is noticeable for the two TBX formats. The performance of the non-augmented model also increased across all scores, and in particular for the term adherence an increase by slightly more than 4% is noticeable.

For paragraph-based translation, the average translation time increased greatly, and was approximately doubled for most terminology formats. This increase in latency is mainly caused by the larger number of tokens that need to be processed and generated by the model. However, it is noticeable that the translation time for the two TBX formats is significantly longer than for the segment-based translation when compared to the other formats, suggesting a longer preprocessing of the context tokens when compared to smaller terminology formats.

GPT-4o-Mini was chosen as the second model for evaluating TAG performance for paragraph-based translation with five different terminology formats. In comparison to the paragraph-based translation with GPT-4o, the performance across all metrics is lower, similar to the performance of the segment-based translation with GPT-4o. The lower performance is most likely due to the fact that GPT-4o-Mini is a smaller model.

Concerning the term adherence, both models performed best with YAML as the terminology format, and the two TBX formats performed the worst for the paragraph-based translation, bolstering the hypothesis that the large amount of tags in TBX formats is suboptimal for TAG.

5 Conclusion

In this work, we performed automatic segment-based translation (de-en, de-cs, de-it) on a custom dataset using different LLMs including GPT-4o, GPT-4o-Mini, Mistral Large 2, Mistral 7B, Mistral NeMo, and Gemma3:12B, with the focus on the correct translation of terminology using TAG.

Furthermore, we tested TAG with five different terminology formats including JSON, Markdown, TBX (DCA), TBX (DCT), and YAML, for segment- and paragraph-based de-en translation with two different models (GPT-4o-Mini, GPT-4o), to evaluate the impact of the format in which the terminological information is provided. The paragraph-based translation with paragraphs of up to 1,000 characters was carried out to evaluate the long-context performance of TAG.

In general, the results indicate that TAG is effective for correctly translating terminology, and compares favorably to LLMs without terminology augmentation. TAG not only led to an increase in the terminology adherence, but also slightly increased the SacreBLEU and COMET scores in most cases.

While we expected to see differences across the three different language pairs (de-en, de-cs, de-it), the scores for all metrics were similarly high for all language pairs. However, the difference in performance of the GPT-4o and GPT-4o-Mini model for the term adherence score varied more with the de-cs and de-it translation, than with the de-en translation, suggesting bigger gaps in multilinguality for the smaller model. Furthermore, our human evaluation determined that the general translation quality of Czech and even Italian was poor, suggesting the LLMs still are not ready to fully replace NMT. However, that was not the main focus of this evaluation.

In conclusion, the results show only a small difference in performance with the different terminology formats, especially for sentence-based translation. However, we were able to observe the tendency that some terminology formats are slightly better than others, with the YAML format emerging as the most effective, especially for paragraph-based translation.

Although most of our research was based on traditional segment-based translation, the results of the paragraph-based translation showed that providing the model with more context and not only single segments improves all scores, including the terminology adherence even for non-augmented models.

Crucially, we observed that most terminology errors could be traced back to terminological incompleteness (e.g. missing definitions), or inconsistencies in the termbase, and incorrect or missing use of terminology in the human source text and reference translation, highlighting the need for good and exhaustive terminology resources and terminologically consistent source texts.

6 Limitations and Future Work

In future work, a comparison between LLMs using TAG and state-of-the-art NMT with Glossaries is something we would like to cover in more depth. Another interesting topic to explore would be to test and evaluate different search parameters for the retrieval, to focus on more advanced ways to retrieve terminology from termbases. For example, by using dense or sparse vector retrieval, or graph-based approaches, making use of relational information in advanced terminological systems. These methods could enhance the accuracy of term recognition.

As already mentioned in Section 3.2, the applied terminology adherence evaluation takes the terminology extracted from the human reference translation regardless of whether it is also present in the source text. This approach was chosen since we assumed that the human reference translation would contain the same terminology as the source text. However, the analysis of the terminology errors revealed that there were differences in the terminology used in human source text and reference translation (see Section 4).

This highlights that a different approach for evaluating the terminology adherence, that considers the terminology in both the human source text and reference translation, could be beneficial and lead to more representative results.

Furthermore, it might make sense to rethink the approach of handling admitted terms that are correct in the given context, but were not used in the human reference translation (e.g., acronyms). We decided to only evaluate the terms in the human reference translation as correct. Alternatively, one could approve all admitted terms that have been correctly used in the provided context, include additional instructions in the system prompt, to restrict the use of certain admitted terms like acronyms, or to add additional information to the termbase like a usage note, to help with the disambiguation.

It would similarly be interesting to also evaluate the false positives of terminology-augmented translation, since we noticed that "with TAG" models have the tendency to use terminology even if there is no terminology present in the sentence. For example, the German term *Distanz* (*spacing*) is in the termbase, but in a sentence it was not used as a term and should have been translated as *distance*.

While we focused mainly on segment-based translation, it might be promising to further investigate paragraph-based translation with TAG, and analyze the general impact of different language varieties (e.g. de-DE vs. de-AT or en-GB vs en-US) on both LLM behavior and TAG effectiveness.

Finally, in this work we only covered the use case of MT, yet there remain many use cases that could be explored with TAG, including checking and revision of terminology, automatically generating texts based on specific terminology, augmenting user prompts for various AI systems, and/or intelligently labeling data. We hope this work can serve both as a foundation for future refinements and evaluations of TAG approaches, as well as an inspiration to explore new applications of terminology in AI.

References

Baek, J., Hussain, A., Liu, D., Vincent, N., Kim, L.H. (2025). *Open webui: An open, extensible, and usable interface for ai interaction.* Retrieved from https://arxiv.org/abs/2510.02546

Dinu, G., Mathur, P., Federico, M., Al-Onaizan, Y. (2019). Training neural machine translation to apply terminology constraints. A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3063–3068). Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1294

Di Nunzio, G.M. (2025). Terminology-Augmented Generation (TAG): Foundations, Use Cases, and Evaluation Paths. *Journal of Digital Terminology and Lexicography*, *1*(1), 97-104, https://doi.org/10.25430/pupj.jdtl.1752566034

Exel, M., Buschbeck, B., Brandt, L., Doneva, S. (2020). Terminology-constrained neural machine translation at SAP. A. Martins et al. (Eds.), *Proceedings of the 22nd annual conference of the european association for machine translation* (pp. 271–280). Lisboa, Portugal: European Association for Machine Translation. Retrieved from https://aclanthology.org/2020.eamt-1.29/

Fleischmann, K. (2025). Terminologiemanagement: Die Schlüsselkomponente für effiziente Kommunikation in Unternehmen. *Information–Wissenschaft & Praxis*, *76*(4), 169–176, https://doi.org/10.1515/iwp-2025-2005

Gupta, S., Ranjan, R., Singh, S.N. (2024). *A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions.* arXiv. Retrieved 2025-01-25, from https://arxiv.org/abs/2410.12837

Heinisch, B. (2025). Large language models for terminology work: A question of the right prompt? *Journal for language technology and computational linguistics*, *38*(2), 13-30, https://doi.org/10.21248/jlcl.38.2025.280

ISO 30042 (2019). *Management of terminology resources — termbase exchange (tbx).* International Organization for Standardization. Retrieved from https://www.iso.org/standard/62510.html

ISO 639 (2023). *Code for individual languages and language groups.* International Organization for Standardization. Retrieved from https://www.iso.org/standard/74575.html

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., … Sayed, W.E. (2023). *Mistral 7B.* arXiv. Retrieved 2024-12-30, from http://arxiv.org/abs/2310.06825

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. D. Lin & D. Wu (Eds.), *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 388–395). Barcelona, Spain: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W04-3250/

Lackner, A., Vega-Wilson, A., Lang, C. (2025). Terminology Augmented Generation: A Systematic Review of Terminology Formats for In-Context Learning in LLMs. F. Vezzani, G.M. Di Nunzio, E. Loupaki, G. Meditskos, & M. Papoutsoglou (Eds.), *Proceedings of the 4th International Conference on Multilingual Digital Terminology Today (MDTT 2025)* (Vol. 3990). Retrieved from https://ceur-ws.org/Vol-3990/

Lang, C., & den Nest, E.V. (2025). *Die prüfung und freigabe von Übersetzungen wird noch lange zeit menschlich bleiben: Christian lang leitet die neu gegründete ki-abteilung des sprachtechnologieunternehmens kaleidoscope und spricht über die zukunft von Übersetzungsdienstleistungen und warum menschen dabei nicht vollständig ersetzbar sind* (AMS info No. 724). Wien. Retrieved from https://hdl.handle.net/10419/324191

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.* arXiv. Retrieved 2025-01-25, from https://arxiv.org/abs/2005.11401

LTAC Global (2021). *Tbxinfo.* Retrieved 2025-02-07, from https://www.tbxinfo.net/

Mistral AI Team (2023). *Mistral 7B.* Retrieved 2025-09-07, from https://mistral.ai/news/announcing-mistral-7b

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318). USA: Association for Computational Linguistics. Retrieved from https://doi.org/10.3115/1073083.1073135

Post, M. (2018). A call for clarity in reporting BLEU scores. O. Bojar et al. (Eds.), *Proceedings of the third conference on machine translation: Research papers* (pp. 186–191). Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Rei, R., Stewart, C., Farinha, A.C., Lavie, A. (2020). COMET: A neural framework for MT evaluation. B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 2685–2702). https://doi.org/10.18653/v1/2020.emnlp-main.213

A System Prompt

A.1 System Prompt with Terminology Format Description

You are a translator and author. The user will provide text to be translated and indications which terminology to use.

Task description
* Translate the text provided by the user from and into the language the user specifies.
* Make sure the translation sounds natural.
* The user specifies the translation direction by prefixing the text to be translated with the following string: 'Translate sourceLanguage to targetLanguage:'
* Use the definition to disambiguate the meaning of the terminology passed by the user and translate accordingly
* Only return the translation

Terminology
* If available, the user will provide indication on what terminology to use
* Follow the suggestions provided within the <tag>-XML elements of the user message

Terminology format
* The terminology will be provided in the format below within the <tag>-XML Elements of the user message:
"'markdown
Concept 1
* Definition of concept 1
source term 1
Possible translations:
1. first possible translation
* Usage note of possible translation 1
2. second possible translation
* Usage note of possible translation 2
source term 2
Possible translations:
1. first possible translation
* Usage note of possible translation 1
2. second possible translation
* Usage note of possible translation 2

Concept 2
source term 2
Possible translations:
1. first possible translation
"'
* Note: not all terms will have a definition or usage note.

Rules
* Use the definition to disambiguate the meaning of term pairs
* Follow the usageNote of each possible translation to choose the most suitable translation, if more than one translation is provided
* If the system returns a term that is not present in the source text, ignore the term.

A.2 System Prompt without Terminology Description

You are a translator and author. The user will provide text to be translated and indications which terminology to use.

Task description
* Translate the text provided by the user from and into the language the user specifies.
* Make sure the translation sounds natural.
* The user specifies the translation direction by prefixing the text to be translated with the following string: 'Translate sourceLanguage to targetLanguage:'
* Use the definition to disambiguate the meaning of the terminology passed by the user and translate accordingly
* Only return the translation

Terminology
* If available, the user will provide indication on what terminology to use
* Follow the suggestions provided within the <tag>-XML elements of the user message
* Note: not all terms will have a definition or usage note.

Rules
* Use the definition to disambiguate the meaning of term pairs
* Use the definition to disambiguate the meaning of term pairs
* Use the usage note to find the most suitable translation, if more than one translation is provided
* If the system returns a term that is not present in the source text, ignore the term.