Terminology Augmented Generation: A Systematic Review of Terminology Formats for In-Context Learning in LLMs

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

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

Abstract

We present our on-going work on a specialized extension of the Retrieval Augmented Generation (RAG) framework focusing on providing knowledge from enterprise terminology databases to generative LLMs: Terminology Augmented Generation (TAG). This study specifically focuses on the role that terminology formatting plays for TAG across common NLP downstream tasks such as translation and terminology revisions of texts. By conducting empirical evaluations using OpenAI's GPT-4o, GPT-4o-mini, and the open-source LLaMA 3.3 and Mistral 7b models, we systematically explore various established terminology formats (including TBXv3) and compare the results to alternative structured and prose formats and their impact on generation quality. Preliminary findings indicate that specific formatting strategies significantly improve model accuracy and recall of in-context knowledge, as well as the disambiguation capabilities in linguistically ambiguous scenarios. This research provides valuable insights into the design of terminology integration methodologies for LLMs, contributing to the development of more effective language processing systems that meet the nuanced demands of professional and technical communication.

Keywords

NLP, LLM, RAG, Retrieval Augmented Generation, TAG, Terminology Augmented Generation, Neural Machine Translation, terminology management, terminology evaluation, terminology revision

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 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)[1]. 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 [1], [2]. 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

---

[*] Corresponding author.
[†] These authors contributed equally.

✉ anna.lackner @kaleidoscope.at (A. Lackner); alena.vega-wilson@eurocom.at (A. Vega-Wilson), christian.lang@kaleidoscope.at (C. Lang)

3rd International Conference on “Multilingual digital terminology today. Design, representation formats and management systems” (MDTT) 2025, June 19-20, 2025, Thessaloniki, Greece."

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3990/short10.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

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 various down-stream tasks with LLMs. We describe TAG in detail in our German publication for the conference proceedings of the DTT-Symposion 2025 [3], but to briefly summarize it, TAG methods should be able to retrieve terminology in real-time from terminology management systems (TMS) like Kalcium Quickterm and/or format the terminological context in a way, that it can be efficiently parsed by the LLM for in-context learning. 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. While LLMs demonstrate remarkable abilities to parse XML – the typical terminology exchange format that is also standardized in the TBX XML specification [4] – 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 structure terminological knowledge to LLMs at run-time.

2. Methodology

We explore two different distinct use cases for TAG: Machine Translation and Automatic Terminology Revision. Since these are two distinct tasks, we follow a slightly different experimental set up for the individual tasks described in the following sub-sections.

2.1. LLM setup

To instruct the LLMs for the task, we set up one system prompt for each of the tasks, that is shared between all models. For the terminology augmentation we explore a variety of possible formats, ranging from the native XML output of terminology systems to other structured language outputs such as JSON, YAML, Markdown or ad-hoc generated "prose" instructions for using the relevant terminology. Since most of our prior testing was done using OpenAI models, our prompting techniques are likely to favor OpenAI trained models. This evaluation is therefore not to be interpreted as a comparison between different models, but rather an exploration of various effects of different prompting formats for TAG with different LLM backends. Nevertheless, we experiment using 4 popular models: 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-source Llama-3.3 70b[2] (4bit int quantized) and Mistral 7b[3] (4bit int quantized) models. For all models we set 3 major hyper-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 *seed* is set to 42. While this still allows for some variance, we found that setting *Temperature* lower tends to reduce the perceived and measured quality of the output.[4]

To access the models, we plan to employ the open-source AI interface Open WebUI [5][5] in combination with open-source ollama[6] framework, which allows us to access both open-source and proprietary models via one single OpenAI-conformant REST-API interface. While Open WebUI also allows us to set up custom pipelines which include TAG via the Kalcium REST-API, we implement the actual TAG code in Python to make the testing reproducible with other OpenAI compatible endpoints. The TAG code itself is based on the Kalcium REST-API[7], which provides ready-made

---

[2] https://ollama.com/library/llama3.3:70b-instruct-q4_0

[3] https://ollama.com/library/mistral:7b-instruct-q4_0

[4] Please note, we might update the models or hyper-parameters for the final release of this paper, to reflect our actual testing conditions.

[5] https://github.com/open-webui/open-webui

[6] https://github.com/ollama/ollama

[7] https://demo.kaleidoscope.at/kalcrest/swagger/index.html

endpoints for advanced term recognition using various established term recognition methods such as fuzzy-matching and stemming. We use the "/kalcrest/terminology/analyze-sentence" endpoint for all of our tests and parse the returned JSON as required.

2.2. Task-specific setup

For Machine Translation, we follow the experimental setup of Dinu et al. [6] using the WMT 2017 English-German news translation task[8] to evaluate our approach to a valid baseline (however, we use LLMs instead of custom NMT models). Additionally, we examine a small custom test set in 3 language directions: German (Austria) → Italian, German (Austria) → Czech and German (Austria) → English (US/GB). We chose these language pairs, since they are common pairs for our customer base. In line with Dinu et al., we evaluate the Machine Translation results with BLEU[7] and additionally also use COMET[8]. Furthermore, we apply a fuzzy matching strategy similar to Exel et al. [9] to detect if the correct terminology was used in the translation. Specifically, we stem words using the stemming engine in Kalcium Quickterm and perform a fuzzy search with a similarity rate of 80%. We acknowledge that this approach will not be perfect and might result in 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.

For terminology revision, we focus on monolingual revision for the same languages, i.e. German, Czech, Italian and English (US/GB), however, to the best of our knowledge, there exists no terminology revision test set, so we create our own based on sentences from public translation projects. Since the terminology revision task requires the model to only replace the invalid terminology with the correct terminology and, if necessary, also adapting the sentence grammar, we evaluate the generated sentences against the "correct" ground-truth sentences and only consider an exact match to the ground-truth sentences to be a successful revision.

2.3. Dataset

As described in 2.2, for the MT evaluation, we use the WMT 2017 English-German news translation task as a baseline, however, we also create a custom dataset to evaluate languages and terminology actually used by our customers. In prior unreleased work, we have already prepared a test set for terminology revision, consisting of various sentences from prior translation projects, which we modified with flawed terminology to be corrected by the LLM. For this first examination, we provided the term replacement pairs directly with the test data. For this work, we plan to rework the dataset to include approximately 200 source and target language sentences for each of the language pairs. Typical test sets are sentence-based, so we make an effort to sentence-align any equivalent sentences, but we also align the dataset on a paragraph and a document-level, allowing for the evaluation of long-context performance with TAG.

As for the terminology, we create three separate termbases: For the WMT 2017 test set we import the glossary created by Dinu et al. [6][9], based on the Wikitionary and IATE terminology. For our custom test set, we use existing terminological resources curated by the customer and our team or extract and create terminological entries from the test set as needed.

Note that the premise of TAG goes beyond term pairs and glossaries and rather aims to augment the generation with concept-oriented terminology, which means that terminological metadata like definition, usage status, usage note and other relevant information from the termbase will be used during the terminological augmentation. This information is not present in the glossaries created by Dinu et al., so to evaluate the effectiveness of TAG for additional capabilities like disambiguation we purposefully include homographs into the custom test sets and the corresponding termbase. The test

---

[8] https://www.statmt.org/wmt17/translation-task.html

[9] https://github.com/mtresearcher/terminology_dataset

set creation is currently an on-going process. We plan to release the test set for reproducibility, but since it will contain customer data, a survey of the legal feasibility needs to be done for the finished dataset.

3. Preliminary results and observations

As this evaluation project is currently still a work in progress, we present preliminary results of our evaluation for the camera-ready version of this paper. For the non-augmented models, we translate the WMT 17 test data (EN-DE) from Dinu et. al with a 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 English to German: {text}*"). For the model with TAG, we use a more complex prompt which is found in Annex B. As the terminology format, we used the *markdown* format, as described in the system prompt (note, that for this evaluation set, no information but the terms themselves were available). We measure the terminology adherence by looking for perfect matches of all terms present in the translation output of each sentence. This is possible, because the terminology contains the translated terminology (mostly) in the same morphological form, in which it is present in the text. In Table 1 we compare the results of GPT-4o, GPT-4o-Mini and GPT4-o with TAG against the *train-by-replacement* approach of Dinu et al. [6] as a baseline, as it had achieved the highest term adherences in their work. For your reference, we also include all sentences where terminology adherence was not achieved with TAG in appendix A.

Model	Term %	BLEU	COMET	Time(s)
Baseline	94.5	26.0	-	0.19
GPT-4o-Mini	87.4	33.7	0.876	1.95
GPT-4o	87.2	**35.7**	**0.880**	2.33
GPT-4o with TAG	**96.37**	35.5	0.877	2.04

Table 1: Comparison of prior methods to non-augmented LLMs and a terminology augmented LLM on the IATE annotated WMT 17 test set.

While the baseline approach manages to outperform non-augmented LLMs in terminology adherence, the LLMs achieve a notably higher BLEU score and crucially, the TAG approach further improves on the already high result on terminology adherence of the baseline approach, failing only in 15 out of the 414 sentences tested. Similarly to Dinu et al. we also observe high terminological adherence even for the non-augmented LLMs (52 and 53 failed examples respectively). However, during our exploration of the terminology used in the Baseline experiments and this preliminary evaluation, we observed several inherent issues with the terminological data itself:

1. The terminology extracted from IATE contains many common nouns like "month", "eggs", "tobacco" or country names like "Syria", which would generally be correctly translated by systems trained on common domain texts.
2. The terminology is not always in the infinite singular form (e.g. "eggs", "victories", "Schweizerin"), is translated as a *noun* when it should be a *verb* (e.g. "covering" → "Bezug" or "arrest" → "Festnahme") or contains articles (e.g. "Die Republikaner"). The last two examples account for 4 issues of the 15 encountered with TAG. (Appendix A: 5,7,11,15)
3. Around 30 homographs (out of 232 total terms) with inconsistent translations are present (e.g. "office" → "Büro" and "office" → "Amt"); since the terminology does not provide any kind of definition or usage recommendation, the LLM has no way to disambiguate the meaning or use a specific translation. This issue accounts for 9 issues of the 15 encountered with TAG. (Appendix A: 1,2,3,4,6,8,9,10,12)

Point 2 and 3 show that the quality and completeness of the terminological data is of high importance for TAG, especially when disambiguation is required. Two issues, which were likely not exclusively caused by deficiencies in the terminological data, but rather stem from either the generative nature of the model or our retrieval method could also be observed: In one instance, the model translated "night" with "Abend" (or rather composited it into "Donnerstagabend") even though "night" was provided with the German translation: "Nacht" (Appendix A: 14). In another instance, the model was provided with both "election campaign" and "campaign" source terms. and chose the wrong translation "Wahlkampf" instead of "Kampagne", which is somewhat related to the "homograph" issue and thus might be resolved with a proper definition of the terms or by using a lower fuzziness for the retrieval (Appendix A: 13).

All the remaining issues were caused through our preliminary way of checking terminology adherence, e.g. the LLM may have predicted a morphological variant or similar composite word instead of the expected terminology verbatim (e.g. "Vormonat" instead of "Monat" (Appendix A: 7). We will likely either use more refined NLP approaches or manual approaches to filter out these false negatives for the final results.

4. Future work

After the creation of the dataset, we aim to finalize our automated testing pipeline as described in section 2. From our experience employing TAG and the preliminary results shown in this paper, we expect to match or beat the baseline result achieved by prior methods for NMT ([6], [9]) with TAG, but are curious to see how well our hypothesis holds up for more complex or ambiguous examples in the custom test set. Especially, we are curious to see how well our hypothesis regarding specific terminology formats for TAG holds up on a larger test set and how the effectiveness of these formats may vary between various LLMs. However, this evaluation covers only a small part of the various systems that come into play during TAG. For example, the terminology retrieval method used for this work is efficient for the tasks at hand, but also rather limited for more open-ended downstream tasks. Future work could focus on more advanced ways to retrieve terminology from termbases, such as 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 but also allow autonomous AI agents to better navigate terminological resources to complete various downstream tasks, such as generating terminologically correct and hallucination-free text from scratch, e.g. for technical documentation, interactive support systems or efficiently navigating a specialized UI.

We hope this work can serve as both as a foundation for future refinements and evaluations to TAG approaches as well as an inspiration to explore new applications of terminology in AI. We look forward to sharing our results and discuss them at the MDTT 2025 in Thessaloniki, Greece.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 for grammar and spelling checks. The authors have subsequently reviewed and edited the content and take full responsibility for the publication's final version.

References

[1] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” 2020, *arXiv*. doi: 10.48550/ARXIV.2005.11401.

[2] S. Gupta, R. Ranjan, and S. N. Singh, “A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions,” 2024, *arXiv*. doi: 10.48550/ARXIV.2410.12837.

[3] K. Fleischmann and C. Lang, “Terminologie für die KI: Wie mit Terminologie der Output von LLMs und GenAI optimiert werden kann.,” in *Akten des Symposions*, vol. 27.–29. März 2025, P. Drewer, F. Mayer, and D. Pulitano, Eds., Worms: Deutscher Terminologie-Tag e.V., 2025.

[4] “Management of terminology resources — TermBase eXchange (TBX),” ISO, Standard ISO 30042:2019, 2019. [Online]. Available: https://www.iso.org/standard/62510.html

[5] “Designing an open-source LLM interface and social platforms for collectively driven LLM evaluation and auditing”.

[6] G. Dinu, P. Mathur, M. Federico, and Y. Al-Onaizan, “Training Neural Machine Translation to Apply Terminology Constraints,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 3063–3068. doi: 10.18653/v1/P19-1294.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL ’02*, Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001, p. 311. doi: 10.3115/1073083.1073135.

[8] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, “COMET: A Neural Framework for MT Evaluation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, 2020, pp. 2685–2702. doi: 10.18653/v1/2020.emnlp-main.213.

[9] M. Exel, B. Buschbeck, L. Brandt, and S. Doneva, “Terminology-Constrained Neural Machine Translation at SAP,” in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, and M. L. Forcada, Eds., Lisboa, Portugal: European Association for Machine Translation, Nov. 2020, pp. 271–280. [Online]. Available: https://aclanthology.org/2020.eamt-1.29/

A. Qualitative evaluation of terminology adherence with TAG

#	Comment	Term	Prediction (with TAG)	Source
1	Two entries for "state of emergency" with "Notstand" and "Ausnahmezustand" as translation (ids: 198, 202)	Notstand	Starker Regen und weit verbreitete Überschwemmungen in Louisiana führten dazu, dass der Gouverneur am Freitag den Ausnahmezustand ausrief, wobei für Samstag weiterer Regen über dem Staat erwartet wird.	Heavy rain and widespread flooding in Louisiana lead the governor to declare a state of emergency on Friday, with more rain expected over the state through Saturday.
2	Three entries for "strength" with "Widerstandsfähigkeit", "Kraft" and "Stärke" in German (ids: 169, 36, 76)	Stärke	Gymnastik entwickelt Kraft, Flexibilität und Koordination für den Körper sowie harte Arbeit, Disziplin und Entschlossenheit für den Geist.	Gymnastics develops strength, flexibility and coordination for the body and hard work, discipline and determination for the mind.
3	Two entries for "sign" with "Signal" and "Zeichen" in German (ids: 133, 220)	Zeichen	Arbeitgeber hofften, dass das anhaltend positive Engagement bei anderen wichtigen Themen - wie Einsatz, Flexibilität in der Ausbildung, zusätzliche Schulungen für Rückkehrer nach einer beruflichen Auszeit, Ausbildungskosten, gegenseitige Anerkennung des Lehrplans, Studienurlaub und das geschlechtsspezifische Entlohnungsgefälle in der Medizin - ein Signal dafür war, wie ernst es den Arbeitgebern, Health Education England und dem Gesundheitsministerium damit war, die mit der BMA im November, Februar und Mai getroffenen Vereinbarungen einzuhalten.	Employers were hopeful that the continued positive engagement on other important topics - such as deployment, flexibility in training, additional training for those returning from career breaks, costs of training, mutual recognition of syllabus, study leave and the gender pay gap in medicine - were a sign of how serious employers, Health Education England and the Department of Health were about honouring the agreements reached with the BMA in November, February and May.
4	Two entries for "election campaign" with "Wahlkampf" and "Wahlkampagne" in German (ids: 78, 11)	Wahlkampagne	Präsidentschaftskandidat Donald Trump hat nun die Schwächen seines Wahlkampfs in Utah, ehemals eine Hochburg der Republikaner, eingeräumt.	Presidential candidate Donald Trump has now admitted to the weaknesses of his election campaign in Utah, formerly a Republican stronghold.
5	Partly correct translated with "Republikaner" but the article wasn't used (which would not have made sense in this context). Only "republicans" (id: 3) is in the termbase and it still correctly translated "senate-republicans" with "Senats-Republikaner"	Die Republikaner	Senats-Republikaner haben die Bestätigung von Garland blockiert, seit Präsident Barack Obama ihn im März nominiert hat.	Senate Republicans have blocked Garland's confirmation since President Barack Obama nominated him in March.
6	Two entries for "sign" with "Signal" and "Zeichen" in German (ids: 133, 220)	Signal	Der Bundesverband deutscher Banken glaubt, dass dies ein Einzelfall ist und sieht es nicht als ein Zeichen: Normalsparer "müssen sich keine Sorgen machen, dass sie mit Strafzinsen auf die Pfennige, die sie beiseitegelegt haben, belastet werden".	The Federal Association of German Banks believes this is an isolated case, and does not see it as a sign: normal savers "need not worry about being hit with penalty interest on the pennies they've put aside".
7	"Vormonat" instead of Monat	Monat	Eine ICM-Umfrage im April ergab, dass fast 50 % der Personen, die im Vormonat ein Vinylalbum gekauft hatten, es noch nicht angehört hatten.	An ICM poll in April revealed that almost 50% of people who bought a vinyl album the previous month had yet to listen to it.
8	Three entries for "strength" with "Widerstandsfähigkeit", "Kraft" and "Stärke" in German (ids: 169, 36, 76)	Kraft	Fidschi gab eine Meisterklasse im Handling, Abspielen, Ausweichen, Unterstützungs-Spiel, Laufwege und brutale Stärke, um Gold zu gewinnen - ihre erste olympische Medaille in irgendeiner Farbe.	Fiji gave a masterclass in handling, off-loading, side-stepping, support play, running lines and brute strength to win gold - their first Olympic medal of any colour.
9	Two entries for "comparison" with "Vergleich" and "Gegenüberstellung" in German (ids: 147, 70)	Gegenüberstellung	Sie fügte hinzu, dass "es schwer ist, gegen den direkten Vergleich der sehr groben asiatischen Karikatur und der Effekte des Filters zu argumentieren."	She added that "it's hard to argue with the side by side comparison of the very gross Asian caricature and the filter's effects.

10	Two entries for "series" with "Serie" and "Reihe" in German (ids: 156, 62)	Reihe	Aufstrebende Meister springen auf einem riesigen Trampolin, schießen bis zu fast 30 Fuß in die Luft und werden aufgefordert, eine Serie von Turnbewegungen (Drehungen, Grätschen und Salti) auszuführen, bevor sie schließlich auf ihren Füßen landen.	Aspiring champions jump on a giant trampoline, shooting up to nearly 30 feet in the air, and are called upon to perform a series of tumbling moves (twists, pikes and somersaults) before eventually landing on their feet.
11	Term was used as a verb (beziehen instead of Bezug), the term in English was also used as a verb	Bezug	Die Beschwerden sind zahlreich und beziehen sich auf alles, von der Bodennutzung bis hin zum Würgegriff der Regierungskoalition auf die Macht.	The complaints are many, covering everything from land use to the governing coalition's stranglehold on power.
12	Three entries for "strength" with "Widerstandsfähigkeit", "Kraft" and "Stärke" in German (ids: 169, 36, 76)	Widerstandsfähigkeit	Moderator und Reiseautor Ian Wright, 51, aus Suffolk, strich mit seiner Zunge über einen Euro-Schein, ein Geländer am Piccadilly Circus, den Spülknopf einer Zugtoilette und ein russisches öffentliches Telefon, um die Stärke seines Immunsystems zu beweisen.	Presenter and travel writer Ian Wright, 51, from Suffolk, scraped his tongue along a Euro note, a Piccadilly Circus hand rail, a train toilet's flush button and a Russian public telephone to prove the strength of his immune system.
13	Incorrectly used the entry for "election campaign" instead of "campaign"	Kampagne	Bordinis Klage fordert Entschädigung vom Wahlkampf für fahrlässige Aufsicht und von Phillip wegen Körperverletzung, tätlichem Angriff und Zufügung emotionalen Leids.	Bordini's complaint seeks compensation from the campaign for negligent supervision, and from Phillip for assault, battery and infliction of emotional distress.
14	Incorrect terminology use	Nacht	Für Menschen, die am Donnerstagabend in Südwestdeutschland Sternschnuppen erhoffen, machte das Wetter ihren Plänen einen Strich durch die Rechnung - aber alles ist nicht verloren.	For people hoping to spot shooting stars in south-west Germany on Thursday night, the weather put paid to their plans - but all is not lost.
15	Term was used as a verb (festnehmen instead of Festnahme), the term was also used as a verb in English	Festnahme	Zivilbeamte der Düsseldorfer Polizei konnten am Donnerstag zwei Frauen und zwei Männer im Alter zwischen 50 und 61 Jahren festnehmen.	Plain clothes officers from Dusseldorf's police force managed to arrest two women and two men, aged between 50 and 61, on Thursday.

B. System prompt and terminology format for preliminary results

~~~
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.
~~~

Prompt 1: System prompt for GPT-4o with TAG