On the Use of LLMs for Specialised Terminology: A Good Alternative to Corpora?

Joachim Minder[1, 2] and **Guillaume Wisniewski**[2] and **Natalie Kübler**[1]
[1]Université Paris Cité, ALTAE, F-75013 Paris, France
[2]Université Paris Cité, CNRS, Laboratoire de linguistique formelle, F-75013 Paris, France

Abstract

Specialised translation relies on the use of documentary and terminological resources, including corpora. These resources are particularly useful for terminology. However, their compilation and exploitation have several limitations: they require time, technical skills and access to data that can be difficult to collect. This study examines the extent to which LLMs can assist specialised translators in finding equivalents from English to French. We evaluate four proprietary models, GPT-4o, GPT-5.2, Claude Sonnet 4.5 and DeepSeek, in two specialised domains, Earth, Environmental and Planetary Sciences (EEPS) and Natural Language Processing (NLP). The experiment is based on 80 terms per domain and compares two prompting strategies: a terminology and a translation mode. The results highlight clear differences between models, prompting strategies and, to a lesser extent, domains. Claude Sonnet 4.5 achieves the best results in the most favourable configuration, while DeepSeek stands out for its greater stability. Analysis of confidence estimates also shows that they are only a partial indicator of terminological accuracy. Overall, the findings suggest that LLMs can be useful tools for specialised translators, but cannot, at this stage, replace specialised corpora. This research therefore paves the way for future work on the real practical usefulness of LLMs for specialised translators in work and educational contexts.

1 Introduction

Specialised translation refers to the translation of texts that are specific to a particular field of knowledge, such as Earth sciences, computer science, law, etc. This type of translation, also known as pragmatic translation, has significant economic implications, as it is used in high-stakes industries.

The main challenges associated with specialised translation relate to aspects of Language for Specific Purposes (LSP). LSP is a language common to a group of specialists that serves the interests of that group (Bowker and Pearson, 2002; Basturkmen and Elder, 2004; Gollin-Kies et al., 2015; Gledhill and Kübler, 2016). Specialised translation–regardless of the domain–presents significant challenges, mainly in terms of terminology (Cabezas-García and León-Araúz, 2023). Each domain is characterised by its own conceptual network and, consequently, its own terminology (Scarpa, 2020). For example, the terms *fault creep*, *effusive volcano* and *very low-frequency earthquake* are part of the terminology of volcanology (and more broadly, Earth sciences). On the other hand, *neural network*, *large language model* and *text mining* refer to natural language processing and computer science. However, some terms may have a different meaning in different specialised languages, such as *cloud*, which does not refer to the same concept in Earth sciences as it does in computer science.

Consequently, specialised translation, even just in terms of bilingual terminology research, is a complex, cognitively demanding and time-consuming task for translators, whether they are still learners or already experienced professionals: "Pragmatically translating LSPs requires not only knowledge of the source and target cultures in general, but also knowledge of very specific areas. Even a very well-educated translator may not know the terminology, phraseology, or even grammar of a particular [specialised] domain" (Kübler, 2011)

When it comes to terminology (but also other aspects such as phraseology, collocations, discursive conventions, etc.), there are several tools available

to translators and translation learners. The simplest tools are arguably term bases and bilingual glossaries. However, these are far from exhaustive and never cover all the terms in a given field. To complement this method, specialised translators must turn to corpora. Studies on the use of corpora for translation and translation teaching began in the late 1990s, notably with Baker (1999) and Aston (1999). Specialised translation is particularly affected by the need to use corpora (Kübler, 2011; Kübler et al., 2018; Bernardini, 2022; Granger and Lefer, 2022). Corpora are useful for acquiring knowledge in these specialised domains, finding linguistic information, and searching for terminological equivalents (Kübler et al., 2018). Two main types of corpora can be compiled to assist specialised translation. Parallel corpora are an ideal resource for translators, as they contain texts aligned from the source language to the target language (Kübler and Aston, 2010; Kübler, 2011; Bernardini, 2022). This makes it relatively easy to perform searches in parallel corpora. However, parallel corpora are complex to compile, as they require bilingual parallel data, which is not easy to obtain for all language pairs and all domains (Corpas Pastor, 2007; Mezeg, 2020). Comparable corpora can also be useful for translators (Kübler et al., 2018; Kübler et al., 2024; Loock, 2016). These corpora are independent of each other (not aligned), but contain texts that deal with the same subject area. These corpora are much easier to collect, but searching for bilingual terminology in comparable corpora is more complicated, since data are not aligned. In the educational context of specialised translation training, the use of corpora has also become a central component of the curriculum. It is one of the skills required for translator training within the competence framework of the European Master's in Translation (EMT) 2022: "Students know how to: make effective use of search engines, corpus-based tools, text analysis tools, computer-assisted translation (CAT) and quality assurance (QA) tools where appropriate." (European Master's in Translation Network, 2022)

In this paper, we explore the use of a different class of tools that may partially replace established practices for terminological research in specialised translation. More specifically, we investigate the use of Large Language Models (LLMs) as an alternative to corpus-based approaches traditionally employed to identify translation equivalents. Given that LLMs are trained on massive amounts of textual data and exhibit emerging reasoning and generalisation capabilities, they may provide translators with direct access to plausible terminological translations without the need to manually compile and query corpora. This raises an important question for both professional translators and trainees: to what extent can LLMs reduce—or even eliminate—the need for corpus-based exploration when searching for specialised terminology, and what are the implications of such a shift for translation practices and training?

2 Related Work

Since 2022, the year in which ChatGPT, arguably the most popular GenAI model, was released (OpenAI, 2022), numerous studies have examined the use of LLMs for translation, with very encouraging results (Vilar et al., 2023; Hendy et al., 2023; Jiao et al., 2023; Wang et al., 2023). Some even believe that the future of machine translation will be closely linked (or is already linked) to the capabilities of LLMs (Lyu et al., 2024). Other studies have used LLMs for related tasks, including the evaluation and annotation of translations (both machine-translated and human-translated) (Kocmi and Federmann, 2023a; Kocmi and Federmann, 2023b; Lu et al., 2024; Fernandes et al., 2023; Moosa et al., 2024; Xu et al., 2023). With regard to our field of study, only a few studies have explored the potential of LLMs for specialised translation, particularly in the area of terminology.

Recently, Pecman (2025) looked at the integration of LLMs into terminology analysis, within the framework of the ARTES (Kübler and Pecman, 2011) term base (a terminology database for research and teaching in specialised translation). This work does not directly address specialised translation, but rather the drafting of definitions for emerging concepts (neologisms) or concepts undergoing semantic change. The study proposes an experimental protocol combining corpus linguistics and interaction with several GenAI tools (ChatGPT, DeepSeek, Perplexity, Claude, Gemini). Pecman (2025) shows that LLMs can be useful for reformulating, improving or structuring terminological definitions, particularly when guided by knowledge-rich contexts extracted from specialised corpora. However, her study emphasises that prior corpus-based analysis remains essential to ensure conceptual accuracy and avoid inaccura-

cies or approximations. The results highlight that LLMs are a relevant assistance tool, but that they cannot replace a rigorous methodology based on authentic data.

A few months prior to this study, an experiment on the annotation of specialised translations in an educational context with LLMs, in this case GPT-4o, was conducted (Minder et al., 2025). The researchers prompted the model to annotate errors in students' translations in the Earth sciences domain, based on an error typology adapted to the annotation of specialised translations, the MeLLANGE error typology (Castagnoli et al., 2011). This typology includes a wide range of terminological errors. After analysing the proportion of errors detected by the LLM, they observed that GPT-4o was able to identify approximately 65 % of the terminological errors (specifically 185 out of 289) contained in the Master's students' translations. This experiment opened up interesting perspectives on the use of LLMs for processing specialised terminology, which led us to conduct the present study.

These findings are particularly relevant to our study, as they already question the role of LLMs in the processing of specialised terminology. However, they focus on derivative tasks (drafting definitions of specialised terms, annotating translation and terminology errors). Our work builds upon this line of thinking by examining more specifically the search for English-French terminology equivalents in specialised translation and by comparing the performance of several LLMs on this task.

3 Background and Goals

Our experiment aims to assess the extent to which various GenAI tools can assist specialised translators performing tasks involving terminology research, primarily when searching for equivalents from English to French. This question is particularly relevant in specialised translation, where terminology accuracy plays a critical role in the quality, reliability and domain appropriateness of the target text.

While LLMs are increasingly explored and used for translation and MT evaluation tasks, their actual usefulness for more specific terminological tasks remains, to our knowledge, poorly documented, especially in LSP with a focus on specialised translation. Unlike the vast majority of studies on LLMs, which tend to focus on general language, our research is grounded in two highly specialised and contrasting fields: Earth, Environmental and Planetary Sciences (EEPS) and Natural Language Processing (NLP). These two domains were intentionally selected to test the performance of LLMs across diverse and varied knowledge areas. By comparing the results across these two domains, we hope to determine whether the models tested here remain stable or vary depending on the domain. Beyond a simple comparison of models' performances, the ultimate goal is more practice-oriented. We aim to explore whether LLMs can realistically serve as assistants for specialised translators in terminology research, or even whether LLMs have the potential to replace corpora.

4 Methods

4.1 Models

In order to assess the potential usefulness of LLMs for specialised terminology research, we examine the performance of four models in the search for equivalents from English into French: GPT-4o, GPT-5.2, Claude Sonnet 4.5 and DeepSeek. We selected these models for pragmatic reasons. As this experiment aims to assess the potential usefulness of LLMs for finding terminological equivalents, we intend to test the tools that are exploited by users. These proprietary models are used in practice by professional translators and translation learners. We test the models on 80 terms per domain, for a total of 160 terms per model.

4.2 Term selection

The first step was to select 80 terms per domain (EEPS and NLP), for a total of 160 terms. To do so, we manually selected terms from several specialised resources, namely texts (research articles) translated by Master's students in specialised translation, a terminology database, ARTES[1], that is fed annually by, among others, Master's students, and other specialised texts that we had used for other experiments on specialised translation (references will be provided once the submission can be de-anonymised). We ensured that terms of different structures and varying levels of complexity were represented in our sample: simple terms as well as compound and complex terms. The 80 terms are identical for all models and prompts. The selected terms are detailed in Figures 8 and 9 in the appendices; the responses generated by each LLM

[1] https://artes.app.univ-paris-diderot.fr/

Scenario	English term	Main equivalent	Secondary equivalent	Rare equivalent
Correct Main Equivalent (CME)	*low-resource language*	langue peu dotée	langue à faibles ressources	langue pauvre en ressources
Main Equivalent Identified (MEI)	*cross-entropy loss*	entropie croisée	perte d'entropie croisée	/
Main Equivalent Not Identified (MENI)	*megathrust earthquake*	séisme de mégaséquence	méga-séisme de subduction	/

Figure 1: Suggestions from LLM – examples for each case: main equivalent correctly identified as main equivalent by the LLM (first example); main equivalent identified as a secondary equivalent by the LLM (second example); main equivalent not identified by the LLM (third example). A green cell indicates that the equivalent is considered appropriate by the professional translator. In the third example, the model failed to identify the correct French equivalent; the French term is *séisme de mégachevauchement* (ou *séisme de méga-chevauchement*).

for each domain and mode, along with the annotations for these responses, will be made available upon publication.

4.3 Prompts and modes

We test two different prompts for each model and each domain. These prompts correspond to different modes. The first prompt is what we call 'terminology mode': it simply instructs the model to search for terminological equivalents, and for each term, we provide in the prompt a context sentence (in English) containing the term. In the second mode, 'translation mode', we first request the model to translate the context sentence provided, then to list the identified terminology equivalents (Figures 10 and 11 in the appendices illustrate the specific differences between terminology mode and translation mode). A total of 16 individual experiments were conducted (4 models, 2 domains and 2 modes). In both prompts, we also request LLMs to indicate their level of confidence for each term processed. In addition, we request LLMs to sort the equivalents found into three categories: main equivalent, and, where applicable, secondary equivalents and rare equivalents. The criteria we give to the models are: "(a) main equivalent – attested (equivalent found in one or more terminology databases) and/or more frequent (frequently found in similar contexts in corpora, in texts of the same register and type); (b) secondary equivalents – less frequent in terminology databases and corpora, but still used; (c) rare equivalents (found only occasionally and with statistically insignificant occurrences). If there is only one equivalent, list it as the main equivalent. There do not necessarily have to be multiple equivalents" (excerpt from the prompt). Two examples have also been added to the prompt to show the desired output presentation.

In addition, we introduced a last variant (referred to as 'documentary justification' in the following) on the best-performing model. On this model, we tested the inclusion of a constraint: we asked the model, for each equivalent it identified, to provide a source (scientific, which could be an article, a book chapter, a conference paper, etc.) containing that equivalent and to give a sentence illustrating the use of the equivalent. We explicitly instructed the model not to invent any sources or sentences. Our idea was to see whether forcing the model to identify a source would yield more accurate and precise results. The full prompts for both modes and for the documentary justification are detailed in Figures 10 and 11 in the appendices.

4.4 Performance, annotation and evaluation

For each term, one professional translator with extensive expertise in corpus linguistics and translation annotation in both domains analyses whether the LLM is able to provide a correct French equivalent. This allows us to precisely quantify the proportion of equivalents correctly identified by each model, and to compare the impact of the prompting strategy and the domain of specialisation.

For performance evaluation, we first classified each equivalent provided by the model as follows. Each term was assigned one of the possible outcomes: the main equivalent was correctly identified as the main equivalent, the main equivalent was identified but classified as a secondary or rare equivalent, or the main equivalent was not identified (see Figure 1 for an example of each case). These three outcomes were weighted differently

in the scoring procedure: full credit was assigned when the main equivalent was correctly identified as the main equivalent, partial credit when it was identified but not ranked as the main equivalent, and no credit when it was not identified. The final score for each model, mode and domain was then obtained by averaging these weighted outcomes across all terms and normalising the result on a 0 to 1 scale.

4.5 Verification resources

In order to check LLM answers, we use several resources: comparable and parallel corpora (English-French) compiled and enriched over the years in our research lab as part of various projects, specialised terminology databases such as TERMIUM[2], and a terminology database developed in our research lab. (We will provide proper references for these resources when the authors' identities may be disclosed). All answers, for the 80 terms per domain, each model and each prompting method, were annotated manually.

5 Results

We analysed the models' performance across different modes and domains according to several statistical indicators.

5.1 Overall performance

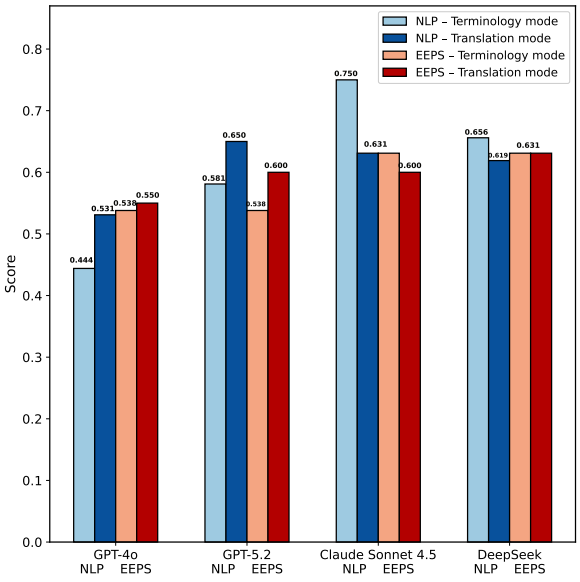

Figure 2: Results achieved on the four models, by mode (terminology or translation) and by domain (EEPS and NLP).

[2]https://www.btb.termiumplus.gc.ca/.

Figure 2 shows that, overall, there are substantial variations in results across models, modes and, to a lesser extent, domains. Of the four LLMs tested, Claude Sonnet 4.5 achieves the highest average score, with peak performance in terminology mode in the NLP domain. However, Claude Sonnet 4.5 shows greater variation across domains and modes than the other models. DeepSeek performs satisfactorily and shows stable results across modes and domains. GPT-5.2 achieves intermediate scores, and GPT-4o remains the weakest model overall for this experiment.

The effect of the prompting strategy (terminology vs translation mode) is not consistent across models. GPT models seem to benefit from translation mode, with a systematic improvement from terminology to translation mode. On the other hand, Claude Sonnet 4.5 performs better in terminology mode, particularly in NLP. DeepSeek shows only very limited variation depending on the mode, suggesting potential robustness to the prompting strategy.

The effects of domain are also noticeable, but less consistent than the effects of mode and prompt. Some models perform slightly better in NLP (GPT-5.2 and Claude Sonnet 4.5), while GPT-4o performs slightly better in STEP. DeepSeek, on the other hand, remains fairly stable across both domains. This suggests that the domain does influence model performance, but that this effect interacts with other conditions: the model and the prompting strategy. DeepSeek, while not achieving the best scores, is the most stable model to changes in mode and domain. Although the difference in performance across different domains appears to be rather minimal, this may be useful for specialised translators: selecting a specific model based on the domain covered may yield better results.

These results do not allow for the conclusion that LLMs can replace corpus-based searches by specialised translators. However, the results are sufficiently positive to suggest that LLMs can indeed provide useful support in the search for equivalents, with some models performing slightly better in one area than in another, in translation mode or terminology mode.

5.2 Confidence estimates

Figure 3 shows the distribution of confidence estimates reported by each model, compared in

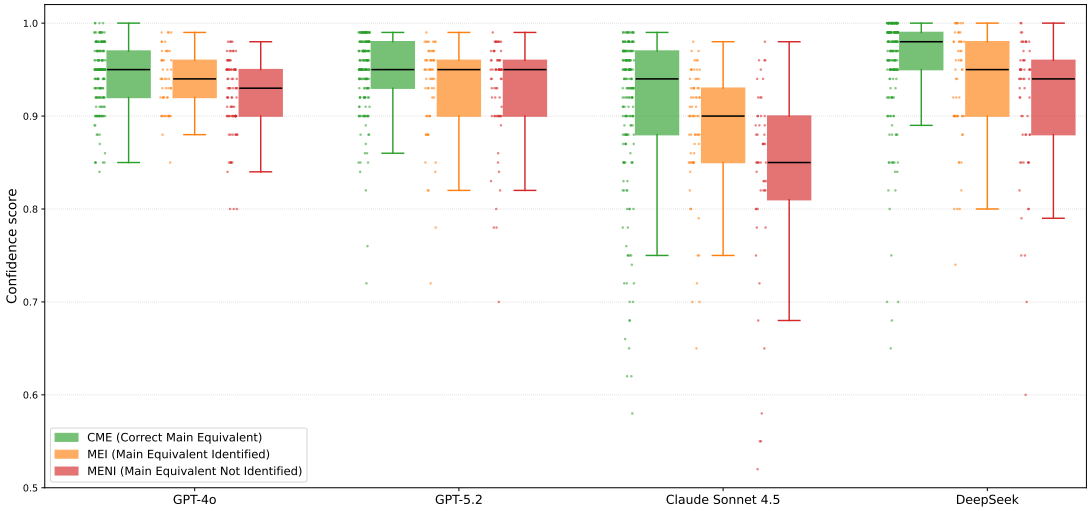

Figure 3: Distribution of confidence estimates for each model, based on the accuracy of the identified equivalents (green: the main equivalent has been identified as the main equivalent; orange: the main equivalent has been identified as a secondary or rare equivalent; red: the main equivalent has not been identified).

terms of the accuracy of equivalents (identified (green), partially identified (yellow) and unidentified (red)). Overall, confidence estimates remain high across all three categories. However, there is a general trend across all four models towards higher average confidence estimates when the main equivalent is correctly identified. In contrast, partially identified and unidentified equivalents tend to be associated with lower average confidence estimates. This trend is particularly noticeable in Claude Sonnet 4.5, where the distribution of estimates becomes progressively weaker and more dispersed as accuracy drops. In other words, when Claude Sonnet 4.5 fails to correctly identify the main equivalent, it reports a generally lower confidence estimate than the other models. GPT-4o, GPT-5.2 and, to a lesser extent, DeepSeek, however, tend to remain confident in all three scenarios. Consequently, for these models, confidence estimates are less clearly aligned with terminological accuracy. From a practical perspective, these results show that confidence estimates are only partially informative as indicators of terminological accuracy.

Figure 4 shows that all four LLMs report higher confidence levels in the EEPS domain, although confidence scores also remain relatively high in NLP. Claude Sonnet 4.5 has a more distinct profile: it has a significantly more dispersed and weaker distribution in NLP, showing that the NLP domain is associated with the presence of many cases of lower confidence. However, Figure 2 shows that Claude Sonnet 4.5 is one of the models with the best results in NLP, particularly in terminology mode, further demonstrating that confidence estimates do not necessarily correlate with accuracy.

From the user's perspective, these results show that confidence estimates alone do not constitute a reliable indicator of terminological accuracy. While there is a slight overall trend–correctly identified equivalents are, on average, accompanied by slightly higher confidence estimates–the significant overlap between the three categories shows that high estimates can also be associated with partially identified, or even unidentified, equivalents. In other words, a model may be highly confident while proposing an inadequate equivalent. From a practical perspective, this means that translators cannot rely on the confidence estimate alone to judge the probability that a suggested equivalent is correct. At best, this estimate can serve as a secondary indicator, but it cannot replace verification in corpora, term bases or other resources.

5.3 Documentary justification

Based on the previous results, we considered Claude Sonnet 4.5 to be the best-performing model according to several variables: good overall performance in both domains (Figure 2) and a stronger correlation between confidence estimates and accuracy of equivalents (Figure 3). Consequently, we tested the addition of an instruction in the prompt for this model: providing a scientific source and a sentence containing the term from that source. The purpose was to determine whether forcing the model to rely on concrete references would yield

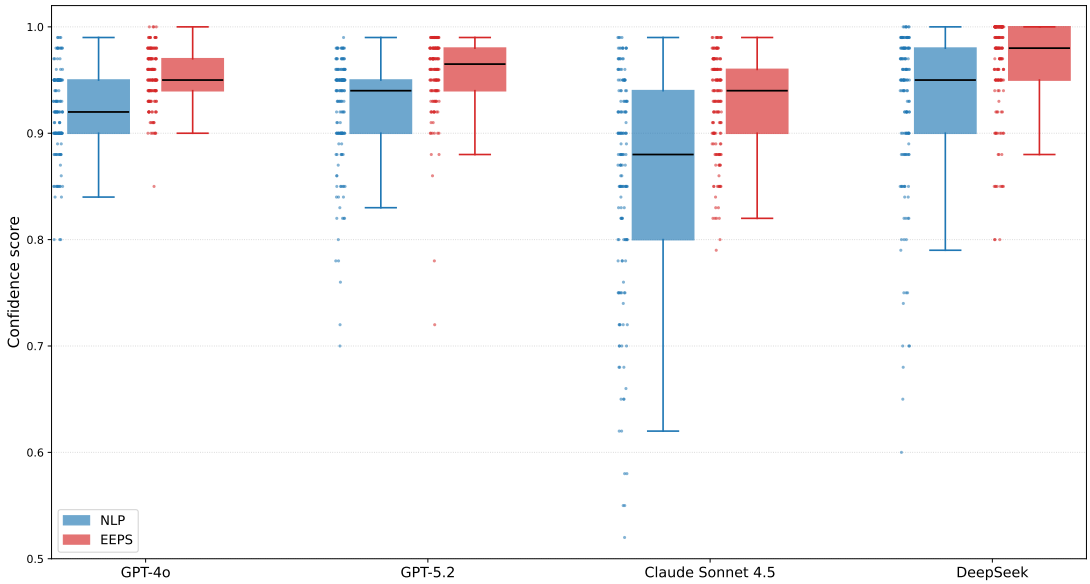

Figure 4: Distribution of confidence estimates for each model, based on the domains (NLP in blue vs EEPS in red).

better results. We only tested this in terminology mode, as this is the mode in which Claude Sonnet 4.5 achieves the best results.

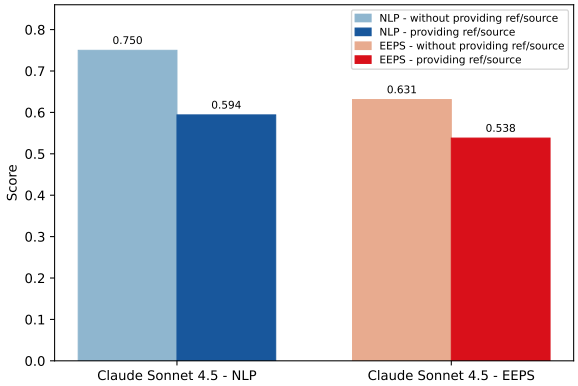

Figure 5: Comparison of the performance of Claude Sonnet 4.5 in terminology mode, with and without the specification of a scientific reference and a sentence illustrating the term.

Figure 5 shows the results obtained with Claude Sonnet 4.5 in terminology mode, with and without the addition of an instruction requiring the model to provide, for each equivalent identified, a scientific source and an example sentence from that source. In both domains, the addition of this constraint leads to a decrease in scores. The decrease is therefore visible in both domains, and particularly marked in NLP. This approach to documentary justification poses an additional issue. Although we explicitly instructed the model not to invent any references or sentences illustrating the term, this instruction was not observed. In most cases, the sources provided by the model do not exist: they are either completely fabricated or are a translated source (for example, the source does not exist in French as provided by Claude Sonnet 4.5, but can be found in English). Furthermore, in every single case without exception, the sentence given as a reference by the model is made up, meaning that, upon verification, we cannot find it anywhere. These results show that, in the framework of this experiment, adding a requirement to reference and justify with a source does not improve the model's performance in finding equivalents. On the contrary, this additional constraint seems to be associated with a decrease in the model's ability to correctly identify the main equivalent. However, it is important to provide some insight into this observation. This drop in performance may not necessarily be (solely) due to the addition of the new instruction. Indeed, LLMs' performances are unpredictable: it has been observed that the performance of models varies between different iterations of the same model (Siu, 2023). It is therefore not directly possible to claim that this decline in performance is the result of the addition of the documentary justification instruction.

5.4 Qualitative analysis

Beyond quantitative and statistical analysis, it is also interesting to observe in concrete terms how the equivalents provided by LLMs, even when incorrect, can potentially guide a translator towards the right solution. Other examples show, on the

contrary, that there are cases—albeit rare—where LLMs produce completely nonsensical or even absurd results.

	English term	Main equivalent	Secondary equivalent	Rare equivalent
(1)	*mineral assemblage*	*assemblage minéral*	*paragenèse minérale, association minérale*	*assemblage de minéraux*
(2)	*melt impregnation*	*imprégnation par le liquide*	*imprégnation par le magma*	/
(3)	*isotope systematics*	*systématique isotopique*	*systématiques isotopiques*	/
(4)	*inductively coupled plasma mass spectrometry*	*spectrométrie de masse à plasma induit*	/	/

Figure 6: Examples of equivalent search results by LLMs that are inaccurate but may still prove useful to a specialised translators. The slash (/) in 'Secondary equivalent' and 'Rare equivalent' indicates that the LLM did not identify any equivalents of this type, and that it identified only a main equivalent, with a secondary equivalent where applicable.

The examples illustrated in the Figure 6 show instances where the equivalents provided by LLMs are inaccurate, but where a simple corpus or term base search based on the equivalents suggested by the LLM quickly provides an accurate solution. The first example (*mineral assemblage*) is the most straightforward: a simple search for *assemblage* associated with the stem *minéral* shows that the most common term is *assemblage minéralogique*, i.e. the same head associated with an adjective that has the same stem as *minéral*. Here, the LLM does not provide the correct equivalent, but it does provide clear clues pointing to the correct solution.

Lines 2 and 3 are similar in terms of approach: they are both examples of confusion between attributive adjectives and noun complements (compléments du nom) in French. In French, modifiers can take several forms: either an attributive adjective (an adjective directly linked to the noun, without a preposition), or a noun complement (a modifier separated from the noun by a preposition). Here, *imprégnation par le magma* and *systématique isotopique* fall into this category: a quick corpus query for the head nouns (*imprégnation* and *systématique*) combined with the stems *magma* and *isotop** provides the answer. This reveals the most frequently used terms: *imprégnation magmatique* and *systématique des isotopes*.

Example 4 is even more straightforward. It is sufficient to search for *spectrométrie de masse* (confirmed in term bases, including Termium) in association with *plasma*, and this should lead to the solution: *spectrométrie de masse à plasma à couplage inductif*.

These examples demonstrate that even when the LLM does not provide the appropriate equivalents, the suggestions can guide a translator in conducting the correct searches (in corpora or term bases). However, this depends on several factors, including experience in the domain (particularly the intuition that can be acquired with experience) and with corpus exploitation tools. Therefore, it should not be assumed that these examples would be useful for any given translator. Only practical experiments in real-world contexts would allow us to assess the usefulness of LLM suggestions.

English term	Context excerpt	Main equivalent	Secondary equivalent	Rare equivalent
aligner	*"... which are very useful for evaluating sentence aligners..."*	*aligner*	*apparier*	/
volatile cycling	*"... behavior of the lithosphere and volatile cycling in the Earth..."*	*recyclage des éléments volatils*	/	/
treebank	*"... successful probabilistic parsers require large treebanks..."*	*arboré syntaxique*	/	/

Figure 7: Examples of LLM hallucinations when searching for equivalents. The slash (/) in 'Secondary equivalent' and 'Rare equivalent' indicates that the LLM did not identify any equivalents of this type, and that it identified only a main equivalent, with a secondary equivalent where applicable.

Figure 7 shows striking examples of hallucinations produced by LLMs in the search for equivalents. The first term, *aligner*, which is a noun referring to an alignment tool, is associated with a verbal equivalent, i.e. the verbs *aligner* (to align) and *apparier* (to pair up) in French; the issue here is that the LLM did not accurately identify the part-of-speech of the term, despite the context sentence clearly illustrating the noun. The correct equivalent here would be *aligneur* or *outil d'alignement* In the second example, the term *cycling* was translated as *recyclage* (recycling) in French, which does not correspond at all to the meaning of the English term. A correct French equivalent for this term would be *cycle des (éléments) volatils* ou *cycle des espèces volatiles*. In the last example, the noun has been translated as an adjectival phrase without a head noun, which makes no sense. The correct equivalent for this term is *corpus arboré*. These three examples show that, although in most cases LLMs identify the correct equivalents or pro-

vide good starting points for the search of the correct equivalent, they can still produce completely disconnected or absurd outputs.

6 Discussion

The results show that LLMs can prove useful tools for specialised terminology research, but that, within the current scope of this experiment, they cannot replace specialised corpora. Even in the best configurations, performance remains limited and uneven depending on the model, mode and domain. These results therefore argue in favour of using LLMs as a complementary tool in specialised translation workflows rather than as a direct substitute for comparable, parallel corpora or terminology databases. LLMs can help to quickly generate potential equivalents, but the verification, validation and contextualisation of terms remain the responsibility of external resources and human expertise.

The effect of the model appears to be the most decisive factor. Claude Sonnet 4.5 achieves the best overall results in the most favourable configuration (terminology mode), while DeepSeek stands out for its greater stability between modes and domains. GPT-5.2 occupies an intermediate position and GPT-4o lags behind overall. In addition, the effect of the prompting strategy is not consistent: GPT models seem to benefit from the translation mode, while Claude Sonnet 4.5 performs better in terminology mode. This suggests that there is no universally ideal prompting strategy for terminological equivalence search, and that effectiveness largely depends on the model being queried.

Analysis of confidence estimates also shows that these are only a partial indicator of terminological accuracy. While Claude Sonnet 4.5 seems to adjust its confidence better when the actual quality of the equivalent decreases, LLMs often maintain high confidence levels, even when the main equivalent is partially identified or unidentified. This highlights possible overconfidence and suggests that self-reported confidence should not be considered a reliable indicator of terminological accuracy. The domain effect also comes into play, but remains moderate and inconsistent. Confidence levels appear to be slightly higher and more stable in EEPS than in NLP for several models, although this trend cannot be generalised to all cases.

Finally, the complementary experiment conducted on Claude Sonnet 4.5 shows that adding a requirement to provide a scientific source and an example sentence does not improve performance, but rather degrades it in both NLP and EEPS. This finding suggests that an additional constraint of documentary justification does not guarantee better terminological quality and may even distract the model from the main task.

7 Conclusion and Future Work

Specialised translation relies heavily on the use of documentation and terminology resources, among which corpora play a central role. However, compiling and exploiting them has significant drawbacks: it requires time, specific technical skills and access to (parallel) data that can sometimes be difficult to obtain, particularly for certain domains and language pairs. It is in this context that this study sought to assess the extent to which proprietary LLMs available to general, non-technical users can assist specialised terminology research.

Based on four models tested on a task of finding English-French equivalents in two specialised domains, the results show that LLMs have real potential for assistance, but that their performance varies depending on the model, the prompting strategy and, to a lesser extent, the domain. They also show that the confidence estimates provided by the models are only a partial indicator of terminological accuracy. Overall, our results therefore support the idea of LLMs playing a complementary role in specialised translation workflows rather than directly replacing specialised corpora.

This study does, however, have several limitations. It focuses on two specialised fields and a single language pair, English-French, and on a specific experimental task of finding equivalents. In future work, it could be useful to extend the analysis to other fields, other languages and other prompting configurations. Another limitation of this study is that it is based on annotations by a single expert, with no measurement of inter-annotator agreement between different expert annotators. This is due to the highly specialised, complex and time-consuming nature of this task. In the future, we aim to involve several expert annotators and incorporate IAA measurements, either across the entire dataset or a subset of the data. This study also lacks testing to determine whether using LLMs as a support tool can reduce time and cognitive load for specialised translators, as opposed to relying, as usual, on corpora and termi-

nology databases. The qualitative analysis of LLM outputs has been undertaken in this study, but in future work, we intend to assess in real-world settings (such as educational contexts involving translation learners) how these outputs can be exploited and how they might complement or be combined with corpus-based alternatives. Most importantly, it would be relevant to go beyond task evaluation to examine more practically the usefulness and usability of these tools for specialised translators, particularly in an educational context, for example with Master's students in specialised translation. Such investigations would make it possible to observe more concretely their effect on the time spent searching for terminology, the quality of the choices made, the verification strategies implemented and, more broadly, their place in translation and training practices.

Acknowledgments

This research was funded by the French Agence Nationale de la Recherche (ANR) under the project MaTOS - "ANR-22-CE23-0033-03".

References

Aston, Guy. 1999. Corpus use and learning to translate. In *Textus*.

Baker, Mona. 1999. The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics*, 4:281–298.

Basturkmen, Helen and Catherine Elder. 2004. The practice of lsp. In Davies, Alan and Catherine Elder, editors, *The Handbook of Applied Linguistics*, chapter 27, pages 672–694. John Wiley & Sons.

Bernardini, Silvia. 2022. How to use corpora for translation. In *The Routledge Handbook of Corpus Linguistics*, page 14. Routledge, 2 edition.

Bowker, Lynne and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge, 09.

Cabezas-García, Melania and Pilar León-Araúz. 2023. Machine versus corpus-based translation of multiword terms. *Digital Scholarship in the Humanities*, 38(Supplement 1):6–16, 06.

Castagnoli, Sara, Dragoș Ciobanu, Kerstin Kunz, Natalie Kübler, and Alexandra Volanschi. 2011. Designing a learner translator corpus for training purposes. In Kübler, Natalie, editor, *Corpora, Language, Teaching, and Resources: From Theory to Practice*, Études contrastives, pages 221–248. Peter Lang, Bern.

Corpas Pastor, Gloria. 2007. Lost in specialised translation: the corpus as an inexpensive and underexploited aid for language service providers. In *Proceedings of Translating and the Computer 29*, London, UK, November 29-30. Aslib.

European Master's in Translation Network. 2022. Emt competence framework 2022.

Fernandes, Patrick, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore, December. Association for Computational Linguistics.

Gledhill, Christopher and Natalie Kübler. 2016. What can linguistic approaches bring to English for Specific Purposes? *ASp - La revue du GERAS*, (69):65–95, March.

Gollin-Kies, Sandra, {David R.} Hall, and {Stephen H.} Moore. 2015. *Language for specific purposes*. Palgrave Macmillan, United Kingdom.

Granger, Sylviane and Marie-Aude Lefer, editors. 2022. *Extending the Scope of Corpus-Based Translation Studies*. Bloomsbury Advances in Translation. Bloomsbury Academic, London.

Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Jiao, Wenxiang, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Kocmi, Tom and Christian Federmann. 2023a. Gemba-mqm: Detecting translation quality error spans with gpt-4.

Kocmi, Tom and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality.

Kübler, Natalie and Mojca Pecman. 2011. ARTES: an online lexical database for research and teaching in specialized translation and communication. In *ESSLLI 2011, International Workshop on Lexical Resources (WoLeR)*, Ljubljana, Slovenia, August.

Kübler, Natalie. 2011. Working with different corpora in translation teaching. In Frankenberg-Garcia, Ana, Lynne Flowerdew, , and Guy Aston, editors, *New Trends in Corpora and Language Learning*, pages 62–80. Continuum.

Kübler, Natalie and Guy Aston. 2010. Using corpora in translation. In O'Keeffe, Anne and Michael McCarthy, editors, *The Routledge Handbook of Corpus Linguistics*. Routledge, London, 1 edition.

Kübler, Natalie, Alexandra Mestivier, and Mojca Pecman. 2018. Teaching specialised translation through corpus linguistics: Translation quality assessment and methodology evaluation and enhancement by experimental approach. *Meta*, 63(3):807–825.

Kübler, Natalie, Hanna Martikainen, Alexandra Mestivier, and Mojca Pecman, 2024. *Chapter 4. Post-editing neural machine translation in specialised languages: The role of corpora in the translation of phraseological structures*, pages 57–78. John Benjamins Publishing Company.

Loock, Rudy. 2016. *La Traductologie de Corpus*. Presses Universitaires du Septentrion.

Lu, Qingyu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models.

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

Mezeg, Adriana. 2020. Parallel corpora vs bilingual dictionaries: Their usefulness in translator training. In Granger, Sylviane and Marie-Aude Lefer, editors, *Translating and Comparing Languages: Corpus-based Insights*, volume 6 of *Corpora and Language in Use Proceedings*, pages 123–140. Presses universitaires de Louvain, Louvain-la-Neuve.

Minder, Joachim, Guillaume Wisniewski, and Natalie Kübler. 2025. Testing LLMs' capabilities in annotating translations based on an error typology designed for LSP translation: First experiments with ChatGPT. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 190–203, Geneva, Switzerland, June. European Association for Machine Translation.

Moosa, Ibraheem Muhammad, Rui Zhang, and Wenpeng Yin. 2024. Mt-ranker: Reference-free machine translation evaluation by inter-system ranking.

OpenAI. 2022. Introducing chatgpt.

Pecman, Mojca. 2025. Drafting definitions for emerging concepts and terms undergoing semantic shift within the artes knowledge base: A protocol for integrating llms into terminological analysis by experimental approach. *Terminologija*, 12.

Scarpa, Federica, 2020. *Introducing Specialised Translation*, pages 1–109. Palgrave Macmillan UK, 09.

Siu, Sai Cheong. 2023. Chatgpt and gpt-4 for professional translators: Exploring the potential of large language models in translation, May. Available at SSRN.

Vilar, David, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance.

Wang, Longyue, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models.

Xu, Wenda, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. Instructscore: Explainable text generation evaluation with finegrained feedback.

8 Appendices

8.1 Terms selected (NLP)

run	aligner	log file
treebank	MT output	text span
fine-tuning	open source	scalability
fine-grained	ground truth	post-editing
priming text	context-aware	crowdsourcing
to outperform	training loss	disambiguation
opinion mining	word embedding	baseline model
neural network	input sequence	inductive bias
computerization	multi-word term	parametrization
dependency tree	ontology mapping	back-translation
attention weight	back propagation	annotation scheme
syntactic pattern	transfer learning	constituency tree
diarization system	k-means clustering	textual entailment
vector space model	chart-based parser	speech recognition
cross-entropy loss	multi-task learning	clustering algorithm
image quality metric	constituency parsing	syntactic dependency
keyphrase generation	adversarial training	summarization model
contextual embedding	information retrieval	knowledge engineering
low-resource language	prosodic highlighting	variational inference
commonsense reasoning	Gaussian mixture model	part-of-speech tagging
reinforcement learning	parallel training data	multivariate time series
named entity recognition	self-supervised learning	masked language modeling
Transformer architecture	discriminative classifier	human-machine interaction
term mismatch probability	computer-aided translation	deep convolutional network
multi-label classification	task-specific output layer	natural language generation
stochastic gradient descent	principal component analysis	state-of-the-art performance
compositional generalization		word-level confidence estimation

Figure 8: 80 terms selected for the Natural Language Processing domain.

8.2 Terms selected (EEPS)

geofluid	xenocryst	flowstone
Bragg peak	ground ice	mantle rock
radiotracer	bioavailable	wet sediment
pore pressure	decarbonation	trace element
igneous crust	solidus curve	forearc region
midocean ridge	melt inclusion	redox reaction
injection well	laser ablation	field sampling
soil aggregate	subduction zone	caprock leakage
megasplay fault	hotspot volcano	upwelling plume
transform fault	aromatic moiety	ultramafic rock
fluid chemistry	plume migration	seismic imaging
fluid inclusion	ice core record	melting behavior
fluid entrapment	volatile cycling	subducting plate
diffraction line	lowermost mantle	Raman scattering
drainage pattern	inflection point	thermal response
magma supply rate	melt-impregnation	deep carbon cycle
X-ray diffraction	superdeep diamond	slab-top geotherm
equation of state	deviatoric stress	nanoscale porosity
hydrothermal fluid	mineral assemblage	diamond anvil cell
breakdown reaction	isotope systematics	carbonate reduction
water contamination	rare earth elements	oxidizing condition
accretionary complex	carbon concentration	carbon-bearing phase
megathrust earthquake	atom probe tomography	convergent plate margin
short-chain hydrocarbons	bulk isotopic composition	grain size-sensitive creep
carbon capture and storage	terminal electron acceptor	wastewater treatment plant
clumped isotope thermometry	transmitted light microscopy	sub-cratonic lithospheric mantle
carbon isotope chemostratigraphy		inductively coupled plasma mass spectrometry

Figure 9: 80 terms selected for the Earth, Environmental and Planetary Sciences domain.

8.3 Prompt: terminology mode

You are a terminologist specialising in the field of Natural Language Processing (NLP)/ Earth, Environmental and Planetary Sciences (EEPS).

Your goal is to find French equivalents for English terms based on their frequency in NLP/EEPS specialised language corpora and their attested translations in glossaries or terminology databases.

Your task is to find one or more French equivalents for each English term in the same specialised NLP language and register (scientific articles), and to sort them by frequency and attestation.

For each term you process, assess your level of confidence. The confidence level, between 0 and 1, is determined according to the following criteria: whether or not the term is used frequently in similar contexts, whether or not the term is attested in one or more relevant term bases, or whether the translation of the term was not found in corpora or term bases and was therefore created.

[For each equivalent term you identify, you must also find a sentence in which that term is used. This sentence must come from an article of the same type and in the same domain. Also provide the full source of the article from which the sentence is taken. Be careful not to invent sentences or references!]

Example 1:

Term: *to annotate*

Context: *A pair was annotated noisy only in the case of serious problems; sentences with single translation errors or relatively poor fluency were still considered clean.*

Equivalents(s):

a) Main equivalent: *annoter*

[Identified context: *Ces trois constats mènent les auteurs à considérer la tâche de QA-SRL (Question Answer Driven Semantic Role Labelling), qui est une tâche similaire au QA, mais où on présente à des annotateurs des relations et des phrases et ceux-ci doivent annoter certaines informations relatives à cette relation.*]

[Reference: Lamarche, Fabrice. (2023). Méthodes d'évaluation en extraction d'information ouverte. Université de Montréal (thèse).]

b) Secondary equivalent: /
c) Rare equivalent: /

Confidence: 0.98

Example 2:

Term: *large language model*

Context: *In recent years, Natural Language Processing (NLP) systems have gotten increasingly better at learning complex patterns in language by pretraining large language models like BERT, GPT-2, and CTRL.*

Equivalent(s):

a) Main equivalent: *grand modèle de langue*
b) Secondary equivalent: *grand modèle de langage*
c) Rare equivalent: *giga modèle de langue*

Confidence: 0.82

Here are the terms to be processed, accompanied by a context. Sort the equivalents you identify according to this taxonomy: (a) main equivalent – attested (equivalent found in one or more term bases) and/or more frequent (frequently found in similar contexts in corpora, in texts of the same register and text type); (b) secondary equivalents – less frequent in terminology databases and corpora, but still used; (c) rare equivalents (found on an ad hoc basis and with insignificant occurrences). If there is only one equivalent, place it as the main equivalent. There are not necessarily several equivalents.

[…]

Figure 10: Prompt used for terminology research on LLMs in "terminology" mode. The segments in square brackets and underlined were added only for the additional documentary justification test on Claude Sonnet 4.5.

8.4 Prompt: translation mode

You are a terminologist and translator specialising in the field of Natural Language Processing (NLP)/ Earth, Environmental and Planetary Sciences (EEPS).

Your goal is to find French equivalents for English terms based on their frequency in NLP/EEPS specialised language corpora and their attested translations in glossaries or terminology databases.

Your task is to find one or more French equivalents for each English term in the same specialised NLP language and register (scientific articles), and to sort them by frequency and attestation.

For each term you process, assess your level of confidence. The confidence level, between 0 and 1, is determined according to the following criteria: whether or not the term is used frequently in similar contexts, whether or not the term is attested in one or more relevant term bases, or whether the translation of the term was not found in corpora or term bases and was therefore created.

You must translate the context sentence by inserting the equivalent term you have identified. If there are several possible equivalents, insert the main equivalent in the translation and note all possible equivalents below the translation, as in the examples below.

Example 1:

Term: *to annotate*

Context: *A pair was annotated noisy only in the case of serious problems; sentences with single translation errors or relatively poor fluency were still considered clean.*

Translation: *Une paire a été annotée comme bruitée uniquement en cas de problèmes graves ; les phrases comportant une seule erreur de traduction ou dont la fluidité était relativement mauvaise ont été considérées comme propres.*

Equivalents(s):

a) Main equivalent: *annoter*
b) Secondary equivalent: /
c) Rare equivalent: *étiqueter*

Confidence: 0.98

Example 2:

Term: *large language model*

Context: *In recent years, Natural Language Processing (NLP) systems have gotten increasingly better at learning complex patterns in language by pretraining large language models like BERT, GPT-2, and CTRL.*

Translation: *Ces dernières années, les systèmes de traitement automatique des langues (TAL) ont fait d'énormes progrès dans l'apprentissage des patrons linguistiques complexes grâce au pré-entraînement de grands modèles de langue tels que BERT, GPT-2 et CTRL.*

Equivalent(s):

a) Main equivalent: *grand modèle de langue*
b) Secondary equivalent: *grand modèle de langage*
c) Rare equivalent: *giga modèle de langue*

Confidence: 0.82

Here are the terms to be processed, accompanied by a context. Sort the equivalents you identify according to this taxonomy: (a) main equivalent – attested (equivalent found in one or more term bases) and/or more frequent (frequently found in similar contexts in corpora, in texts of the same register and text type); (b) secondary equivalents – less frequent in terminology databases and corpora, but still used; (c) rare equivalents (found on an ad hoc basis and with insignificant occurrences). If there is only one equivalent, place it as the main equivalent. There are not necessarily several equivalents.

[…]

Figure 11: Prompt used for terminology research on LLMs in "translation" mode. The underlined segments are those that have been added for 'translation' mode compared to 'terminology' mode.