Exploring the power of generative artificial intelligence for automatic term extraction from small samples

Lena De Pourcq, Marie Grégoire, Leonardo Zilio

UCLouvain, Louvain-la-Neuve, Belgium
E-mail: {lena.depourcq,marie.gregoire}@student.uclouvain.be, leonardo.zilio@uclouvain.be

Abstract

This study explores the use of several chatbots based on recent generative large language models for automatic term extraction (ATE) from smaller text samples. The samples were selected from three domains: board games, ice hockey, and kitesurfing; and they cover three languages: English, French, and Portuguese. We used four prompting strategies: zero shot, one shot, few shots, and few shots with context. A single prompt with placeholders for language, domain and examples (when available) was used for all settings, and, in the case of French and Portuguese, we tested the ATE prompt in English and in the respective language. Results were calculated in terms of f-measure, and we further tested the best models with five consecutive runs to calculate a mean f-measure and a standard deviation. No clear best system was verified for the task. Each of the domains and languages had different best systems. In terms of prompting strategy, more information did not always lead to better results, as zero-shot and one-shot attempts had the best results in several scenarios. The main contribution of the study is an overview of the ATE capacity of several chatbot systems across multiple scenarios.

Keywords: automatic term extraction; ATE; chatbots; generative artificial intelligence; GenAI

1. Introduction

Automatic term extraction (ATE) is an important task in the translation and localisation industry (Giguere, 2023), and it is well-researched in the context of larger monolingual and bilingual corpora (cf. Pazienza et al., 2005; Costa et al., 2016; Oliver, 2017; Di Nunzio et al., 2023). However, for translators, having to generate *ad hoc* glossaries based on smaller text samples is a much more common task (Šajatović et al., 2019). In this context, chatbots powered by generative artificial intelligence (GenAI) offer a flexible alternative for extracting specialised information from a document. It is however known that web-based GenAI chatbots can produce variable answers, and the current inability of the user to properly change some important settings in these models makes it impossible to have less variability in some of these systems' replies.

In this context, the main aim of this study was to test six freely available online GenAI chatbots and evaluate the quality of their outputs when used to perform ATE with a series of pre-set prompts across three different languages and domains. As evaluation dataset, we used short samples of specialised texts from three domains, covering three languages, and then we evaluated the ATE output of GenAI chatbots against human-generated glossaries. The languages in focus are English, French, and Portuguese, and we tested the extraction in these domains: board games, ice hockey, and kitesurfing. Our research questions were the following:

- How good (in terms of f-measure) are GenAI chatbots for extracting terms from a small text sample?
- Is there a dominant system across the three tested languages or domains?
- Does providing more information in the prompt increase output quality?
- Do ATE prompts work better when written in the target language?
- How impactful is the variability in output for ATE when using GenAI chatbots that have a temperature setting above zero?

The main contribution of this paper is thus a comparison of the output of multiple GenAI chatbots when prompted to perform ATE with shorter documents across different domains and languages. This area of research has remained much less explored, and such an overview of how chatbots perform in ATE can provide good insights to translators, terminologists, and translation managers in their choice of ATE methods when confronted with a single text from an unfamiliar domain.

The rest of this paper is structured as follows: Section 2 focuses on recent work that uses large language models and chatbots for ATE; Section 3 presents the construction of the corpora and the smaller sample data, and it also describes the chatbots and prompting strategies that were used; in Section 4, we present and briefly discuss the results with the help of several comparison tables, highlighting the GenAI contrast across the domains, languages and prompting strategies; finally, Section 7 contains some further considerations about the results, a discussion about the limitations of the study, and a few pointers for future work.

2. Related work

The field of automatic term extraction (ATE) continues to receive considerable attention, as can be observed in the numerous surveys that appeared in the past few years (cf. Di Nunzio et al., 2023; Andrade et al., 2023; Tran et al., 2023; Xu et al., 2025). While these surveys present an overview of recent ATE approaches, tools, and techniques, in this section, we focused on papers that specifically applied GenAI chatbots or generative large language models (LLMs) for ATE and ATE-related tasks.

This is a relatively recent methodology, and not many studies have already tested the use of GenAI chatbots for ATE, possibly because generative LLMs have been shown to perform rather poorly for tasks where token classification is a common approach. For instance, Wang et al. (2023) tested GPT3 for named-entity recognition, and the results were worse in comparison to supervised baselines. Still, a few authors have addressed the use of GenAI chatbots, mainly ChatGPT, for ATE and ATE-related tasks and here we present their findings.

Siu (2023) showed how ChatGPT can be used for several tasks related to translation, including ATE associated with term explanation. The results were, however, not quantified in the paper, as the authors were focusing on showing the system's capacity for helping in translation tasks, but not on quantifying the quality of this help.

In another study, de Paiva et al. (2023) tried to complement the annotation of mathematical concepts in a corpus by comparing human annotations with ChatGPT's ATE capabilities. Results were evaluated in terms of Jaccard similarity and they showed that ChatGPT had

an overall similarity of 0.5 with the human annotators, while the three human annotators achieved results upwards of 0.74 among themselves.

Martínez-Cruz et al. (2025) compared ChatGPT with KeyBART (Kulkarni et al., 2022) in case studies that used shorter and longer documents from different domains, which bears some similarities with our approach. They reported varying results in extraction quality.

In a study related to terminology work, Vidal Sabanés & da Cunha (2025) also tested ChatGPT for medical terminology extraction in Spanish. They organise the terminology in several categories, provide a gold standard (which is taken from medical glossaries that are also used as sample for ATE), and the results of the extraction are provided in terms of accuracy, precision and recall. Their f1-score for ATE with ChatGPT lies at 0.51, with a precision of 0.66 and a recall of 0.42, considering the 180 terms in the gold standard.

These two last papers are the most similar to our task, as the output from ChatGPT is confronted with manual annotations or a gold standard. In their study, de Paiva et al. (2023) focused on annotations by three human annotators and used similarity scores to check the inter-annotator agreement and the agreement of the system with the annotators. In our study, as further discussed below, in Section 3, we preferred to have each specialist annotating terms in their own field of expertise, instead of having several annotators for each domain, because the other annotators were not domain experts. In the case of Vidal Sabanés & da Cunha (2025), the gold standard was already available in the sample itself, as the corpus they used was composed of several glossaries, so no annotation was needed. They opted however, for an evaluation not in terms of agreement, but in terms of precision and recall, similar to what we propose in our study.

Moving to a closely related task, Song et al. (2023) and Martínez-Cruz et al. (2025) tested the power of ChatGPT for keyphrase extraction in a series of datasets. Both papers compare the results of ChatGPT with fine-tuned baselines and provide an evaluation in terms of f1-score (in this case f1@5 and f1@M). While ChatGPT only had the best result in one dataset for Song et al. (2023), for Martínez-Cruz et al. (2025), ChatGPT had the best results in most of the cases. One explanation for the difference in results could be the use of a newer GPT version: Martínez-Cruz et al. (2025) were able to access *GPT-3.5 Turbo*, while Song et al. (2023) report a test date of 1 March 2023. Other plausible explanations include GPT's variation in its replies or simply the difference in prompts.

More recently, Chun et al. (2025) released an arXiv paper proposing a new method for ATE using LLMs. They test the capacity of LLMs to learn from examples with syntactically similar sentences from other domains. They compare the extraction results of three LLMs (Llama-3.1-8B-IT, Gemma-2-9B-IT, and Mistral-Nemo) over three benchmark datasets, using four prompting methodologies. The authors also observed that, in general, generative LLMs had worse performance in the benchmarks when compared to pre-trained language models (PTLMs), such as RoBERTa-large and BART-large, adding to the evidence that generative LLMs tend to perform worse than PTLMs for token-based classification tasks.

Each of these studies contribute in their own way to the field of ATE using GenAI chatbots or generative LLMs, but a more comprehensive analysis of GenAI chatbots for ATE such as we propose in this study is still missing, especially when considering the needs of a translation task, which involves smaller text samples. Most studies either consider only one GenAI chatbot (usually ChatGPT) or a single language. This points to one of the

main contributions of our study, but also exposes the need for further studies in the area, because there are many variables involved in ATE using genAI chatbots, as we further discuss later in Section 7.

3. Methodology

In this section, we first describe the composition of the corpora and how they were created, and then we discuss the chatbots and prompting strategies used in the study.

3.1 Corpus and sample data

In this study, we work with three separate domains that reflect the specialisation of the authors, who have a background in translation studies, with emphasis on terminology. The three domains are *board games*, *ice hockey* and *kitesurfing*. This makes for an interesting mix, in which we have two corpora from the broader sports domain, and one corpus that relates to the hobby domain. Each of the three corpora was created with Sketch Engine (Kilgarriff et al., 2014) using the *Find texts on the web* feature, which allows for the automatic scraping of data guided by a few seeds or keyphrases[1]. Each domain was scraped in two languages: English and French for the two sports-related domains, and English and Portuguese for the board games domain. The selection of languages also reflected the specialisation of the authors. This decision of focusing on the specialisation of the authors was made to ensure a higher quality in the manual annotation of data for the creation of the evaluation datasets, as we will further discuss in Subsection 3.2.

From the scraped data, which amounted to between 60k and 200k words for each language and domain (as verified with Sketch Engine), we then proceeded with a light cleaning of extraneous sources (that inevitably get selected in a Web scraping) and a selection of a few portions of data to form a sample of each corpus. The sample was selected based on the types of terms that were present and on the term density of the excerpts. We preferred selection over random sampling, because we did not want to risk ending up with meaningless passages that contained only a few terms from the domains in focus. We then proceeded to an annotation process, which we discuss in the next subsection.

3.2 Evaluation datasets

Each of the selected domains corresponds to domains of interest for the authors, and the same applies to the languages in focus, so we did not proceed with multiple annotations of each sample. Considering that terminology annotation is a complex task, it can lead to a very low inter-annotator agreement, as discussed in Zilio et al. (2020), so we opted for having one specialist annotating both single-word and multi-word terms in each of the samples[2].

Table 1 contains descriptive and terminological information related to each sample. The selected samples have varying degrees of term density, but they all contain a high percentage

[1] The terms that were used as seeds for scraping each corpus were selected from a single text that belonged to the domain in focus. We used between 4 and 5 key phrases per corpus.

[2] The prompts (discussed in Subsection 3.3), the sample data and the lists of terms for each language and domain are available in the following GitHub repository: https://github.com/uebelsetzer/GenAI_chatbots_for_ATE.

	Board games		Ice hockey		Kitesurfing	
	EN	**PT**	**EN**	**FR**	**EN**	**FR**
Word tokens*	1984	2227	2078	2571	1888	2047
Word types*	579	718	572	576	525	587
Term tokens**	412	436	181	135	97	101
Term types**	210	222	63	52	17	25

*Verified with AntConc (Anthony, 2005), version 3.5.9, released in 2020.

*** Term tokens consist of one or more word tokens, and term types are based on surface form.

Table 1: Sample data from each domain and language

of specialised tokens. Regarding the ease of term extraction in the smaller samples, kitesurfing seems to offer more repetition, with a terminological type-token ratio (TTR) of 0.175 for English and 0.248 for French. On the other extreme, the board games domain presents a high number of terms and not much repetition, with a TTR just above 0.5 for both languages. Ice hockey offers a middle ground, with a good variety of terms and some repetition, presenting a TTR of 0.348 for English and 0.385 for French.

3.3 Chatbots and prompts

In this ATE experiment, we tested well-known GenAI chatbots: *ChatGPT*[3], *Copilot*[4], *Gemini*[5], *NotebookLM*[6], *DeepSeek*[7] and *Le Chat*[8]. ChatGPT and Copilot are both built upon a GPT model, although the first belongs to OpenAI and the latter to Microsoft. Gemini and NotebookLM are both developed by Google and use a Gemini model on the back-end. We then pit these systems against Le Chat, developed by the French start-up Mistral AI, and DeepSeek, which is developed by an homonymous Chinese company. These latter two figure as the European and Chinese responses to the Silicon Valley giants.

An important methodological decision was taken in regard to the use of the free version of each of these chatbots. We did not use the APIs for direct access to the underlying generative LLMs. As such, we used default settings for all parameters, including the underlying models themselves and their temperature parameter. This decision was taken with the idea of better simulating a potential interaction between a translator and a chatbot.

All chatbots were tested following four prompting strategies: zero shot, one shot, few shots, and few shots with a context sentence for each example-term. These strategies are applied in a cross-domain learning approach (Tran et al., 2023), as the LLMs that are trained for the chatbots are not specifically trained for ATE, but are then requested to perform such task with a few pieces of information (one shot and few shots), or no extra information at all (zero shot).

[3] https://chatgpt.com/.
[4] https://copilot.microsoft.com/.
[5] https://gemini.google.com/app.
[6] https://notebooklm.google.com/.
[7] https://chat.deepseek.com/.
[8] https://chat.mistral.ai/.

The main textual body of the prompt remained the same in all four approaches, and it instructs the system to perform a thorough terminological extraction from the sample data, and only the number and type of examples change across the four approaches. It starts with information about the background, the language and the domain under scope. It then provides information about the task and introduces the sample data. There is then a definition of what is expected regarding form, comprehensiveness, and output format. Here is the main prompt that we used, with placeholders for language and domain:

> I am working on a terminology extraction project that considers the <LANGUAGE> terminology used in <DOMAIN>. Attached, I am providing a text sample of around 2 thousand tokens from the domain of <DOMAIN>. Considering the principles of Terminology, I need you to create a comprehensive list of <DOMAIN> terms present in the sample. These terms can be formed by single words, or they can be multiword terms, and a multiword term might be formed by different single-word terms (and all of them should be extracted). It is very important for the extracted list to be comprehensive and to contain all terms that appear in the provided sample. The list should be formatted as a single-column output, with one term per line. I don't need definition nor context of occurrence, but the extracted terms must appear in the sample that was provided.

This initial zero-shot template was then extended to have one example, for the one-shot strategy, with the following add-on sentence:

> Here is an example of term from the <DOMAIN> domain (this term does not appear in the provided sample), for reference: <EXAMPLE-TERM1>.

And it was also similarly extended to have six term examples, for the few-shot approach, and then again to have example-sentences, for the few-shot approach with contexts:

> Here is a list of terms from the <DOMAIN> domain (these terms do not appear in the provided sample), each accompanied by an example-sentence, for reference:
> Example 1: <EXAMPLE-TERM1>
> <EXAMPLE-SENTENCE1>

As explicitly mentioned in the prompts, the example-terms used in three of the four approaches were not present in the sample data. They were collected, along with their contexts of occurrence, from the larger corpus. As such, we did not give a better starting point for the approaches with examples, we merely provided extra information about the domain and the language in focus. We used the prompts in English even when querying the French and Portuguese corpora, but we also experimented using prompts written in the target language, ensuring that the French prompt was the same for both sports domains. As such, we used the English prompt for English, French, and Portuguese (with slight adjustments of the example-terms and context-sentences, which were compiled according to the domain and language), and then we also tested a prompt in Portuguese and French for these target languages. The sample data was fed to the chatbots as a single-file attachment containing raw text.

Because we are dealing with several systems that function differently, a few adjustments had to be made for two of the chatbots. In the case of NotebookLM, it simply cannot produce a single-column output, and it would always provide a space-separated list (with no clear term boundaries). We then added a single extra prompt for NotebookLM requiring the conversion of the space-separated list into a comma-separated list, which was easier to work with. For Le Chat, we noticed that using raw text input was resulting in the system often getting into an infinite loop. It would generate thousands of candidates on end. We observed that this behaviour did not occur when we attached a DOCX file instead of TXT so we ended up using a DOCX file for Le Chat, with the exact same content and no extra formatting.

In summary, we had a total of six chatbots, three domains, two languages per domain, and four prompting strategies, totalling 144 interactions with the chatbots using prompts in English, plus 72 interactions using prompts written in the target languages. Each interaction was created within a new chat in each of the systems. This was done in order to not contaminate the results of one strategy with inputs from previously used strategies. Apart from the two slight adjustment required by NotebooLM to output a list of terms as explained above, all the systems were tested following exactly the same one-prompt procedure. The result from each interaction was then recorded and compared to the human-annotated datasets and evaluated in terms of precision, recall and f-measure. All tests were run between 20 May 2025 and 26 June 2025.

4. Results and analysis

The results for each of the domains in each of the languages under study, including the tests with prompts in the target language are presented in Tables 2 to 10. These tables are grouped by domain, and the best result in each column is highlighted in boldface. For the column of ”unique terms”, the best result was evaluated as the closest to the actual number of unique annotated terms in the evaluation datasets.

Because the number of extracted terms varies for each system, the f-measure result is the one we used for evaluating the best systems. For instance, we can see in Table 2 that DeepSeek achieves high precision (94%) when we used a few-shot strategy with added context sentences, but it has only extracted 100 terms, which is then reflected in a recall of 44.76%. Inversely, on that same prompting strategy, NotebookLM achieves the highest recall (81.90%) with its 364 extracted terms, but its precision is much lower, at 47.25%. The trade-off between precision and recall is well known, and that is why the f-measure (a harmonic mean) is important. Considering the f-measure, it is NotebookLM, with the one-shot strategy, that has the best overall result for English in the board games domain, with a score of 73.98%.

Such evaluation also depends on the requirements of the task. For instance, if the objective of the extraction is not necessarily to get all the terms, but to have a list that can be trusted without too much cleaning, then precision would have a higher weight. On the other hand, if cleaning the list is not a problem, but the extraction should include as many terms as possible, then more weight would be given to recall. This is also why we report all three values, and not just the f-measure, even if here we give more importance to the f-measure.

After testing all the selected systems across all the proposed prompting strategies, we have overall clear results for each language and domain, but no clear overall best system. The best systems actually varied quite a lot, from domain to domain, and from language to language, so there is no emerging pattern for "the" best system. We have NotebookLM working fairly well for the board games domain, but having a rather poor result in the other two domains, and we have Gemini working well both for board games and kitesurfing, but having bad results in ice hockey, except when the prompt was in the target language. ChatGPT, which was the only system tested by most of the previous studies, only excelled for kitesurfing in English, where it got the best overall score, but had an otherwise average performance on the other domains, with a few highlights for recall in some scenarios.

The systems varied wildly also in terms of the number of extracted terms. However, in a quantity-over-quality scenario, NotebookLM seems to be recommended, as it provided the most extracted terms in most of the tested scenarios, and it usually presented an above-average recall. On the other hand, Copilot presented outputs with overall fewer items, but usually high precision.

Overall, the prompts written in French and Portuguese did not improve the results for these languages, as their English counterpart was most of the time better for ATE. The only exception to this was in the ice hockey domain, where the results for a zero-shot extraction with Gemini achieved the highest f-measure for French, even if most of the other tests decreased in value.

Although we highlighted the best results in each of the strategies, many times the results are fairly close to each other, and due to the temperature setting in GenAI chatbots, which allow the systems to produce non-deterministic results, we could not ascertain whether the differences in the results were statistically significant. For that reason, and also to test the variability of system outputs, we selected the two best overall results for each language and domain to undergo a more complete test run, consisting of five consecutive runs using exactly the same settings, each on a new chat. This process follows in the footsteps of a five-fold cross-validation in classic machine learning. It provides us with a mean f-measure and a standard deviation for the tested chatbots, which in turn allow us to verify whether there is statistical significance in the difference observed between the systems, as well as how much the outputs vary within the same chatbot. The selected systems and their respective mean f-measure and standard deviation are displayed in Table 11.

The results of the five-run test show that most of the systems have enough variation in their answers to present no statistical difference from the second best, and sometimes even from lower ranked systems[9]. Only Gemini, for Kitesurfing in French, was statistically better than Copilot, the other system that was selected for the five-run test.

This five-run sub-experiment also seems to confirm that NotebookLM has its temperature set to zero, or very close to zero, as the outputs for both English and Portuguese in the board games domain were exactly the same over the five runs. However, there was also one issue that became apparent: Gemini (and thus also NotebookLM) had a system update

[9] In the case of systems with only one run, we consider that there is no statistical difference when their f-score is within two standard deviations from the mean f-score of the system that was tested with the five-run. When both systems were tested with five runs, then statistical difference was observed when the systems' mean f-scores plus or minus one standard deviation do not overlap. This would give us around 95% of statistical confidence.

happening in-between the tests that we originally ran and the five-run test. The large language model behind the two chatbots was changed from Gemini 2.0 to Gemini 2.5[10]. Because it was outside the scope of this study, and because keeping up with all the system updates would possibly turn into an infinite catch-up game, we simply ran the five-run test with the new model, which ended up performing worse than its older counterpart.

[10] As of 17th June 2025: https://developers.googleblog.com/en/gemini-2-5-thinking-model-updates/.

Board games - English																
	Zero shot				One shot				Few shots				Few shots + context			
	Unique terms	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**
ChatGPT	147	0.8367	0.5857	0.6891	159	0.8239	0.6238	0.7100	122	**0.8689**	0.5048	0.6386	125	0.8720	0.5190	0.6507
Copilot	63	**0.9206**	0.2762	0.4249	88	0.8068	0.3381	0.4765	79	0.7722	0.2905	0.4221	73	0.8630	0.3000	0.4452
DeepSeek	135	0.8815	0.5667	0.6899	157	0.8217	0.6143	0.7030	157	0.8471	0.6333	**0.7248**	100	**0.9400**	0.4476	0.6065
Gemini	**194**	0.7629	0.7048	**0.7327**	**174**	0.7931	0.6571	0.7188	173	0.7977	0.6571	0.7206	**189**	0.7196	0.6476	**0.6817**
Le Chat	145	0.7517	0.5190	0.6141	123	**0.8374**	0.4905	0.6186	150	0.5467	0.3905	0.4556	101	0.7525	0.3619	0.4887
NotebookLM	269	0.6283	**0.8048**	0.7056	255	0.6745	**0.8190**	**0.7398**	**238**	0.6092	**0.6905**	0.6473	364	0.4725	**0.8190**	0.5993

Table 2: Number of unique terms, precision (P), recall (R) and f-measure (F) for the domain of Board Games in English

Board games - Portuguese																
	Zero shot				**One shot**				**Few shots**				**Few shots + context**			
	Unique terms	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**
ChatGPT	106	0.2547	0.1216	0.1646	**125**	0.7760	0.4369	0.5591	**117**	0.6752	0.3559	0.4661	139	0.5612	0.3514	0.4321
Copilot	70	0.5571	0.1757	0.2671	113	0.6903	0.3514	0.4657	85	0.8353	0.3198	0.4625	62	0.7581	0.2117	0.3310
DeepSeek	134	0.7687	0.4640	0.5787	124	0.8548	0.4775	0.6127	131	0.7481	0.4414	0.5552	116	**0.8448**	0.4414	0.5799
Gemini	149	0.7919	0.5315	0.6361	124	**0.9113**	**0.5090**	**0.6532**	**117**	**0.8803**	0.4640	**0.6077**	**172**	0.7733	0.5991	**0.6751**
Le Chat	88	**0.8750**	0.3468	0.4968	121	0.8099	0.4414	0.5714	109	0.8532	0.4189	0.5619	124	0.8226	0.4595	0.5896
NotebookLM	**232**	0.7026	**0.7342**	**0.7181**	0*	N/A	N/A	N/A	372	0.4005	**0.6712**	0.5017	446	0.3341	**0.6712**	0.4461

* System provided the following answer: "NotebookLM can't answer this question. Try rephrasing it, or ask a different question."

Table 3: Number of unique terms, precision (P), recall (R) and f-measure (F) for the domain of Board Games in Portuguese

Board games - Portuguese (with prompt in Portuguese)																
	Zero shot				**One shot**				**Few shots**				**Few shots + context**			
	Unique terms	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**
ChatGPT	**271**	0.2768	0.3378	0.3043	**175**	0.5314	0.4189	0.4685	177	0.5819	0.4640	0.5163	**177**	0.1808	0.1441	0.1604
Copilot	71	0.8451	0.2703	0.4096	26	**0.8846**	0.1036	0.1855	83	0.7590	0.2838	0.4131	73	0.8356	0.2748	0.4136
DeepSeek	120	**0.8500**	0.4595	**0.5965**	153	0.7908	**0.5450**	**0.6453**	143	0.8462	0.5450	**0.6630**	132	0.8106	**0.4820**	**0.6045**
Gemini	95	0.7158	0.3063	0.4290	120	0.8833	0.4775	0.6199	**194**	0.6959	0.6081	0.6490	118	0.8475	0.4505	0.5882
Le Chat	113	0.8319	0.4234	0.5612	94	0.8723	0.3694	0.5190	98	**0.8673**	0.3829	0.5313	104	**0.8654**	0.4054	0.5521
NotebookLM	290	0.5103	**0.6667**	0.5781	0**	N/A	N/A	N/A	321	0.5452	**0.7883**	0.6446	0*	N/A	N/A	N/A

* System provided the following answer: "NotebookLM can't answer this question. Try rephrasing it, or ask a different question."

** System provided an answer completely outside the scope of the prompt.

Table 4: Number of unique terms, precision (P), recall (R) and f-measure (F) for the domain of Board Games in Portuguese, using a prompt in the target language

Ice Hockey - English																
	Zero shot				One shot				Few shots				Few shots + context			
	Unique terms	P	R	F	Unique terms	P	R	F	Unique terms	P	R	F	Unique terms	P	R	F
ChatGPT	91	0.4725	**0.6825**	0.5584	111	0.4144	**0.7302**	0.5287	91	0.4505	0.6508	0.5325	72	0.5139	0.5873	0.5481
Copilot	**54**	**0.6111**	0.5238	0.5641	58	0.5345	0.4921	0.5124	42	**0.6429**	0.4286	0.5143	68	0.4118	0.4444	0.4275
DeepSeek	93	0.4624	**0.6825**	0.5513	99	0.4242	0.6667	0.5185	83	0.4819	0.6349	0.5479	85	0.5294	**0.7143**	**0.6081**
Gemini	129	0.2713	0.5556	0.3646	106	0.3774	0.5763	0.4734	98	0.3878	0.6032	0.4720	87	0.4023	0.5556	0.4667
Le Chat	74	0.5405	0.6349	**0.5839**	**59**	**0.5763**	0.5397	**0.5574**	**56**	**0.6429**	0.5714	**0.6050**	**68**	**0.5652**	0.6190	0.5909
NotebookLM	269	0.1413	0.6032	0.2289	231	0.1818	0.6667	0.2857	377	0.1220	**0.7302**	0.2091	0*	N/A	**N/A**	N/A

* System provided the following answer: "the system doesn't respond".

Table 5: Number of unique terms, precision (P), recall (R) and f-measure (F) for the domain of Ice Hockey in English

Ice Hockey - French																
	Zero shot				**One shot**				**Few shots**				**Few shots + context**			
	Unique terms	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**
ChatGPT	88	0.3977	0.6731	0.5000	125	0.3040	**0.7308**	0.4294	72	0.4444	0.6154	0.5161	77	**0.4675**	0.6923	**0.5581**
Copilot	**59**	0.4407	0.5000	0.4685	**63**	**0.4762**	0.5769	**0.5217**	41	0.4634	0.3654	0.4086	**62**	0.4355	0.5192	0.4737
DeepSeek	78	0.4615	0.6923	0.5538	80	0.4000	0.6154	0.4848	81	0.4198	0.6538	0.5113	73	0.4658	0.6538	0.5440
Gemini	122	0.3361	**0.7885**	0.4713	103	0.3495	0.6923	0.4645	95	0.3474	0.6346	0.4490	115	0.3217	**0.7115**	0.4431
Le Chat	69	**0.5072**	0.6731	**0.5785**	83	0.3976	0.6346	0.4889	**53**	**0.5741**	0.5962	**0.5849**	103	0.3495	0.6923	0.4645
NotebookLM	283	0.0709	0.3846	0.1198	116	0.2931	0.6538	0.4048	118	0.3136	**0.7115**	0.4353	0*	N/A	**N/A**	N/A

* System provided the following answer: "the system doesn't respond".

Table 6: Number of unique terms, precision (P), recall (R) and f-measure (F) for the domain of Ice Hockey in French

Ice Hockey - French (with prompt in French)																
	Zero shot				**One shot**				**Few shots**				**Few shots + context**			
	Unique terms	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**
ChatGPT	119	0.3361	**0.7692**	0.4678	104	0.3654	**0.7308**	0.4872	113	0.3363	**0.7308**	0.4606	89	0.3933	**0.6731**	0.4965
Copilot	94	0.3085	0.5577	0.3973	90	0.3778	0.6538	0.4789	95	0.3789	0.6923	0.4898	79	0.3671	0.5577	0.4427
DeepSeek	94	0.3936	0.7115	0.5068	84	**0.4286**	0.6923	**0.5294**	79	0.4430	0.6731	0.5344	95	0.3053	0.5577	0.3946
Gemini	**49**	**0.6327**	0.5962	**0.6139**	115	0.3304	**0.7308**	0.4551	**45**	**0.6444**	0.5577	**0.5979**	**45**	**0.6444**	0.5577	**0.5979**
Le Chat	69	0.4928	0.6538	0.5620	**75**	0.4267	0.6154	0.5039	95	0.3750	0.6923	0.4865	63	0.5397	0.6538	0.5913
NotebookLM	200	0.1800	0.6923	0.2857	170	0.2176	0.7115	0.3333	138	0.2609	0.6923	0.3789	0*	N/A	**N/A**	N/A

* System provided the following answer: “the system doesn’t respond”.

Table 7: Number of unique terms, precision (P), recall (R) and f-measure (F) for the domain of Ice Hockey in French, using a prompt in the target language

Kitesurfing - English																
	Zero shot				One shot				Few shots				Few shots + context			
	Unique terms	P	R	F	Unique terms	P	R	F	Unique terms	P	R	F	Unique terms	P	R	F
ChatGPT	85	0.0941	0.4706	0.1569	87	0.1034	**0.5294**	0.1731	59	**0.2034**	**0.7059**	**0.3158**	82	0.1220	0.5882	0.2020
Copilot	82	0.0732	0.3529	0.1212	59	0.1017	0.3529	0.1579	54	0.1111	0.3529	0.1690	82	0.1098	0.5294	0.1818
DeepSeek	67	0.0597	0.2353	0.0952	60	0.0833	0.2941	0.1299	75	0.0800	0.3529	0.1304	85	0.0706	0.3529	0.1176
Gemini	58	**0.1552**	**0.5294**	**0.2400**	60	0.1500	**0.5294**	0.2338	62	0.1774	0.6471	0.2785	**35**	**0.2286**	0.4706	**0.3077**
Le Chat	**47**	0.0851	0.2353	0.1250	**49**	**0.1837**	**0.5294**	**0.2727**	**49**	0.1633	0.4706	0.2424	128	0.0703	0.5294	0.1241
NotebookLM	288	0.0243	0.4118	0.0459	280	0.0250	0.4118	0.0471	170	0.0235	0.2353	0.0428	141	0.0780	**0.6471**	0.1392

Table 8: Number of unique terms, precision (P), recall (R) and f-measure (F) for the domain of Kitesurfing in English

Kitesurfing - French																
	Zero shot				One shot				Few shots				Few shots + context			
	Unique terms	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**	**Unique terms**	**P**	**R**	**F**
ChatGPT	62	0.2742	0.6800	0.3908	62	0.2903	0.7200	0.4138	74	0.1892	0.5600	0.2828	61	0.3115	0.7600	0.4419
Copilot	64	0.2656	0.6800	0.3820	**35**	**0.4000**	0.5600	**0.4667**	**27**	**0.5185**	0.5600	**0.5385**	62	0.2742	0.6800	0.3908
DeepSeek	66	0.2576	0.6800	0.3736	60	0.2667	0.6400	0.3765	53	0.3019	0.6400	0.4103	70	0.2429	0.6800	0.3579
Gemini	**35**	**0.4000**	0.5600	**0.4667**	90	0.2111	**0.7600**	0.3304	42	0.4048	0.6800	0.5075	**45**	**0.4000**	0.7200	**0.5143**
Le Chat	**35**	0.3429	0.4800	0.4000	60	0.2667	0.6400	0.3765	36	0.3889	0.5600	0.4590	41	0.3415	0.5600	0.4242
NotebookLM	61	0.2951	**0.7200**	0.4186	70	0.2571	0.7200	0.3789	61	0.3115	**0.7600**	0.4419	79	0.2532	**0.8000**	0.3846

Table 9: Number of unique terms, precision (P), recall (R) and f-measure (F) for the domain of Kitesurfing in French

Kitesurfing - French (with prompt in French)																
	Zero shot				One shot				Few shots				Few shots + context			
	Unique terms	P	R	F	Unique terms	P	R	F	Unique terms	P	R	F	Unique terms	P	R	F
ChatGPT	87	0.1494	0.5200	0.2321	60	0.2833	**0.6800**	0.4000	66	0.1970	0.5200	0.2857	75	0.2400	0.7200	0.3600
Copilot	80	0.2000	0.6400	0.3048	65	0.2308	0.6000	0.3333	69	0.1884	0.5200	0.2766	78	0.2179	0.6800	0.3301
DeepSeek	88	0.2045	0.7200	0.3186	75	0.2267	**0.6800**	0.3400	95	0.1895	0.7200	0.3000	70	0.2429	0.6800	0.3579
Gemini	**35**	0.3143	0.4400	0.3667	**44**	**0.3636**	0.6400	**0.4638**	**48**	**0.3542**	0.6800	**0.4658**	49	0.3469	0.6800	0.4595
Le Chat	**35**	**0.4000**	0.5600	**0.4667**	60	0.2333	0.5600	0.3294	49	0.2857	0.5600	0.3784	**47**	**0.3830**	0.7200	**0.5000**
NotebookLM	111	0.1982	**0.8800**	0.3235	77	0.2078	0.6400	0.3137	83	0.2771	**0.9200**	0.4259	89	0.2135	**0.7600**	0.3333

Table 10: Number of unique terms, precision (P), recall (R) and f-measure (F) for the domain of Kitesurfing in French, using a prompt in the target language

	Board games				Ice hockey				Kitesurfing			
	English		Portuguese		English		French		English		French	
	Gemini Zero shot	NotebookLM One shot	NotebookLM Zero shot	Gemini Few shots +	DeepSeek Few shots +	Le Chat Few shots	DeepSeek Zero shot	Le Chat Few shots	ChatGPT Few shots	Gemini Few shots +	Copilot Few shots	Gemini Few shots +
Mean F1	0.6334	0.6489	0.4617	0.5794	0.5563	0.5922	0.4894	0.4915	0.1942	0.2256	0.3210**	0.4753**
StDev	0.0620	N/A*	N/A*	0.0716	0.0439	0.0174	0.0368	0.0161	0.0360	0.0377	0.0338	0.0447

*NotebookLM did not present any variation along the five runs. | **Statistically different.

Table 11: Mean f1-scores and standard deviation based on five consecutive runs of the best systems for each domain and language.

5. Discussion and final remarks

This study focused on an analysis of ATE from small text samples using several available chatbots across different languages and domains. We could observe that there was no emerging dominant system. Gemini was among the best f-measures for a few of the scenarios we tested, but the recent release of a new model already showed a score decrease on the five-run test.

In general, systems were hard-pressed to achieve an f-score above 0.6. While many systems were good in providing a lot of candidates (usually achieving a good recall) and others were good at providing fewer, more targeted, candidates, no system achieved a consistently high precision and recall. Even so, the results show which systems are prone to providing more candidates, and which ones are conservative in their extraction. As such, although the f-scores were not off the charts, there is some useful information for future ATE attempts using chatbots. The variation in the five-run tests also shows that, apart from NotebookLM, which has no variation in the output (probably due to a zero or close to zero temperature setting), the results from the best systems in the original single run were from the systems "on a good day", and the five-run test showed that those f-scores were close to, or even beyond, one standard deviation above the mean.

We also observed that writing the prompt in the target language does not seem to improve the results when compared to writing it in English, as most of the tests with a prompt written in Portuguese or French resulted in poorer performance for these languages in comparison to the use of a prompt in English. Still, this only seems to be a tendency, as there were some results that were indeed better with the prompt in the target language, such as for ice hockey, where the best result was achieved with the prompt written in French.

In terms of providing more information with one-shot and few-shot prompts, our expectation was that, with added information, the GenAI chatbots would provide better results due to their in-context learning capabilities, as demonstrated, for instance, by Brown et al. (2020) in tests with GPT3. In our tests, however, more information did not immediately translate into better results, as the zero-shot prompts produced, in some scenarios, even better outputs than the prompts with examples and context-sentences. Figures in Annex A show the curves for the prompting strategies, and there doesn't seem to be a clear pattern, except from a usual raise from zero-shot to one-shot prompts.

5.1 Limitations and future research

As a piece of a larger research puzzle, this study has some limitations, but it also opens the door for some future research possibilities:

- Because of the non-determinism in the systems' replies, the results presented here could potentially differ in a second run. To mitigate this issue, we performed a five-run test on the best systems, but, for a more complete overview, such a technique would be in order for all the systems and all the scenarios. This would, however, have required an amount of time and work that we unfortunately could not spare.
- The results from this study, while fairly comprehensive in the comparison of several systems, languages and prompting strategies, remain as an image frozen in time

of the systems' performances as they were available to the general public in their free versions. As with all technologies, further developments will make these results obsolete when newer models are released, requiring further studies to keep the comparison up to date. In this regard, we hope that the methodology presented here, although not fully innovative, but pieced together from several sources, will be useful for further studies that delve into the ATE performed by GenAI models.

- One of the main causes for variation in chatbot outputs, besides the temperature setting, is the prompt itself. Simple changes in the prompt can lead to different results (Salinas & Morstatter, 2024). We cannot, however, test the chatbots for all possible variations, so we instead created a rather complete prompt to cover all the details of the task, from background to output format. However, it would be an interesting study to observe the variation in ATE generated by several runs of slightly different prompts.
- Because we were not working with a full corpus (except for extracting the samples, and a few terms and sentences for the prompts), we cannot compare this extraction with the extraction of a traditional statistical ATE system, such as Sketch Engine (Kilgarriff et al., 2014) or TBXTools (Oliver, 2017). As a future comparison, it would be possible, however, to perform a simple pattern-based extraction, relying on part-of-speech annotation and a previous evaluation of patterns in a larger corpus. This could be performed, for instance, with the linguistic features in TBXTools or with mwetoolkit (Ramisch et al., 2010), or even with pre-trained language models that are fine-tuned for ATE, but each of these options would come with their own limitations. Another option would be to use available benchmarks, as in Chun et al. (2025), but then the idea of simulating a translator's task would be lost.

To sum up, this study contributed to giving a broad overview of the GenAI chatbot systems and their quality for ATE using small text samples. The results presented here can prompt further research in the area and also provide a point of reference for researchers and translators who would like to have more information about the quality of GenAI for ATE and token classification tasks in general.

6. Acknowledgements

We would like to thank UCLouvain, the Institute of Language and Communication, and the Louvain School of Translation and Interpreting for the support and for providing the necessary conditions for this research to be developed in a pedagogical context.

Software

Ammon, U., Bickel, H. & Lenz, A.N. (eds.) (2016). *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen.* Berlin/Boston: De Gruyter Mouton, 2 edition.

Andrade, J.C.B., Otálvaro, C.M.M., Jaramillo, C.M.Z. & Ríos, A.M. (2023). Approaches, tools, algorithms, and methods for automatic term extraction: A systematic literature mapping. *Research Square, preprint.*

Anthony, L. (2005). AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *IPCC 2005. Proceedings. International Professional Communication Conference, 2005.* IEEE, pp. 729–737.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, pp. 1877–1901.

Chun, Y., Kim, M., Kim, D., Park, C. & Lim, H. (2025). Enhancing Automatic Term Extraction with Large Language Models via Syntactic Retrieval. *arXiv preprint arXiv:2506.21222*.

Costa, H., Zaretskaya, A., Corpas Pastor, G. & Seghiri Domínguez, M. (2016). Nine terminology extraction Tools: Are they useful for translators? *Multilingual*, 27(3), pp. 14–20.

de Paiva, V., Gao, Q., Kovalev, P. & Moss, L.S. (2023). Extracting Mathematical Concepts with Large Language Models. In *CEUR Workshop Proceedings*. pp. 1–13.

Di Nunzio, G.M., Marchesin, S. & Silvello, G. (2023). A systematic review of Automatic Term Extraction: What happened in 2022? *Digital Scholarship in the Humanities*, 38(Supplement_1), pp. i41–i47.

Giguere, J. (2023). Leveraging large language models to extract terminology. In *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*. pp. 57–60.

Hellrich, J. & Hahn, U. (2016). An Assessment of Experimental Protocols for Tracing Changes in Word Semantics Relative to Accuracy and Reliability. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Berlin, Germany: Association for Computational Linguistics, pp. 111–117. URL http://anthology.aclweb.org/W16-2114.

Kerremans, D., Stegmayr, S. & Schmid, H.J. (2011). The NeoCrawler: identifying and retrieving neologisms from the internet and monitoring ongoing change. In *Current Methods in Historical Semantics*. Berlin/Boston: De Gruyter, pp. 59–96. URL http://www.degruyter.com/view/books/9783110252903/9783110252903.59/9783110252903.59.xml.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P. & Suchomel, V. (2014). The sketch engine. *Lexicography*, 1(1), pp. 7–36.

Kulkarni, M., Mahata, D., Arora, R. & Bhowmik, R. (2022). Learning Rich Representation of Keyphrases from Text. In M. Carpuat, M.C. de Marneffe & I.V. Meza Ruiz (eds.) *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, pp. 891–906.

Martínez-Cruz, R., López-López, A.J. & Portela, J. (2025). Chatgpt vs state-of-the-art models: a benchmarking study in keyphrase generation task. *Applied Intelligence*, 55(1), p. 50.

Oliver, A. (2017). A system for terminology extraction and translation equivalent detection in real time: Efficient use of statistical machine translation phrase tables. *Machine Translation*, 31(3), pp. 147–161.

Pazienza, M.T., Pennacchiotti, M. & Zanzotto, F.M. (2005). Terminology extraction: an analysis of linguistic and statistical approaches. In *Knowledge mining: Proceedings of the NEMIS 2004 final conference*. Springer, pp. 255–279.

Ramisch, C., Villavicencio, A. & Boitet, C. (2010). mwetoolkit: A framework for multiword expression identification. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), pp. 662–669.

Šajatović, A., Buljan, M., Šnajder, J. & Bašić, B.D. (2019). Evaluating automatic term extraction methods on individual documents. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*. pp. 149–154.

Salinas, A. & Morstatter, F. (2024). The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance. In *Findings of the Association for Computational Linguistics ACL 2024*. pp. 4629–4651.

Schmidlin, R. (2011). *Die Vielfalt des Deutschen: Standard und Variation - Gebrauch, Einschätzung und Kodifizierung einer plurizentrischen Sprache*. Berlin: De Gruyter.

Siu, S.C. (2023). ChatGPT and GPT-4 for professional translators: Exploring the potential of large language models in translation. *Available at SSRN: https://ssrn.com/abstract=4448091*.

Song, M., Jiang, H., Shi, S., Yao, S., Lu, S., Feng, Y., Liu, H. & Jing, L. (2023). Is chatgpt a good keyphrase generator? a preliminary study. *arXiv preprint arXiv:2303.13001*.

Südtiroler Kulturinstitut (2017). Sprachstelle im Südtiroler KULTURinstitut. URL http://www.kulturinstitut.org/hauptnavigation/sprachstelle.html.

Tran, H.T.H., Martinc, M., Caporusso, J., Doucet, A. & Pollak, S. (2023). The recent advances in automatic term extraction: A survey. *arXiv preprint arXiv:2301.06767*.

Vidal Sabanés, L. & da Cunha, I. (2025). AI as a resource for the clarification of medical terminology: An analysis of its advantages and limitations. *Terminology*, 31(1), pp. 37–71.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J. & Wang, G. (2023). GPT-NER: Named Entity Recognition via Large Language Models. *arXiv e-prints*, pp. arXiv–2304.

Webber, W., Moffat, A. & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4), pp. 1–38.

Xu, K., Feng, Y., Li, Q., Dong, Z. & Wei, J. (2025). Survey on terminology extraction from texts. *Journal of Big Data*, 12(1), pp. 1–40.

Zilio, L., Paraguassu, L.B., Hercules, L.A.L., Ponomarekano, G.L., Berwanger, L.P. & Finatto, M.J.B. (2020). A lexical simplification tool for promoting health literacy. In *1st Workshop on Tools and Resources to Empower People with REAding DIfficulties*.

Annex A

Figure 1 in this annex illustrates the f-measure evolution over the different prompting strategies for each GenAI chatbot. The line graphs show how the f-measure of each system increases or decreases as we used the prompts written in English with an increasing amount of information, from a zero-shot prompt to a few-shot prompt with added context sentences.

As we can see in the graphs, there is no clear picture emerging from the systems. In several scenarios, we can see an increase in f-measure from zero- to one-shot prompts, but there is no clear tendency for an increase in performance from one-shot to few-shot prompting, or from few-shot prompting to the few-shot prompts with context sentences.

(a) Board games - English

(b) Board games - Portuguese

(c) Ice hockey - English

(d) Ice hockey - French

(e) Kitesurfing - English

(f) Kitesurfing - French

Figure 1: Evolution of f-measure with increasing information on the prompt.