Large Language Models are Good Term Extractors: A Systematic Evaluation

Ayla Rigouts Terryn
Université de Montréal
Mila - Quebec Artificial Intelligence Institute
ayla.rigouts.terryn@umontreal.ca

Abstract

This paper systematically evaluates modern large language models for automatic term extraction (ATE), examining GPT-5 and Mistral across four domains and three languages using the ACTER corpus. The study compares model sizes, evaluates reasoning-enhanced variants, and tests prompting strategies aligned with human annotation guidelines. Beyond extracting term lists, models provide term labels, confidence scores, and terminology management remarks. Current large language models achieve F1 scores of .36-.72; while seemingly low, this is competitive with supervised approaches and approaches the human inter-annotator agreement ceiling of 0.59. Larger models outperform smaller variants, with reasoning-enhanced models showing modest improvements. Qualitative error analysis reveals that evaluation methodology partly misrepresents model capabilities: many extractions classified as errors represent defensible boundary judgements, and apparent hallucinations are predominantly (though not exclusively) valid normalisations. Limitations remain in fine-grained categorisation and handling overly general expressions. However, the convergence of model scores with each other and with human inter-annotator agreement suggests that, for high-resource languages, basic ATE may no longer be the bottleneck in terminology management pipelines, and research should shift toward downstream tasks such as definition generation and ontology construction.

Keywords: automatic terminology extraction, large language models, evaluation, reasoning models

1. Introduction

Automatic term extraction (ATE) identifies domain-specific terminology in specialised texts, supporting language professionals such as translators and technical writers, and enabling downstream NLP tasks including taxonomy learning (Lefever, 2016) and aspect-based sentiment analysis (De Clercq et al., 2015). The field has evolved from rule-based approaches, whether statistical, linguistic, or hybrid (Drouin, 2003; Macken et al., 2013), through supervised machine learning (Rigouts Terryn et al., 2021; Hazem et al., 2022), to recent explorations with large language models (LLMs) (Tran et al., 2024; Banerjee et al., 2024).

However, research on LLM-based ATE remains limited. Key practical questions remain unexplored, including the impact of model size on ATE performance and whether reasoning-enhanced models outperform non-reasoning models. Only a few prompting strategies have been tried. Moreover, most prompting for ATE has been sentence-by-sentence, whereas the increasing context windows of LLMs now allow for larger chunks of texts to be added simultaneously, thus leveraging more relevant information. Additionally, the potential of LLMs for more advanced terminology work beyond the narrow information extraction task of ATE has barely been studied. Are LLMs capable of offering useful information to terminologists beyond pure term identification?

This study provides an updated evaluation of current LLMs for ATE, testing both GPT-5 and Mistral model families on the multilingual ACTER dataset (Rigouts Terryn et al., 2020b) across three languages and four technical domains. It includes a systematic comparison of model sizes (small vs. large) and types (standard vs. reasoning), with multiple prompts and using larger chunks of texts within the limits of the available context windows. Beyond extracting candidate term lists, models are prompted to include term labels (Specific, Common, or Out-of-Domain terms), confidence scores, and optional remarks for terminology management.

The evaluation examines traditional metrics (precision, recall, F1) alongside practical considerations such as hallucination rates and the reliability of confidence scores. By testing different prompting strategies in line with the human annotation protocols of the dataset, the study assesses how well LLMs can adapt to specific terminology extraction requirements. The results provide practical insights for researchers and practitioners considering LLMs for terminology work, illustrating current capabilities and areas requiring further development.

2. Related Research

Early explorations of LLM-based ATE have established promising baselines. Tran et al. (2024) and Banerjee et al. (2024) showed that early LLMs could compete with supervised fine-tuned models like XLM-RoBERTa and mBART. Both studies

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7631–7643
11–16 May 2026.

used the English part of the ACTER dataset and processed texts sentence-by-sentence. Tran et al. (2024) followed the TermEval shared task procedures (Rigouts Terryn et al., 2020a) training on corruption, equitation, and wind energy domains and testing on heart failure. They compared GPT-3.5-Turbo and Llama-2-Chat with few-shot prompting against XLM-RoBERTa and mBART trained as sequence labellers, finding comparable performance. Banerjee et al. (2024) compared GPT-3.5-Turbo with 5 to 30-shot prompting against XLM-RoBERTa-base fine-tuned on the same limited samples across all four ACTER domains. Testing on 150-sample subsets across domains, they found GPT-3.5-Turbo outperformed both the fine-tuned model and unsupervised baselines (C-Value (Frantzi and Ananiadou, 1999) and ComboBasic (Astrakhantsev et al., 2015)). Both studies also identify hallucinations, i.e., extracted candidate terms that are not present in the original texts, as challenges (Huang et al., 2025) potentially solved through rule-based post-processing.

Prompting strategies for ATE remain limited. Tran et al. (2024) experimented with brief versus more elaborate task instructions and compared three output formats: sequence labelling, term lists, and marking terms within copied text, finding the best strategy depended on the model. Banerjee et al. (2024) tested eight prompt templates varying in task instruction complexity (from "identify terms" to more detailed domain-specific descriptions), finding that detailed instructions yielded modest improvements of approximately 1-2%. Tran et al. (2025) expanded their earlier work with Llama-2-chat across all ACTER languages, introducing self-verification mechanisms where models assessed their own extractions. They found that marking terms within copied text reduced hallucinations compared to generating lists, and notably, cross-lingual few-shot examples performed comparably to in-lingual ones. Chun et al. (2025) explored syntactically-informed example selection for 10-shot prompting on English ACTER data, achieving marginal improvements over random selection. All studies used English prompts, regardless of text language.

Model comparison studies for terminology tasks remain sparse. Breton et al. (2025) compared GPT-4 against Mistral variants for French legal information extraction with predefined categories (a related but distinct task from open-ended ATE), finding GPT-4 superior and confirming that larger models generally performed better. No studies have systematically compared model sizes or reasoning capabilities specifically for terminology extraction across multiple domains and languages.

Current evaluation practices rely on precision, recall, and F1 metrics, with no consistent findings on higher precision or recall. Error analyses remain relatively limited, with Banerjee et al. (2024) noting better performance on abbreviations and highly specialised terms, Breton et al. (2025) finding partial matches problematic, and Tran et al. (2025) reporting better recall on shorter terms. Most studies have focused on single languages or domains using sentence-level processing. Adding supplementary information to the extracted terms like confidence scoring and structured metadata generation (term categorisation, usage notes) have not been explored.

3. Methodology

Models and Configurations The study initially planned to evaluate two model families, each with two sizes (small and large) and two types (standard and reasoning), yielding four configurations per family. For GPT, all four configurations were tested: GPT-5-mini-2025-08-07[1] (GPT-5-mini) and GPT-5-2025-08-07[2] (GPT-5-large), each with minimal and high *reasoning effort* settings. For Mistral, preliminary tests of smaller models (mistral-small-2506 and magistral-small-2507) showed insufficient instruction-following capabilities, leading to their exclusion. The final evaluation therefore used only Mistral-large-2411[3] (Mistral) and Magistral-medium-2507, resulting in six model configurations total. All experiments used snapshot versions for reproducibility with default temperature settings.

Dataset and Processing The evaluation used the complete ACTER dataset, consisting of twelve corpora in four domains (Corruption (corp), Equitation (equi), Heart Failure (htfl), Wind Energy (wind)) and three languages (English (EN), French (FR), Dutch (NL)). Documents were processed at the document level when possible, unlike the sentence-level processing in previous studies. For the largest documents, the same splits used during the human annotation process (Rigouts Terryn et al., 2020b) were applied. Only when this failed due to context window limitations was a text split into additional equal parts using simple character-based division, which may have caused errors when splitting mid-term. Further details are reported in Section 5.

Prompting Strategy Concise and elaborate prompts were developed based on ACTER's human annotation guidelines and refined using GPT-

[1] https://platform.openai.com/docs/models/gpt-5-mini

[2] https://platform.openai.com/docs/models/gpt-5

[3] https://docs.mistral.ai/getting-started/models/models_overview/

5's *prompt optimizer*[4]. Given the subjectivity of term extraction and moderate inter-annotator agreement, prompts were based on the specific guidelines used by human annotators to ensure models had access to the same conceptual framework rather than generic terminology definitions. The elaborate prompt included detailed definitions of lexicon-specificity and domain-specificity, extraction guidelines, and term category examples. Both prompts requested TSV output with four fields: term, label (Specific/Common/Out-of-Domain Term, optionally Named Entity), confidence score (50-100; terms with lower confidence should be discarded), and optional remarks for information deemed crucial for terminology management. Prompts and examples can be found in the Appendix. Code and prompts are available at https://github.com/AylaRT/llmate.

Experimental Design Preliminary experiments on Heart Failure (English and Dutch) tested multiple variables to optimise the final configuration: model size (small and large) and type (standard or reasoning), prompt elaborateness (elaborate or concise), example provision (zero-shot, few-shot in-domain, few-shot out-of-domain), term scope (including/excluding Named Entities), and prompt language (same as data or different). Examples were always in the same language as the data. Examples were not optimised, as earlier research found minimal impact (Chun et al., 2025). Instead, all out-of-domain examples used the same paragraph of text in the fitness domain (equivalent across all three languages). In-domain examples were paragraphs from real documents similar to, but not part of, those in the corpus. Examples always included at least one term with remarks and example annotations for each label. Based on preliminary results, optimal settings were selected for full evaluation across all domains and languages. Each document was processed in a new conversation. Due to cost constraints, each experiment was run once. Since temperature is no longer a controllable parameter in GPT-5, default settings were maintained.

Evaluation Performance was measured using precision, recall, and F1 score through case-insensitive matching. Evaluation was conducted at both document and corpus levels (with deduplication), analysing hallucination rates, confidence score reliability, performance for different types of terms, and remarks.

[4] https://platform.openai.com/chat/edit?models=gpt-5&optimize=true

4. Preliminary Experiments

To identify optimal configurations for full-scale experiments, preliminary tests on the English and Dutch heart failure corpora systematically evaluated the following variables:

- Model configurations: size (small/large) and type (standard/reasoning)
- Prompting: elaborateness (concise/elaborate) and language
- Examples: zero-shot, few-shot in-domain/out-of-domain
- Term scope: with/without Named Entities

For each corpus, this meant a total of 144 experiments. Given the 190 (EN) and 175 (NL) documents per corpus and the fact that a few documents had to be split in certain settings, this resulted in a total of 52,574 runs.

4.1. Baseline Performance and Task Complexity

To contextualize the preliminary findings, Table 1 presents state-of-the-art results from previous studies on the same heart failure corpus. The supervised systems (including recurrent neural networks (RNN) and random forest classifiers (RFC)) were trained on the three other domains in the ACTER corpus in the same language and tested on the heart failure domain. The rule-based system is based on Universal Dependencies (UD) grammars. The final study uses few-shot prompting with Llama-2-Chat, reporting results for the most advanced setup: monolingual domain transfer with self-verification and explanation, using explicit in-domain examples.

Notably, human inter-annotator agreement on this task averages only .59 F1 score across English domains (Rigouts Terryn et al., 2020b), indicating substantial subjectivity in term identification. Moreover, all studies report considerable variation (differences of 0.20 or more) across languages and domains (e.g., higher scores in Dutch than in English, likely due to compounding rules). This ceiling effect, combined with inherent cross-domain and cross-linguistic variation, suggests the task's ambiguity limits achievable performance. The preliminary experiments therefore targeted performance in the .50–.65 F1 score range as both realistic and competitive with existing approaches, while acknowledging that optimal configurations may differ by language and domain.

4.2. Variable testing

Model Size and Type Table 2 shows all models achieve F1 scores at the lower end of the expected range (.47-.60), with only 5-6 point differ-

model	type	EN			NL		
		p	r	f1	p	r	f1
Rigouts Terryn et al. (2021)	features-based RFC	.53	.37	.43	.61	.50	.55
Rigouts Terryn et al. (2022)	RNN with BERT embeddings	.52	.63	.58	.59	.70	.64
Marciniak et al. (2023)	UD dependency rules + C-value ranking	.14	.20	.17	.07	.09	.08
Marciniak et al. (2025)	token classification with fine-tuned RoBERTa	.68	.51	.58	-	-	-
Tran et al. (2025)	few-shot prompting with Llama-2-Chat	.55	.52	.54	.53	.48	.50

Table 1: Precision (p), recall (r) and F1 score ($f1$) obtained by previous studies on the EN and NL parts of the ACTER heart failure corpus, specifically the version including named entities.

model	EN			NL		
	p	r	f1	p	r	f1
standard						
GPT-5-large	.39	.69	.50	.50	.70	.58
GPT-5-mini	.41	.66	.50	.50	.63	.55
Mistral	.38	.64	.48	.50	.64	.56
average	**.39**	**.67**	**.49**	**.50**	**.66**	**.57**
reasoning						
GPT-5-large	.41	.75	.52	.52	.72	.60
GPT-5-mini	.36	.70	.47	.46	.69	.55
Magistral	.37	.63	.47	.51	.62	.56
average	**.38**	**.69**	**.49**	**.50**	**.68**	**.57**
average (all)	**.39**	**.68**	**.49**	**.50**	**.67**	**.57**

Table 2: Precision, recall, and F1 scores for preliminary experiments, averaged over prompting strategies; evaluated against gold data including named entities.

ences between best and worst performers per corpus. Larger GPT models very marginally outperform smaller variants, and Mistral models perform on par with GPT-5-mini. The reasoning variants do not perform better in the cases of GPT-5-mini and Magistral, and even perform worse for the English corpus. Yet, for GPT-5-large, higher reasoning effort does improve results slightly. All models favour recall over precision. Given the minimal performance differences and the centrality of these variables, **all models are retained** for final experiments.

Named Entities Consistent with previous studies, including named entities improved F1 scores by 0.02 on average. This is unsurprising given NE extraction is less ambiguous than term extraction. Final experiments **include named entities** to maintain comparability with prior work.

Prompting English prompts very marginally outperformed Dutch prompts for both English (+0.007 F1 score) and Dutch (+0.001 F1 score) data, though the differences are negligible. Elaborate prompts showed clearer benefits over concise variants (+0.024 F1 score), particularly for reasoning models (+0.035 vs +0.013 for standard models). Final experiments use **elaborate English prompts only**.

Examples Surprisingly, few-shot examples provided minimal benefit: average F1 scores were .53 for all conditions (few-shot in-domain: +0.004, few-shot out-of-domain: +0.002, zero-shot: baseline). Given this negligible impact, final experiments use **both zero-shot and few-shot out-of-domain prompting** to ensure identical prompts across domains.

5. Results

For the final experiments, all 6 model configurations and 2 prompting types (zero-shot and few-shot out-of-domain) were tested on all corpora (4 domains in 3 languages), totalling 144 experiments. The corpora in other domains contain fewer but larger documents, which sometimes had to be split for processing. This resulted in a total of 10,933 runs. Only 32 of these runs required splitting a document (into 2-6 parts) to fit within context window limits, occurring exclusively for Magistral, mostly in the wind energy and corruption domains, demonstrating that document-level processing fits within current LLM context limits, though the impact of context length on recall (e.g., the 'lost in the middle' phenomenon) requires further study.

Zero-Shot vs. Few-Shot The limited impact of few-shot prompting found in preliminary experiments persists across all corpora: few-shot prompting improves F1 scores by only +0.006 on average. The benefit is marginally larger for reasoning models (+0.008) than standard models (+0.004), with the largest effect for GPT-5-large (+0.02). Domain-specific variation is small: few-shot prompting helps most for corruption (+0.012) but slightly harms performance for heart failure (-0.003). Given these consistently small differences, subsequent analyses report **averages over both conditions**, which also helps mitigate variability from single runs with non-zero temperature.

	CORP			EQUI			HTFL			WIND			AVG
model	**EN**	**FR**	**NL**	**EN**	**FR**	**NL**	**EN**	**FR**	**NL**	**EN**	**FR**	**NL**	**All**
standard													
GPT-5-large	.40	.40	.42	.60	.54	.70	.50	.54	.59	.48	.39	.52	.51
GPT-5-mini	.36	.33	.33	.55	.49	.62	.51	.52	.57	.43	.35	.47	.46
Mistral	.35	.34	.39	.50	.46	.60	.52	.50	.57	.44	.33	.45	.45
average (standard)	**.37**	**.36**	**.38**	**.55**	**.49**	**.64**	**.51**	**.52**	**.58**	**.45**	**.35**	**.48**	**.47**
reasoning													
GPT-5-large	.42	.42	.46	.63	.54	.72	.56	.54	.63	.52	.43	.58	.54
GPT-5-mini	.40	.38	.42	.58	.47	.65	.52	.48	.58	.47	.34	.49	.48
Magistral	.37	.36	.39	.49	.47	.61	.51	.50	.57	.42	.33	.47	.46
average (reasoning)	**.40**	**.39**	**.42**	**.57**	**.49**	**.66**	**.53**	**.51**	**.59**	**.47**	**.37**	**.51**	**.49**
average (all)	**.38**	**.37**	**.40**	**.56**	**.49**	**.65**	**.52**	**.51**	**.58**	**.46**	**.36**	**.50**	**.48**

Table 3: F1 scores for all models across domains and languages, averaged over prompting conditions. The final column shows macro averages over all domains and languages.

Results per Corpus per Model Table 3 shows F1 scores for all model configurations per corpus. Consistent with previous research, scores are highest for equitation (mean .57) and heart failure (mean .54), followed by wind energy (mean .44) and corruption (mean .39). These differences reflect annotation difficulty and the ease of distinguishing terms from general words (Rigouts Terryn et al., 2020b). Across languages, Dutch achieves the highest scores (.53), followed by English (.48) and French (.44), likely due to compounding rules: Dutch single-word compounds (e.g., 'ejectiefractie') correspond to multi-word terms in English ('ejection fraction') and French ('fraction d'éjection'), making extraction easier as all components must be correct in multi-word terms.

Model differences are small compared to cross-domain and cross-linguistic variation: only 0.08 separates the best and worst performing models overall. Model size matters most, with GPT-5-large outperforming all others regardless of reasoning effort. Mistral and GPT-5-mini perform similarly (within 0.05 points), except Mistral scores 0.06 higher on Dutch corruption and GPT-5-mini with reasoning scores 0.08 higher on English equitation. Reasoning variants average only +0.02 better than standard versions (+0.01 for Magistral, +0.02 for GPT-5-mini, +0.03 for GPT-5-large), with the largest single improvement being +0.09 for GPT-5-mini with reasoning on Dutch corruption. In seven cases, standard models outperform their reasoning counterparts.

Precision and recall patterns vary by corpus. Heart failure shows persistently higher recall (.70) than precision (.44), matching preliminary experiments. However, corruption shows the reverse: precision (.44) exceeds recall (.36). For equitation and wind energy, the pattern depends on language, with precision highest in English and recall highest in French.

Scores per Term Label Gold standard terms carry one of four labels: Specific Term (lexicon- and domain-specific), Common Term (domain-specific only), Out-of-Domain Term (lexicon-specific only), or Named Entity (proper name). Named entities achieve the highest F1 scores across all models and corpora, as they are less ambiguous to detect. Among term types, Specific Terms score highest (.44), followed by Common Terms (.22) and Out-of-Domain Terms (.10). This distribution aligns with expectations: Specific Terms represent the primary target of most terminology extraction applications and the strictest definition of terminology, whereas Common and Out-of-Domain Terms are more debatable. The pattern remains consistent across corpora and models. Model-assigned labels match the gold standard well: F1 scores drop by only 12 points on average when requiring strict matches for both term and label (consistent across all model configurations). However, model-assigned labels show a concerning pattern with a systematic overapplication of the OoD Term label among false positives, which is examined in detail in the qualitative analysis (Section 6).

Hallucinations Hallucinated terms were detected through case-insensitive exact string matching between extracted candidates and source text: candidates not found in the source were marked as hallucinations. Only 2% of candidate terms on average received this label, but with clear differences between models. For both GPT models, high reasoning effort notably reduced hallucinations: GPT-5-mini dropped from 3.7% to 0.7%, and GPT-5-large dropped from 1.8% to 0.4%. In contrast, Mistral (3.1%) and Magistral (3.4%) both showed high hallucination rates. Spearman correlation coefficients show moderate negative correlations between hallucination rate and recall ($\rho = -0.54$) and F1 score ($\rho = -0.49$), but only weak negative correlation with precision ($\rho = -0.18$).

Confidence Scores All models were prompted to include confidence scores between 50 and 100 for each extracted candidate term, excluding candidates below this threshold. This mirrors current ATE tools that sort candidate terms (often by termhood scores) to facilitate manual validation. All models assign higher confidence scores to true positives than false positives, with an average difference of 6 points and the highest difference for GPT-5-large standard (8.1 points). Though modest, this difference may prove useful for validation workflows. Average confidence scores show negligible correlation with F1 scores ($\rho = -0.05$).

Remarks Models were prompted to provide free-text remarks per candidate term "if truly crucial for terminology management". On average, 7% of terms received remarks, with large differences between models. Mistral and all reasoning models rarely added remarks (0.8%-3.2%), while standard GPT models added substantially more: GPT-5-large for 9.9% of candidates and GPT-5-mini for 28.1%. The percentage of remarks shows no correlation with performance ($\rho = 0.01$) but weak negative correlation with hallucination rates ($\rho = -0.05$).

6. Qualitative Analysis

Error analysis focused on the heart failure corpus using the best-performing model (GPT-5-large with high reasoning effort) to ensure lasting relevance, using the configuration from final experiments (elaborate English prompt, few-shot out-of-domain, including named entities). Though examples are from English for clarity, parallel analyses of Dutch and French results showed similar patterns.

For this corpus, there are 2,581 unique gold standard terms across 190 documents. GPT-5-large extracted 4,964 unique candidate terms, of which 2,109 were correct, yielding recall of .82, precision of .43, and F1 score of .56.

6.1. Hallucinations

In the English experiments, only 31 candidate terms were automatically classified as hallucinations (0.8% of all false positives; 18 in French, 15 in Dutch). However, upon further analysis, only one represents a true fabrication: 'out-pient follow-up' extracted instead of 'out-patient follow-up'. The remaining 30 fall into five categories:

- Ellipsis resolution (15 instances): e.g., 'heart failure with preserved ejection fraction' extracted from 'heart failure with reduced and preserved ejection fraction'
- Punctuation removal (11 instances): e.g., 'NT-proBNP assays' from '(NT-proBNP) assays'; 'congestion-like symptoms' from "congestion-like' symptoms'
- Parenthetical removal (4 instances): 'adverse cardiovascular events' from 'adverse cardiovascular (CV) events'
- Modifier variation (3 instances): from 'prospective randomized multicenter trial', extracted 'prospective randomized trial', 'randomized trial', and 'prospective trial'
- Number format (1 instance): '30-day mortality' from 'thirty-day mortality'

Multiple phenomena can co-occur: from 'genetic (AUC = 0·67) or the clinical (AUC = 0·69) models', the model extracts 'genetic model', resolving the ellipsis, removing parentheticals, and converting from plural to singular. thus, only a single extraction from GPT-5-large with high reasoning effort represents a real fabrication, and it is merely a typographical error ('out-pient follow-up').

Smaller models show slightly more problematic patterns: extracting abbreviations not present in the text ('LVEF' when only 'left ventricular ejection fraction' appears), though remarks sometimes flag these (e.g., 'abbreviation'). These models also normalise spelling variants (extracting 'follow-up' when the text contains 'follow up').

Similar patterns appear in other languages. In Dutch, extractions classified as hallucinations include 'intravenuze medicatie' (misspelling the source text's correct 'intraveneuze medicatie'), falling into the same normalisation category as English examples. However, one Dutch extraction represents a more serious error: 'afferente arteriole' when the text stated 'efferente arteriole'. Both terms exist but denote different anatomical structures, making this substitution potentially misleading unlike simple misspellings or formatting variations. Other Dutch hallucinations fall into the same categories as English ones: ellipsis resolution, punctuation removal, and orthographic normalisation.

6.2. False Positives

Given the volume of false positives (2,855) and the inherent subjectivity of term boundaries, an exhaustive manual analysis was unfeasible. Instead, a qualitative examination identified four primary categories of false positives:

1. **Overly general expressions**: Single words or short phrases lacking sufficient specificity (e.g., 'guidelines', 'management', 'age', 'exacerbation', 'smoking', 'studies', 'research', 'kg', 'ml', 'water')
2. **Coordinated term phrases**: Expressions combining multiple distinct terms through conjunction or disjunction (e.g., 'signs and symptoms', 'death or HF hospitalisation', 'a- and b-natriuretic

peptides', 'tachy- or bradyarrhythmias', 'N- and C-terminally processed forms')
3. **Terms with non-terminological modifiers**: Domain-specific terms qualified by generic quantifiers or evaluative adjectives (e.g., 'total adiponectin', 'strong acid', 'worsened prognosis', 'symptoms of heart failure', 'reductions in LVEF')
4. **Descriptive clinical phrases**: Phrases describing patient populations or clinical contexts rather than reified terminological concepts (e.g., 'hospitalised for HF', 'elderly patients', 'patients with heart failure', 'healthy volunteers', 'worsening of heart failure')

However, many false positives do not fall into these problematic categories. Examples include 'clinical Doppler echocardiography' (the gold standard includes only 'clinical' and 'Doppler echocardiography' separately), 'iron deficiency', 'endothelial repair', 'receiver operating characteristic curve', and 'anterior leaflet opening'.

A pattern emerges in model-assigned labels for this experiment: 55% of false positives were labelled Specific Terms, 16% Common Terms, 25% Out-of-Domain Terms, and 4% Named Entities. The proportion of OoD labels is notably high given OoD terms comprise only 6% of the gold standard for this corpus. Examination reveals the model applies the OoD label to three distinct cases: valid OoD terms (e.g., 'anova', 'baseline covariates'), terms that should be Common Terms (e.g., 'blood loss' could be considered domain-specific but not lexicon-specific), and clear extraction errors (e.g., 'beagle dogs', which is neither lexicon- nor domain-specific). This suggests the model struggles with the three-way distinction between Specific, Common, and OoD terms, using OoD as a default for technical language of uncertain domain relevance.

To investigate whether this represents a systematic issue, label distributions were examined across all final experiments. The over-application of the OoD label proves systematic: 36% of false positives are labelled OoD on average across all corpora (50% for Mistral and Magistral, 30% for GPT models), while OoD terms comprise only 0.1-6.8% of gold standards.

6.3. False Negatives

The 472 false negatives are distributed among the four term labels in proportion that is much more similar to the gold standard. They fall into three main categories. First, non-nominal terms are systematically under-extracted: adjectives account for 106 false negatives (22%), alongside 24 verbs and 15 adverbs. Examples include 'atrial', 'venous', 'coronary', 'ischaemic', 'diastolic', 'myocardial' (adjectives); 'prescribed', 'followed-up', 'correlate', 'diagnoses' (verbs); and 'orally', 'medically', 'non-invasively' (adverbs). Adjectives are reliably extracted when embedded in multi-word noun phrases ('left ventricular ejection fraction', 'venous pressure') but missed in isolation.

Second, named entities among false negatives predominantly comprise non-domain-specific proper names: author names, geographical locations, and hospital names. The gold standard annotation guidelines required exhaustive named entity extraction, whereas the prompt simply mentioned including named entities. The models prioritised domain-specific named entities (drug names, procedure names) over incidental proper names.

Third, complex multi-word expressions show boundary judgement differences. Examples include 'exercise-induced pulmonary artery systolic pressure', 'in vivo coronary perfusion', and 'fully magnetically levitated centrifugal-flow chronic LVAS'. The gold standard includes 'LV diastolic stiffness' where the model extracted 'increased LV diastolic stiffness'; 'erythropoiesis-stimulating agent' versus 'erythropoiesis-stimulating agent therapy'; 'proteomics' versus 'quantitative proteomics'. When multi-word expressions contain several domain-specific elements, the model tends to extract maximal phrases, whereas the gold standard often atomises them into smaller components.

6.4. Remarks

For the English heart failure experiment, GPT-5-large with high reasoning effort added 35 remarks, categorised as follows:

1. **Preferred variant indication** (10): The model extracts the term as it occurs but notes a preferred variant. Examples: for 'hf-ref', remark 'preferred modern abbreviation is HFrEF'; for 'pulmonary resistance', remark 'usually referred to as pulmonary vascular resistance'; for 'right atrium mass', remark 'nonstandard; preferred term is right atrial mass'
2. **Alternative variant** (3): Similar to the first category, without indicating preference. Example: for 'implantable cardiac defibrillator', remark 'variant of implantable cardioverter defibrillator'
3. **Misspelling flagging** (9): The model extracts misspellings as they occur but flags them. Examples: for 'bisphophonates', remark 'misspelling of bisphosphonates'; for 'pdrelated complications', remark 'preferred: PD-related complications (typo in text)'
4. **Abbreviation explanation** (8): Mostly when the full form is not explicitly mentioned. Examples: for 'HR', remark 'Used as hazard ratio, not heart rate'; for 'RAS', remark 'stands for renal artery stenosis in this text'
5. **Other explanations** (3): Examples: for 'aldactone', remark 'brand name of spironolactone'; for

'significant' and 'significantly', remark 'statistical usage implied'

6. **Singular form preference** (2): Example: for 'ejection fractions', remark 'preferred singular is ejection fraction (EF)'

7. Discussion

he central finding of this evaluation is not merely that LLMs achieve strong performance at term extraction, but rather that ATE for high-resource languages is approaching a performance ceiling imposed by the task's inherent ambiguity rather than by model limitations. All six configurations reach F1 scores of .36–.72, comparable to previous supervised and few-shot approaches. The gap between models is narrow: only 0.08 F1 separates the best and worst configurations, with model size mattering more than reasoning capabilities (+0.02 F1 on average for reasoning variants). These findings must be qualified, since smaller Mistral models were excluded after preliminary tests, meaning results reflect only current large-scale, advanced models. Still, the convergence of scores across architectures, combined with the fact that individual corpus scores range from well below to above human inter-annotator agreement (.59 F1 across English domains; Rigouts Terryn et al., 2020b), suggests that remaining performance variation is driven more by task and language characteristics than by model choice.

The qualitative analysis supports this interpretation. Many errors reflect annotation framework challenges rather than extraction failures. Boundary judgement differences for multi-word expressions are common: the gold standard may annotate 'Doppler echocardiography' and 'clinical' separately, while the model extracts 'clinical Doppler echocardiography' as a single term, with neither segmentation being inherently wrong. More broadly, numerous extractions classified as errors represent defensible terminological choices that resemble inter-annotator disagreement rather than system mistakes. What automated checks flag as "hallucinations" (i.e., sequence not found in corpus) are predominantly valid normalisations (ellipsis resolution, punctuation standardisation), with only one genuine fabrication among 31 flagged cases. With only moderate human agreement, gold standards represent one valid interpretation rather than objective truth, yet string-matching treats any deviation as error. Current evaluation methodology therefore partly misrepresents model performance, and metrics likely underestimate actual utility. Models also demonstrate valuable capabilities beyond extraction itself: confidence scores show modest discrimination (6-point average difference between true and false positives), and generated remarks include potentially useful terminological information, including variant identification and misspelling detection.

Nevertheless, LLM-based ATE does have clear limitations. Overly general single words ('age', 'water', 'management') and clear mistakes ('beagle dogs' as a term in the heart failure corpus) are consistently over-extracted. The systematic over-application of the OoD Term label (36% of false positives vs. 0.1–6.8% in gold standards) indicates difficulty with fine-grained semantic categorisation, not merely annotation disagreement. Similarly, under-extraction of non-nominal terms (22% of false negatives are adjectives) represents a real limitation. These are substantive weaknesses, though they also represent practically useful findings: the systematic patterns identified (over-extraction of general terms, under-extraction of non-nominal terms, over-application of the OoD label) could inform more targeted prompts for users for whom these distinctions matter most, though this strategy was not tested in the current study.

Given this situation, incrementally optimising F1 scores against a fixed gold standard offers limited insight; evaluation approaches that distinguish extraction errors from defensible boundary judgements would be more informative. ATE was always intended as one step in a terminology management pipeline, not as an end in itself. The convergence of LLM performance with both supervised approaches and inter-annotator agreement suggests that, for high-resource languages and well-represented domains, basic term extraction may no longer be the bottleneck in such pipelines. Whether the remaining systematic weaknesses propagate meaningfully into downstream tasks remains to be tested, but the research focus should shift accordingly toward tasks such as definition generation (San Martín, 2024), cross-lingual equivalent identification (Heinisch, 2025), and ontology construction (Babaei Giglou et al., 2023). These tasks better leverage the models' reasoning capabilities and present open challenges for future research.

8. Limitations

The qualitative analysis examined only one corpus (heart failure) with one model configuration (GPT-5-large with high reasoning effort), though quantitative patterns proved consistent across corpora and spot-checking showed no significant differences in other corpora. Heart failure was selected because it is the standard test domain in previous ACTER-based studies, facilitating comparison with prior work. Potential training data contamination cannot be ruled out, though substantial misalignment with gold standards suggests this is not a significant issue.

Each experiment was run only once, and GPT-5

no longer exposes temperature as a controllable parameter, introducing unquantifiable variance. Averaging over zero-shot and few-shot conditions partially mitigates but does not eliminate this issue. The evaluation prioritised breadth of experimental settings over breadth of model families, covering only GPT and Mistral, which limits the generalisability of findings across architectures.

This work should not be interpreted as advocating LLM adoption for term extraction. Supervised fine-tuned models (e.g., RNNs, as shown in Related Research) achieve comparable F1 scores using a fraction of the computational resources. Where relevant training data exist, small dedicated models remain far more efficient, both financially and environmentally, than deploying large-scale LLMs for this narrow task. This evaluation's value lies in establishing what current LLMs can achieve, and in demonstrating capabilities beyond basic extraction (confidence scoring, remarks, terminology management support) that point toward more appropriate uses of these models.

9. Conclusions

Evaluated across four domains and three languages, current large-scale LLMs achieve term extraction performance approaching human inter-annotator agreement. Many apparent errors reflect the ambiguity of term boundaries rather than genuine extraction failures, suggesting that standard evaluation metrics underestimate actual utility. Real limitations remain, particularly in fine-grained categorisation and non-nominal term extraction, but these point toward targeted improvements. For high-resource languages, basic ATE may no longer be the bottleneck in terminology management pipelines. Future work should focus on more nuanced evaluation methods, and move toward downstream tasks such as definition generation, cross-lingual equivalent identification, and ontology construction.

10. Acknowledgements

This project was undertaken thanks to funding from IVADO and the Canada First Research Excellence Fund.

11. Bibliographical References

Nikita Astrakhantsev, D. Fedorenko, and D. Yu. Turdakov. 2015. Methods for Automatic Term Recognition in Domain-specific Text Collections: A Survey. *Programming and Computer Software*, 41(6):336–349.

Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. 2023. LLMs4OL: Large Language Models for Ontology Learning. In *The Semantic Web – ISWC 2023*, pages 408–427, Cham. Springer Nature Switzerland.

Shubhanker Banerjee, Bharathi Raja Chakravarthi, and John Philip McCrae. 2024. Large Language Models for Few-Shot Automatic Term Extraction. In Amon Rapp, Luigi Di Caro, Farid Meziane, and Vijayan Sugumaran, editors, *Natural Language Processing and Information Systems. NLDB 2024*, volume 14762 of *Lecture Notes in Computer Science*, pages 137–150. Springer Nature Switzerland, Cham.

Julien Breton, Mokhtar Mokhtar Billami, Max Chevalier, Ha Thanh Nguyen, Ken Satoh, Cassia Trojahn, and May Myo Zin. 2025. Leveraging LLMs for legal terms extraction with limited annotated data. *Artificial Intelligence and Law*.

Yongchan Chun, Minhyuk Kim, Dongjun Kim, Chanjun Park, and Heuiseok Lim. 2025. Enhancing Automatic Term Extraction with Large Language Models via Syntactic Retrieval. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9916–9926, Vienna, Austria. Association for Computational Linguistics.

Orphée De Clercq, Marjan Van de Kauter, Els Lefever, and Veronique Hoste. 2015. LT3: Applying Hybrid Terminology Extraction to Aspect-Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 719–724, Denver, Colorado. Association for Computational Linguistics.

Patrick Drouin. 2003. Term Extraction Using Non-Technical Corpora as a Point of Leverage. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 9(1):99–115.

Katerina T. Frantzi and Sophia Ananiadou. 1999. The C-value/NC-value Domain-independent Method for Multi-word Term Extraction. *Journal of Natural Language Processing*, 6(3):145–179.

Amir Hazem, Merieme Bouhandi, Florian Boudin, and Beatrice Daille. 2022. Cross-lingual and Cross-domain Transfer Learning for Automatic Term Extraction from Low Resource Data. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 648–662, Marseille, France. European Language Resources Association.

Barbara Heinisch. 2025. Large language models for terminology work: A question of the right

prompt? *Journal for Language Technology and Computational Linguistics*, 38(2):13–30.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.

Els Lefever. 2016. A Hybrid Approach to Domain-Independent Taxonomy Learning. *Applied Ontology*, 11(3):255–278.

Lieve Macken, Els Lefever, and Véronique Hoste. 2013. TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 19(1):1–30.

Małgorzata Marciniak, Piotr Rychlik, and Agnieszka Mykowiecka. 2023. TermoUD — a Language-independent Terminology Extraction Tool. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 178–186, Dubrovnik, Croatia. Association for Computational Linguistics.

Małgorzata Marciniak, Piotr Rychlik, and Agnieszka Mykowiecka. 2025. Do transformer-based token classification methods solve the problem of terminology extraction? *Natural Language Processing*, pages 1–26.

Ayla Rigouts Terryn, Véronique Hoste, Patrick Drouin, and Els Lefever. 2020a. TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset. In *Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020)*, pages 85–94, Marseille, France. European Language Resources Association.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2020b. In No Uncertain Terms: A Dataset for Monolingual and Multilingual Automatic Term Extraction from Comparable Corpora. *Language Resources and Evaluation*, 54(2):385–418.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2022. Tagging Terms in Text: A Supervised Sequential Labelling Approach to Automatic Term Extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 28(1).

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2021. HAMLET: Hybrid Adaptable Machine Learning approach to Extract Terminology. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 27(2):40.

Antonio San Martín. 2024. What Generative Artificial Intelligence Means for Terminological Definitions.

Hanh Thi-Hong Tran, Carlos-Emiliano González-Gallardo, Antoine Doucet, and Senja Pollak. 2025. LlamATE: Automated terminology extraction using large-scale generative language models. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 31(1):5–36.

Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Julien Delaunay, Antoine Doucet, and Senja Pollak. 2024. Is Prompting What Term Extraction Needs? In Elmar Nöth, Aleš Horák, and Petr Sojka, editors, *Text, Speech, and Dialogue*, volume 15048, pages 17–29. Springer Nature Switzerland.

A. Prompts and Examples

This appendix presents the two prompt templates used in the experiments, followed by the in-domain and out-of-domain examples. In both templates, `[DOMAIN]` is replaced with the target domain (e.g., "heart failure"), `[TEXT]` with the text to be processed, and `[EXAMPLE]` with either nothing (zero-shot) or a worked example consisting of a short paragraph and its expected TSV output (few-shot out-of-domain). The French and Dutch prompts and examples can be found in the shared repository.

A.1. Concise Prompt Template

```
Extract all terminology from the
provided text in the domain of
[DOMAIN], respecting the following
guidelines:
- List one term per line and
deduplicate case-insensitively.
- Do not lemmatise or otherwise
normalise.

# Output format
<term>\t<term_label>\t<confidence_score>
\t<remarks>

Field definitions:
```

```
- <term> = substring that occurs
verbatim in the text (no quotation
marks).
- <term_label> = {Specific Term
(lexicon- and domain-specific), Common
Term (only domain-specific), OoD Term
(only lexicon-specific,
Out-of-Domain)[, Named Entity (proper
name)]}
- <confidence_score> = integer 50-100
inclusive (lower confidence terms
should be excluded).
- <remarks> = - by default; add remark
(<=20 words) only if it is truly
crucial for terminology
  management (when unsure, stick to -).

Respect the following:
- Plain TSV only: no markdown, no
bulleted or numbered lists
- Actual tab characters as separators,
exactly three tabs per line.
- No other separators or styling (no
commas, pipes, quotes, **).
- Output only TSV rows (exactly 4
tab-separated fields per line); no text
before or after.

[EXAMPLE]

Text: [TEXT]
```

A.2. Elaborate Prompt Template

```
# Objective
Extract all terminology from the
provided text in the domain of
[DOMAIN], knowing that a term can be
defined as a word or phrase used to
express a concept in a specialised
domain.

# Background
Termhood is determined based on two
main criteria:
1. "Lexicon-specificity" indicates
whether a lexical unit is part of
common language or known primarily by
specialists.
2. "Domain-specificity" indicates
whether a lexical unit is relevant to
the targeted domain.

These criteria define three types of
terms:
1. Specific Terms: both lexicon- and
domain-specific; vocabulary primarily
known by specialists in the relevant
domain.
2. Out-of-Domain (OoD) Terms:
lexicon-specific but not
domain-specific (not relevant to the
subject of the text).
3. Common Terms: not lexicon-specific
(part of common vocabulary) but
domain-specific (relevant to the
subject of the text). A lexical unit
that is neither lexicon-specific nor
domain-specific is not a term.

For instance, when extracting terms in
the domain
of fitness:
- "hypertrophy" and "heart rate
variability" are Specific Terms
- "dumbbell" and "heart rate" are
Common Terms
- "p-value" and "NLP" are Out-of-Domain
Terms
- "Matthew" and "Planet Fitness" are
Named Entities

# Term Identification and Labelling
Guidelines
- Terms have no minimum or maximum
length.
- Terms are not limited to nouns or
noun phrases: verbs, adverbs,
adjectives, etc. can also be terms.
- A practical heuristic to decide
between Common Term or Specific Term:
would a popular magazine or newspaper
use the expression without further
explanation? If not, it is likely
lexicon-specific.
- Label terms according to their
meaning in this text. Common words can
become terms when used with a domain
meaning. For example, in statistics,
"significant" means having a p-value
below alpha; in that statistical
context the word is lexicon-specific,
so label Specific Term (or OoD Term if
the document's domain is not
statistics). In general, in
non-statistical usage, "significant"
just means "notable" and is not a term.

# Extraction Guidelines
- List one term per line and
deduplicate case-insensitively.
- Do not lemmatise or otherwise
normalise.
- Nested candidates: emit both the
longer term and any contained shorter
term if the shorter term is a
recognised standalone term in the
domain (even if it does not occur
independently in this text); otherwise
emit only the longer term.
- Do not annotate below word level. Do
not split compounds into morphemes
(e.g., do not split "football" into
"foot" and "ball"). Hyphens and
apostrophes may delimit multiword-like
expressions (e.g., part-of-speech):
```

```
extract sub-components only if they
also occur as standalone terms
elsewhere in the text.
- Be consistent: the same term should
receive the same label within the same
domain (unless polysemy requires
otherwise). Variants of a term
(abbreviations, derivations, adjectival
forms, etc.) should receive the same
label.
- Apply the same rules to units of
measurement, misspellings, and
adjectival/adverbial modifiers; do not
carve out exceptions.

# Output Format
Format the output as TSV only (no
header,
explanations, or extra text; if no
terms, output
a single -):
<term>\t<term_label>\t<confidence_score>
\t<remarks>

- <term> = substring that occurs
verbatim in text (no quotation marks)
- <term_label> = {Specific Term, Common
Term, OoD Term[, Named Entity]} (pick
only one)
- <confidence_score> = integer 50-100
inclusive (lower confidence terms
should be excluded).
- <remarks> = - by default; add remark
(<=20 words) only if it is truly
crucial for terminology management
(when unsure, stick to -).

Respect the following:
- Plain TSV only: no markdown, no
bulleted or numbered lists
- Actual tab characters as separators,
exactly three tabs per line.
- No other separators or styling (no
commas, pipes, quotes, **).
- Output only TSV rows (exactly 4
tab-separated fields per line); no text
before or after.

# Stop Conditions
Return only when all terms are
extracted, classified, and formatted as
required, or a single dash - if no
terms are identified.

[EXAMPLE]

# Text
[TEXT]
```

A.3. In-Domain Example (Heart Failure)

```
# Example Text
During the follow-up period, there were
94 deaths (46.3%). Deceased patients
were older (p < 0.001), commonly in New
York Heart Association (NYHA) stage III
or up (p < 0.001), had lower 6-minute
walk distances (p = 0.014), higher
prevalence of type 2 diabetes mellitus
(T2DM) (p = 0.018), raised creatinine
(p = 0.001), and lower hemoglobin (p =
0.004).

# Example Output
follow-up Common Term 90 -
deaths Common Term 81 -
deceased Common Term 52 -
patients Common Term 88 -
p OoD Term 89 -
New York Named Entity 95 -
New York Heart Association Named Entity
99 -
NYHA Named Entity 99 -
heart Common Term 96 -
New York Heart Association stage III
Specific Term 87 preferred term would
be "class" instead of "stage"
6-minute walk distances Specific Term
77 -
prevalence Common Term 79 -
type 2 diabetes mellitus Specific Term
99 -
type 2 diabetes  Specific Term 99 -
diabetes Common Term 97 -
diabetes mellitus Specific Term 99 -
T2DM Specific Term 99 -
creatinine Specific Term 99 -
hemoglobin Specific Term 99 -
```

A.4. Out-of-Domain Example

```
# Example Text
During the warm-up, coach Mike explains
progressive overload and asks us to
estimate our one-repetition maximum. We
then do high-intensity interval
training (HIIT) to improve VO2 max. All
strength training sessions start with a
warm-up and end with a warm-down. I
track everything with my Garmin, though
the accelerometer is broken.

# Example Output
warm-up Common Term 90 -
coach Common Term 60 -
Mike Named Entity 95 -
progressive overload Specific Term 98 -
one-repetition maximum Specific Term 99
-
repetition Specific Term 75 -
high-intensity Common Term 55 -
high-intensity interval training
Specific Term 99 -
interval training Common Term 70 -
training Common Term 69 -
```

HIIT Specific Term 99 –
VO2 max Specific Term 92 –
strength Common Term 61 –
strength training Common Term 72 –
warm-down Common Term 65 accepted, but
preferred term is cool-down
Garmin Named Entity 99 –
accelerometer OoD Term 73 –