“Let’s Tackle This Translation Task”: Legal Homonym Disambiguation with Reasoning Models

Paolo Di Natale[1,2], Elena Chiocchetti[2,*] and Marlies Alber[2]

[1]Free University of Bolzano/Bozen, Bolzano/Bozen, Italy
[2]Eurac Research, Institute for Applied Linguistics, Bolzano/Bozen, Italy

Abstract

This paper aims to help fill a current research gap by assessing the quality of legal translation produced by LLMs. It focuses on the handling of legal terminology and the disambiguation of legal homonyms with the help of reasoning models (RMs). The paper starts from a real-world use case: South Tyrol is a bilingual province in Italy where Italian and German are co-official languages and legal translation from Italian into a non-dominant legal variety of German constitutes a daily activity for many organisations.

Results indicate that RMs show a strong capacity to effectively distinguish between different legal contexts or legal subdomains. Enabling reasoning in LLMs increases the rate of accurate homonym disambiguation. Longer reasoning also correlates with correct homonym disambiguation. However, complying with system-bound legal terminology, especially in a non-dominant variety, remains a challenge. RMs tend to translate as if there were only one variety of legal German. Homonym disambiguation may rely on legally debatable criteria (e.g. frequency, formality). While reasoning successfully helps disambiguate homonyms, the rest of the text is not necessarily adapted in consequence of the decision. Further research is needed to assess whether systematic pivoting via English (or reasoning in English) influences translation quality. Results stress the importance of feeding LLMs not only with domain-specific terminology but also with data that enable them to distinguish between legal varieties (e.g. from terminology databases).

Keywords

Reasoning language models, legal terminology, terminological variation, homonym disambiguation

1. Introduction

Large language models (LLMs) are increasingly being used for translation purposes [1, 2, 3], including in the field of legal translation [4, 5]. However, research on the use of LLMs for legal translation remains scarce [6].

In multilingual societies, legal translation is a daily activity for many public and private organisations, and any effective support provided by translation tools is welcome [7, 8]. However, LLMs are primarily trained on general language data, which results in lower translation quality in highly specialised domains such as law [9]. Quality also varies across language pairs, with performance typically declining for low-resource languages and non-English language combinations [10, 11]. Many multilingual societies do not have English as one of their official languages, such as Belgium (Dutch, French, German), Switzerland (German, French, Italian, Romansh) and the Italian province of Bolzano/Bozen (Italian, German). This paper focuses on the latter area, which is also known as “South Tyrol”.

One of the major challenges in legal translation relates to legal terminology [12], which may show various forms of ambiguity. For example, certain words can be used either in their general language sense or in their specific legal meaning (e.g. “trust”). Legal terms may have different meanings depending on the specific context or legal subdomain. For example, “charge” may refer to the formal statement of the crime a party is accused of or to the oral instructions given by a judge to the jurors before deliberations. Correctly disambiguating such legal homonyms during translation poses a challenge for both humans and machines.

5th International Conference on “Multilingual digital terminology today. Design, representation formats and management systems” (MDTT) 2026, June 25-26, 2026, Zadar, Croatia.

*Corresponding author.
pdinatale@eurac.edu (P. D. Natale); echiocchetti@eurac.edu (E. Chiocchetti); marlies.alber@eurac.edu (M. Alber)
0009-0000-5840-4954 (P. D. Natale); 0000-0002-1309-7759 (E. Chiocchetti); 0009-0009-6255-6710 (M. Alber)

A further challenge arises from the system-bound nature of legal terminology [13, 14]. Each legal system has developed own rules and terminology, resulting in conceptual and linguistic differences even between legal systems that use the same official language. For example, the strict separation between solicitors and barristers does not apply in the United States, and what is called "corporate law" in the US corresponds to "company law" in the United Kingdom.

The aim of this paper is to begin addressing the research gap concerning the use of LLMs for legal translation from Italian into the specific (legal) variety of German used in South Tyrol. In particular, it examines how LLMs disambiguate legal homonyms during the translation process, a typical legal translation challenge. The recent emergence of Large Reasoning Models (RMs) provides an excellent opportunity to gain insight into the internal processes of LLMs when applied to translation tasks. While being aware that reasonings may not be faithful [15], our quantitative and qualitative analyses shed light on both the legal translation performance and homonym disambiguation process of RMs as well as the overall disambiguation success rate. In addition, we argue for the importance of feeding LLMs not only with domain-specific terminology but also with data that enables them to distinguish between legal varieties.

2. Background

2.1. Reasoning LLMs

Since LLMs have been found to perform a wide range of tasks even without specific training, researchers found prompting an effective strategy to impart generic instructions [16]. However, the limitations of prompting become apparent in more specialised or cognitively demanding settings. To avoid computationally onerous post-training procedures, the objective of instilling reasoning capabilities to resolve semantically and logically ambiguous challenges was established.

The first approaches relied on explicitly instructing models to "think step by step" (elicitive prompting) [17] or provided templates of the reasoning patterns to be applied (in-context learning) [18]. Under these conditions, the models tended to constrain text output to the formats and lengths they were pre-trained to follow. Additionally, it offered little incentive for exploring alternative reasoning routes that could better leverage the world knowledge of LLMs.

By allowing models to generate an unbounded sequence of intermediate reasoning steps in a "scratch-pad" (later accessible for qualitative inspection) [19], a greater portion of tokens can be devoted to the thinking phase itself. Strategies such as iterative "self-asking" to reassess provisional answers suggested not only a positive correlation between reasoning depth and answer accuracy, but also the emergence of a form of self-reflection [20]. Models were observed to revise their initial responses without external intervention [21].

Recent findings [22] have demonstrated that reasoning abilities can be injected and optimised for directly during training with no need for human supervision. The release of RMs introduced "native thinking" [23] achieved through test-time scaling [24]: when users set a thinking effort according to task difficulty, the models correspondingly allocate a variable number of tokens at inference time to an explicit reasoning phase. From the outset, these models are trained to defer final answers by producing intermediate reasoning steps leading up to the output. While doubts remain about their genuine logical inference skills [25], we argue that RMs may benefit lower-resourced settings [26]. By decomposing tasks into sub-units, models may be steered to focus on sensitive passages, locale-oriented choices, or user-provided terminology.

As a matter of fact, inducing terminology into translated text is a persistent challenge in machine translation (MT). This issue persists in LLMs [27]: inserting user-specified terms often compromises fluency, as the generation probabilities of the surrounding context are reshaped to accommodate terminology requirements. We maintain that augmenting and focusing the reasoning effort on the appropriateness of specific terminology may help redress this gap. Through a qualitative analysis of reasoning traces, we also aim to shed some light on the decision processes that undergird the generation

of translated text when implementing statistically under-represented terminology in non-dominant language varieties.

2.2. Legal translation

Legal language has always posed translation challenges to both humans and machines and has long been considered unsuitable for machine translation [12, 28]. Sentences in legal texts tend to be long and syntactically complex. The legal domain has highly domain-specific terminology (e.g. "felony") and phraseology (e.g. "uphold a verdict"). However, there are also legal terms that coincide with general-language words (e.g. "claim", "opinion"). In addition, legal language is system-bound and differences between the conceptual systems of various countries are the rule rather than an exception [13, 14]. This means that concepts from a specific source legal system may not exist in the target legal system or may be applied and embedded differently within it. At the lexical level, the same term may have different meanings in different legal systems.

Conceptual and terminological misalignments are also present when comparing legal systems that use the same official language. For example, there are as many German legal language varieties as there are countries using German as a (co-)official language, including Germany, Austria, Switzerland, Italy, Belgium and the EU. Conceptual and terminological differences can vary according to the specific legal subdomain, with domains like administrative and family law varying much more than domains that have undergone a process of harmonisation (e.g. privacy law and consumer protection law within the EU). For example, the concept of *Milizparlament*, a legislative body whose members are not full-time politicians but serve part-time in parliament while maintaining their regular professions, is typically Swiss and not part of any other German-speaking legal system. A collective bargaining agreement is termed *Tarifvertrag* in Germany, *Kollektivvertrag* in Austria (and South Tyrol) and *Gesamtarbeitsvertrag* in Switzerland, with different national rules applying to collective bargaining.

Translating legal texts is a high-stakes activity because mistakes can have serious consequences, including loss of reputation, financial liability, legal disputes and even more serious consequences [28]. Despite these challenges, machine translation in the legal domain is gaining prominence and LLMs are being tested for legal translation tasks [29]. LLMs are increasingly successful in producing fluent and contextually adequate legal translations [30]. However, details and cultural nuances that may be extremely relevant in legal translation still pose challenges, especially for some low-resource language combinations [31]. Approaches relying on a human-in-the-loop and AI agents may address existing shortcomings in future [32].

2.3. South Tyrolean (legal) German

German – like Arabic, English and Spanish, to name a few examples [33] – is a pluricentric language with several codified standard forms. South Tyrolean German is a non-dominant standard variety of German spoken as a first language by a minority of about 300,000 Italian citizens [34] living in the Autonomous Province of Bolzano/Bozen (South Tyrol), Italy. It is a southern German variety that has distinctive terminology in the domain of food as well as law and administration [35] due to the contact with Italian and the area being part of the Italian legal system.

For legal translation in South Tyrol, the system-boundness of legal language implies both the language-related challenge of coping with regional language variation [36] and domain-related challenges due to cross-systemic conceptual incongruity of the German-speaking legal systems [14]. It is not only necessary to respect regional language preferences (e.g. the choice of the auxiliary *sein* vs *haben* for some position verbs) but also to strictly adhere to the legal terminology of the specific legal system [37]. Otherwise, texts might be misunderstood or not considered legally valid.

In addition, in South Tyrol there is a Terminology Commission [38] that officially standardises legal terminology. The Terminology Commission's decisions have a binding nature for public authorities, even though not all standardised terms are equally successful and quickly implemented. South Tyrolean

terminology is publicly available via the Information System for Legal Terminology *bistro*[1] [39]. *bistro* contains about 14,000 Italian legal concepts with their definitions and contexts of use, together with terms used in South Tyrol, including standardised terms, and – where available – also the terms used for equivalent legal concepts in other German-speaking legal systems (Austria, Germany, Switzerland and the EU).

Like other non-dominant (legal) varieties, South Tyrolean German has fewer language resources and is underrepresented in reference works (e.g. dictionaries) and in the training data of current technologies, such as neural MT systems and generative AI applications [40, 33, 41]. However, it is specifically in multilingual societies like South Tyrol or Switzerland that such tools are needed and widely used [7, 8] in language combinations that often exclude English. MT quality in less-resourced languages and varieties and non-English language combinations shows a known quality gap [42] but specific research is needed for each language (variety) and language combination.

3. Method

We conduct our experiments on the homonym subsection of the LegISTyr test set [27]. It was created to evaluate the MT quality of legal translation from Italian into South Tyrolean German, with a focus on local terminology. It contains over 2000 Italian legal sentences with terms that have specific equivalents in South Tyrol, due either to customary use or to the Terminology Commission's decisions. There is a subset of 250 examples with legal homonyms, which need to be correctly disambiguated based on the surrounding context for correct translation into South Tyrolean German (see Table 1). For example, the Italian term *procedura concorsuale* can mean *Insolvenzverfahren* (insolvency proceedings) within insolvency law or *Wettbewerbsverfahren* (open competitive examination) within administrative law. Each source term in LegISTyr is associated with two or three possible South Tyrolean German equivalents. For each word sense, five instances are provided. By design, the example sentences in the test set provide the necessary context cues to enable disambiguation.

Table 1
Examples from the LegISTyr test set, more specifically from the homonym subset. For each example **Source sentence**, the subset contains information on the **Legal subdomain**, the Italian **Source term**, the preferred South Tyrolean German **Target term** and – where available – any **Other terms** from South Tyrol or **Other terms from other legal systems** using German.

	Example 1	**Example 2**
Source sentence	Il diploma è rilasciato dall'assessorato provinciale competente in materia di formazione sanitaria ed è sottoscritto anche dal presidente della commissione giudicatrice.	Nell'appalto concorso la nomina di una commissione giudicatrice diversa dagli uffici dell'amministrazione appaltante non è obbligatoria, ma facoltativa.
Legal subdomain	administrative law	insolvency law
Source term	commissione giudicatrice	commissione giudicatrice
Target term	Prüfungskommission	Bewertungskommission
Other terms from South Tyrol	/	Preisgericht
Other terms from other legal systems	Prüfungsausschuss	/

We evaluate both general-purpose LLMs and reasoning models (RMs) (see Table 2) to appreciate the effect of reasoning-oriented training by comparing them to architecturally matched counterparts that have not been reinforced or fine-tuned for explicit reasoning behaviours.

[1] https://bistro.eurac.edu/

We experiment with setting a long (max 2,000 tokens) and a short (max 100 tokens) reasoning budget to control for performance variations along thinking effort. For models that do not let us limit the reasoning tokens count, we approximate these conditions by setting the thinking effort to "medium" and "low", respectively. The average number of generated thinking tokens is reported in Table 2.

We also probe the effectiveness of prompting in the target language (German), considering that legal concepts and terminology are system-bound.

We generate translations via the OpenRouter API[2], which interfaces between the user and various model backend providers. This allows us to access both commercial and open-sourced models through a single entry point. We concede that this setup reduces transparency when it comes to assessing the computational costs and throughput delays associated with the increased compute demands of reasoning models. It is difficult to establish the actual additional compute end users would incur on consumer-grade GPUs, as OpenRouter distributes computation across providers under the hood. However, we report the total monetary costs charged by the reasoning tokens, based on the OpenRouter rates at the time of model access.

Generation parameters are fixed across all experiments. Following prior work on constrained generation tasks [43], we set the temperature to 0.2 and the top-p value to 0.95. We employ a uniform prompt for all models (see Appendix A), providing both the correct and incorrect target terms as candidate translation and leaving the model to resolve homonymy through contextual disambiguation.

For automatic evaluation (see Section 4.1), we measure terminology accuracy rate by verifying whether the correct target term appears in the machine-generated translation. To evaluate output fluency, we compute MetricX-XL [44], a state-of-the-art automatic evaluation metric. The resulting score corresponds to the predicted penalty a translation would receive on the 0–25 MQM scale; lower scores thus indicate higher-quality translations. In this way, we track the trade-off between terminological accuracy and overall fluency [45]. We advise against interpreting MetricX scores in absolute terms, but rather as relative points of comparison across models and experimental conditions.

For qualitative evaluation (see Section 4.2), we focus on Qwen-Plus and DeepSeek-R1. Most commercial models encrypt reasoning traces, while these models provide access to intermediate reasoning processes. Our analyses focus on which homonym disambiguation and translation decision criteria are explicitly mentioned, how the South Tyrolean legal variety is implemented and what external sources are (claimed to be) consulted. We also assess whether homonym disambiguation is based on sensible reasoning and decision criteria and check the accuracy and fluency of the surrounding translation.

4. Results

4.1. Quantitative analyses

We find that enabling reasoning improves the rate of accurate homonym selection compared to both deactivated mode and models without this feature. We also observe a positive correlation between longer thinking traces and correct homonym rate selection, which confirms that allocating additional tokens at inference time brings benefits vis-à-vis increased computing and time demands. However, prompting in a language other than English does not seem to impact the reasoning trajectory of the model, especially considering that reasoning generally continues to happen in the pivot language English. This should not be surprising, as reasoning data used during training almost only includes English.

To assess whether reasoning improvements are statistically significant, we apply a one-sided Wilcoxon signed-rank test. Because each term is represented by five instances, we can perform a paired test at the term level, comparing the median accuracy rates for each term. In this manner, we make sure that gains are consistent across all terms, rather than on raw total insertion rates alone, which may be inflated by some good outliers. We find that reasoning outperformance is statistically significant ($p < 0.05$) for all models, except for Deepseek-R1.

[2]https://openrouter.ai/

Regarding the fluency metric, we do not find consistent improvements related to thinking mode. As discussed in Section 4.2, the thinking effort zooms in on the aspects evidenced in the prompt, namely keeping a standard language register and implementing the requested terminology. The remainder of the sentence remains largely unaffected by the iterative editing actions triggered across reasoning steps, which cause minor variations in the metric score.

For deployment costs, affordability depends on user's resources and priorities. The central question revolves around the trade-off between the length (and cost) of reasoning trajectories and actual performance gains. In the medium range of reasoning effort, Gemini charges roughly half the cost of GPT, although its higher number of total reasoning tokens can lead to increased generation latency. By contrast, GPT-5.1 achieves largely competitive performance to Gemini and Qwen3-Plus with fewer than 100 reasoning tokens, resulting in a significantly lower (or comparable) monetary expense. A comprehensive estimation should account not only for the nominal cost per output token, but also for a model's ability to achieve comparable performances with fewer, more effective reasoning tokens. DeepSeek R1, being an open model, can also be deployed locally given sufficient hosting capacity. Under these conditions, the burden of cost optimisation is placed onto the user, who must appropriate and manage sufficient computational resources.

Table 2
Evaluation of tested models on our test set. The **Thinking** column indicates whether the thinking mode has been activated, while **Thinking effort** counts the average number of thinking tokens generated in the thinking traces. **Correct Homonym** reports the terminology accuracy rate, while **Fluency penalty** reports the MetricX score. **Reasoning cost in $** reports the inference cost of reasoning tokens, based on OpenRouter rates at the time of the model access request. We do not collect the reasoning token count for prompts in German.

Model	Thinking	Thinking effort (avg. tokens)	Prompt language	Correct homonym	Fluency penalty	Reasoning cost in $
Deepseek R1	Yes	531	English	76.80%	-2.279	0.33
Deepseek V3.1 terminus	No	0	English	71.60%	-2.357	n/a
Gemini 2.5 flash	No	0	English	68.80%	-2.234	n/a
	Yes	94	English	78.40%	-2.152	0.08
	Yes	869	English	87.20%	-2.129	0.56
	Yes	n/a	German	85.20%	-2.069	n/a
Gpt 4.1	No	0	English	64.80%	-2.080	n/a
Gpt 5.1	Yes	80	English	85.60%	-2.114	0.33
	Yes	313	English	86.80%	-2.044	0.91
	Yes	n/a	German	85.60%	-2.106	n/a
Qwen3 Plus (2025-07-28)	No	0	English	73.60%	-2.170	n/a
	Yes	93	English	61.20%	-2.543	0.01
	Yes	1547	English	82.80%	-2.123	0.30
	Yes	n/a	German	80.00%	-2.120	n/a

4.2. Qualitative analyses

The qualitative evaluation of DeepSeek-R1 and Qwen-Plus outputs focused on their reasoning processes regarding homonym disambiguation and overall translation quality. In general, both models tend to separate reasoning on terminology, sentence structure and grammar, sometimes creating small bilingual glossaries. In terms of translation quality, they generally prefer sticking to literal translations (e.g. "However, legal language might prefer the literal translation."). With respect to terminology, both models encounter considerable difficulties regarding the two major challenges of legal translation mentioned in Section 2.2: the inherent ambiguity of legal terms and their system-bound nature.

Observations reveal that both Deepseek-R1 and Qwen-Plus typically initiate the disambiguation of homonymous terms using contextual cues as the basis for subsequent choices. In most cases, the models show a strong capacity to effectively distinguish between different legal contexts or legal subdomains (e.g. “So between the two options, ""Insolvenzverfahren"" is the correct term here because it directly relates to insolvency proceedings. ""Wettbewerbsverfahren"" would translate to competition procedures, which doesn’t fit the context.”). However, although they successfully manage to assign the correct meaning to the legal homonyms under examination, the criteria underpinning these decisions often rely on generic factors, such as frequency (e.g. “I’ll choose ""Preis"" since it’s more common for awards in legal contexts like this.”) or register (e.g. “""Bestandnehmer"" might be a more formal or specific legal term.”).

Language variation was rarely employed as a criterion for homonym disambiguation. For instance, Deepseek – unable to resolve the correct distinction between *Mieter* (tenant) and *Bestandnehmer* (lessee) based on the context (see Section 5) – resorted to language variation, presuming *Bestandnehmer* to be specific to South Tyrolean legal terminology, as it appeared to be less frequent. Furthermore, terminological decisions were not necessarily applied consistently throughout the sentence. For example, as a consequence of opting for *Bestandnehmer* during homonym disambiguation, the other party in the contract should have become *Bestandgeber* (lessor) and not remain *Vermieter* (landlord).

Regarding the reasoning on language variation in law, both systems rarely engaged with its distinctive features, often acting as if there were only one “legal German” or set of “German legal terminology”. In most cases, the reasoning indicated that South Tyrolean German was equated generically with standard German (e.g. “Since it’s a standard variety, standard German terms should suffice.” or “I need to translate it naturally into standard German, as South-Tyrolean German aligns with that.”). They thus disregarded both regional variation and the system-boundness of legal language (e.g. “The translation should adhere to standard German legal terminology, which is the same as in Germany, Austria, etc., for legal contexts.”). In general, “standard variety” was interpreted as an instruction to avoid dialectal expressions (e.g. “The user said it’s a standard variety, so no special dialect terms.”). Only occasionally was terminological variation considered possible (e.g. “In South Tyrol, which is a German-speaking region in Italy, they might use specific terms.” or “In German legal terminology, ""Zivilgesetzbuch"" is sometimes used, though ""Bürgerliches Gesetzbuch"" is more common in Germany. However, in South Tyrol, which follows Italian law but has German as an official language, they might use ""Zivilgesetzbuch"" for the Italian civil code.").

Another limitation was the lack of systematic consultation of external resources, particularly those pertinent to South Tyrolean legal language. Although the models occasionally claimed to consult external data, such as an unspecified legal dictionary, an Austrian legal text or the Italian Civil Code, there was little evidence that these sources actually influenced the reasoning process. Search and verification results were not always reported and were sometimes incorrect (e.g. “Another check: ""bene mobile"" is ""bewegliches Gut"", correct.”; however, *bene mobile*, movable property, has a different standardised equivalent in South Tyrol, i.e. *bewegliche Sache*). Additionally, the models rarely referred to South Tyrolean legal texts or websites, as a human translator would most probably do to identify appropriate terminology.

Finally, a further complication arises from the strong bias of LLMs towards English. Generally, both systems rendered the Italian source sentence (or sentence chunks) quite literally into English as an intermediate step before translating it into (South Tyrolean) German.

5. Discussion

By analysing the reasoning traces of two RMs, we gained valuable insights into how LLMs approach legal translation tasks and, in particular, the disambiguation of legal homonyms. Based on our quantitative results, enabling reasoning in LLMs appears to be a successful strategy for achieving accurate legal homonym disambiguation, with longer thinking efforts correlating positively with correct disambiguation results. However, it should be noted that the reasoning does not necessarily reflect the

actual behaviour of these models. For instance, they sometimes claim to search for documents and check information, but also state that they do not have access to external sources. In addition, several false or unreliable statements can be found (e.g. “"Assessorato provinciale" is Landesamt in South Tyrolean German.”, whereas the term normally used in South Tyrolean German is *Landesressort*). Other statements simply do not provide any clues about the underlying thinking process (e.g. “Check if "am Titel" is correct. Yes, "am Titel" [on the title] is correct.”).

Translating legal texts is a major challenge on several levels, not only for machines but also for humans. Considering the differences between legal systems and (legal) language varieties is of particular importance. Our analyses indicate that LLMs possess only a limited capacity to adhere to the specificities and conventions of various legal systems. However, respecting the system-boundness of legal language is an essential quality factor in legal translation, as each variety represents a distinct legal language with its own system-bound concepts and terminology. For example, while *esame di stato* (state exam) can be translated as *Staatsexamen* for Germany, it should be *Staatsprüfung* for South Tyrol.

In addition to linguistic variation, legal terms can also be characterised by semantic ambiguities and their meaning may differ depending on context or legal subdomain. Generally, our study indicates that RMs are able to effectively recognise these differences in meaning and choose the correct term for translation in most cases. However, the decisions are often based on problematic criteria such as frequency, diastratic variation (formal register, no dialect) or diaphasic variation (legal language). While this approach might yield satisfactory results for more generic translation tasks, it is not necessarily appropriate for translating legal texts, where accuracy is crucial for avoiding misunderstandings, ensuring legal certainty, and adhering to system-specific terminology conventions. In fact, less frequent terms may be more appropriate in certain contexts, especially when dealing with highly specialised terminology and a specific language variety. Literal or frequent translations are therefore not always the safer option, as repeatedly assumed by the two RMs under examination. Additionally, decisions seem to be made in isolation, as they are not always propagated consistently through the same sentence. This behaviour stands in contrast to the translation strategies typically employed by humans.

Translation of legal texts is particularly relevant for multilingual regions. Indirect translations via a pivot language (e.g. English) are therefore not a realistic scenario for contexts like South Tyrol. Moreover, pivoting through a third language (and legal system) introduces an additional source of risk. Especially in the legal domain, full equivalence between terms is rare. Therefore, the more languages are involved, the higher the risk of severe mistranslations. At any rate, it remains to be studied whether there was an actual indirect translation process or whether English is just the language used for the reasoning output. Some examples suggest no relevant influence of English for the final translation. For instance, *sintetico* (concise) was incorrectly translated into English with the false friend “synthetic” (artificial, not natural) but still correctly rendered as *kurz* (short) in German.

Regarding translation tasks in general, RMs seem to address terminology, structure and grammar separately. Hence, there may be scope for targeted terminology injection. Direct access to terminological resources such as the South Tyrolean Information System for Legal Terminology *bistro* could solve many of the issues addressed in this paper. This would help provide officially validated terminology or information on legal language varieties. For example, the term entry for *codice civile* (Civil Code) in *bistro* lists the correct equivalent for different German-speaking legal systems: *Bürgerliches Gesetzbuch* for Germany, *Allgemeines bürgerliches Gesetzbuch* for Austria and *Zivilgesetzbuch* for Switzerland and South Tyrol. It would also provide precious information for homonym disambiguation. For example, the term entry for *locatario* in *bistro* explains that *Mieter* is to be used only for tenancy agreements while *Bestandnehmer* applies to lease agreements in general in South Tyrol.

6. Limitations

We are aware of several limitations of this paper that could be partly addressed in future work. The dataset is small and limited to one language combination and legal system. A larger dataset and comparisons with other non-dominant varieties and different language combinations would help assess

whether the results are generalisable. Claims that the models are checking external resources cannot be verified, so that the (real) reasons underlying specific decisions often remain partly unclear. We used the same prompt to enhance comparability, but crafting tailored prompts for each model may lead to different results and may better exploit their specific strengths. Finally, more comprehensive human evaluations of the target texts following an error-annotation framework like MQM [46] would shed further light on the most frequent and problematic categories of translation mistakes produced by LLMs in our specific translation scenario and could guide targeted improvements.

7. Conclusion and outlook

Working with language combinations that exclude English, with low-resource languages and with highly specialised domains are well-known MT challenges that can be confirmed by our study. We tested the homonym disambiguation and legal translation capabilities of different LLMs in the language combination Italian to South Tyrolean German. We show that also non-dominant varieties of high-resource languages like German face the risk of being overshadowed by dominant varieties, both in terms of regional variation in general language and system-bound variation in legal language. Addressing this issue would benefit several multilingual communities where legal translation is a daily activity for many public and private organisations.

Based on our quantitative analyses, RMs exhibit promising results in legal translation and homonym disambiguation. However, wrong outcomes and even correct outcomes achieved through faulty reasoning highlight the need to inject domain-related and system-specific information into LLMs. Much of the information required to correctly disambiguate legal homonyms and steer the choice of legal terms is already available in terminology databases. For South Tyrol, *bistro* even provides information on the different varieties of legal German. Future work should therefore focus on finding and implementing efficient strategies that enable LLMs to access such terminological data, for example via retrieval-augmented generation based on knowledge graphs or ontologies. Real-time interaction with LLMs may also require that terminological databases conform to specific standards and formats (e.g. TermBase eXchange, Linguistic Linked Data). As LLMs and terminological resources evolve, it will also be necessary to repeat the evaluation following major developments.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-5.2 and Microsoft Copilot for grammar and spelling checks and to improve the writing style. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, L. Li, Multilingual machine translation with large language models: Empirical results and analysis, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, p. 2765–2781. doi:10.18653/v1/2024.findings-naacl.176.

[2] R. Zhang, W. Zhao, S. Eger, How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs (2024). doi:10.48550/arXiv.2410.18697.

[3] B. Zhang, B. Haddow, A. Birch, Prompting large language model for machine translation: A case study (2023). doi:10.48550/arXiv.2301.07069.

[4] A. Larroyed, Redefining patent translation: The influence of ChatGPT and the urgency to align patent language regimes in europe with progress in translation technology, GRUR International 72 (2023) 1009–1017. doi:10.1093/grurint/ikad099.

[5] R. Sousa-Silva, 'We attempted to deliver your package': Forensic translation in the fight against cross-border cybercrime, International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique 37 (2024) 1323–1349. doi:10.1007/s11196-023-10102-2.

[6] M. Bajčić, D. Golenko, Applying large language models in legal translation: The state-of-the-art, International Journal of Language Law (JLL) 13 (2024) 171–196. doi:10.14762/jll.2024.171.

[7] F. D. Camillis, La traduzione non professionale nelle istituzioni pubbliche dei territori di lingua minoritaria: il caso di studio dell'amministrazione della Provincia autonoma di Bolzano, Ph.D. thesis, alma, 2021. URL: https://amsdottorato.unibo.it/id/eprint/9695/.

[8] R. Martínez-Domínguez, M. Rikters, A. Vasiļevskis, M. Pinnis, P. Reichenberg, Customized neural machine translation systems for the Swiss legal domain, in: J. Campbell, D. Genzel, B. Huyck, P. O'Neill-Brown (Eds.), Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track), Association for Machine Translation in the Americas, Virtual, 2020, pp. 217–223. URL: https://aclanthology.org/2020.amta-user.11/.

[9] A. Mukherjee, M. Shrivastava, IIIT HYD's submission for WMT23 test-suite task, in: P. Koehn, B. Haddow, T. Kocmi, C. Monz (Eds.), Proceedings of the Eighth Conference on Machine Translation, Association for Computational Linguistics, Singapore, 2023, p. 246–251. doi:10.18653/v1/2023.wmt-1.24.

[10] N. Robinson, P. Ogayo, D. R. Mortensen, G. Neubig, ChatGPT MT: Competitive for high- (but not low-) resource languages, in: P. Koehn, B. Haddow, T. Kocmi, C. Monz (Eds.), Proceedings of the Eighth Conference on Machine Translation, Association for Computational Linguistics, Singapore, 2023, p. 392–418. doi:10.18653/v1/2023.wmt-1.40.

[11] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, H. H. Awadalla, How good are GPT models at machine translation? a comprehensive evaluation (2023). doi:10.48550/arXiv.2302.09210.

[12] J. Killman, Machine translation and legal terminology. data-driven approaches to contextual accuracy, in: Biel, H. J. Kockaert (Eds.), Handbook of Terminology. Legal Terminology, volume 3, John Benjamins, Amsterdam/Philadelphia, 2023, p. 485–510.

[13] D. Cao, Translating Law, Multilingual Matters, Clevedon, 2007.

[14] Biel, H. Kockaert, Introduction. legal terminology, in: Biel, H. Kockaert (Eds.), Handbook of Terminology. Legal Terminology, volume 3, John Benjamins, Amsterdam / Philadelphia, 2023, p. 1–14.

[15] Y. Chen, J. Benton, A. Radhakrishnan, J. Uesato, C. Denison, J. Schulman, A. Somani, P. Hase, M. Wagner, F. Roger, V. Mikulik, S. R. Bowman, J. Leike, J. Kaplan, E. Perez, Reasoning models don't always say what they think (2025). doi:10.48550/arXiv.2505.05410.

[16] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3560815. doi:10.1145/3560815.

[17] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 22199–22213. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

[18] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, X. Sun, L. Li, Z. Sui, A survey on in-context learning, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 1107–1128. URL: https://aclanthology.org/2024.emnlp-main.64/. doi:10.18653/v1/2024.emnlp-main.64.

[19] M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, C. Sutton, A. Odena, Show your work: Scratchpads for intermediate computation with language models, 2021. URL: https://arxiv.org/abs/2112.00114. arXiv:2112.00114.

[20] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, P. Clark, Self-

refine: Iterative refinement with self-feedback, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 46534–46594. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

[21] O. Press, M. Zhang, S. Min, L. Schmidt, N. Smith, M. Lewis, Measuring and narrowing the compositionality gap in language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 5687–5711. URL: https://aclanthology.org/2023.findings-emnlp.378/. doi:10.18653/v1/2023.findings-emnlp.378.

[22] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, B. et al., Deepseek-r1 incentivizes reasoning in LLMs through reinforcement learning, Nature 645 (2025) 633–638. URL: http://dx.doi.org/10.1038/s41586-025-09422-z. doi:10.1038/s41586-025-09422-z.

[23] J. Wang, A tutorial on LLM reasoning: Relevant methods behind ChatGPT o1, 2025. URL: https://arxiv.org/abs/2502.10867. arXiv:2502.10867.

[24] N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, T. Hashimoto, s1: Simple test-time scaling, 2025. URL: https://arxiv.org/abs/2501.19393. arXiv:2501.19393.

[25] F. Wu, W. Xuan, X. Lu, M. Liu, Y. Dong, Z. Harchaoui, Y. Choi, The invisible leash: Why RLVR may or may not escape its origin, 2026. URL: https://arxiv.org/abs/2507.14843. arXiv:2507.14843.

[26] S. Frontull, T. Ströhle, Compensating for data with reasoning: Low-resource machine translation with LLMs, 2025. URL: https://arxiv.org/abs/2505.22293. arXiv:2505.22293.

[27] P. Di Natale, E. W. Stemle, E. Chiocchetti, M. Alber, N. Ralli, I. Stanizzi, E. Benini, The LegISTyr test set: Investigating off-the-shelf instruction-tuned LLMs for terminology-constrained translation in a low-resource language variety, in: K. Gkirtzou, S. Žitnik, J. Gracia, D. Gromann, M. P. di Buono, J. Monti, M. Ionov (Eds.), Proceedings of the 5th Conference on Language, Data and Knowledge: TermTrends 2025, Unior Press, Naples, Italy, 2025, pp. 1–15. URL: https://aclanthology.org/2025.termtrends-1.1/.

[28] H. E. Mattila, Legal Language, in: J. Humbley, G. Budin, C. Laurén (Eds.), Languages for Special Purposes: An International Handbook, De Gruyter Mouton, Berlin, Boston, 2018, pp. 113–150.

[29] M. Bajčić, D. Golenko, Applying Large Language Models in Legal Translation: The State-of-the-Art, International Journal of Language & Law (JLL) 13 (2024) 171–196. doi:10.14762/jll.2024.171.

[30] V. Briva-Iglesias, G. Dogru, J. L. Cavalheiro Camargo, Large Language Models "ad referendum": How good are they at machine translation in the legal domain?, MonTI. Monographs in Translation and Interpreting 16 (2024) 75–107. doi:https://doi.org/10.6035/MonTI.2024.16.02.

[31] A. M. Badah, C. N. Khalaf, F. J. Dwaikat, N. AlQbailat, The usage of artificial intelligence in legal translation: Bridging the gap between law and language, Ampersand 16 (2026) 100248. URL: https://www.sciencedirect.com/science/article/pii/S2215039025000323. doi:https://doi.org/10.1016/j.amper.2025.100248.

[32] V. Briva-Iglesias, Are AI agents the new machine translation frontier? Challenges and opportunities of single- and multi-agent systems for multilingual digital communication, 2025. URL: https://arxiv.org/abs/2504.12891. arXiv:2504.12891.

[33] B. Schuppler, M. Adda-Decker, C. Cucchiarini, R. Muhr, An introduction to pluricentric languages in speech science and technology, Speech Communication 156 (2024) 103007. doi:10.1016/j.specom.2023.103007.

[34] Astatinfo, Ergebnisse Sprachgruppenzählung - 2024 / Risultati Censimento linguistico - 2024, 56, Bozen/Bolzano, 2024. URL: https://assets-eu-01.kc-usercontent.com/b5376750-8076-01cf-17d2-d343e29778a7/5deec178-b2a3-4e2d-8795-d37635c7e0f7/pressnote_1160209_mit56_2024.pdf.

[35] U. Ammon, H. Bickel, A. N. Lenz, Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen, 2 ed., de Gruyter, Berlin, 2016.

[36] M. Magris, L. Rega, Auswirkungen der sprachvarietäten auf das Übersetzen mit besonderer

rücksicht auf diatopie, in: C. Di Meola, J. Gerdes, L. Tonelli (Eds.), Sprachvariation im Deutschen zwischen Theorie und Praxis. Fachsprachlichkeit, Inklusion, Didaktik, Übersetzung, Kontrastivität, Frank Timme, Berlin, 2025, p. 535–552.
[37] L. Biel, Variation of legal terms in monolingual and multilingual contexts, in: L. Biel, H. J. Kockaert (Eds.), Handbook of Terminology. Legal Terminology, volume 3, John Benjamins, Amsterdam/Philadelphia, 2023, p. 90–123.
[38] E. Chiocchetti, Effects of social evolution on terminology policy in South Tyrol, Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 27 (2021) 110–139. URL: https://www.jbe-platform.com/content/journals/10.1075/term.00060.chi. doi:https://doi.org/10.1075/term.00060.chi.
[39] N. Ralli, N. Andreatta, bistro – ein Tool für mehrsprachige Rechtsterminologie, trans-kom 11 (2018) 7–44.
[40] G. Rehm, A. Way, European language equality. Introduction, in: G. Rehm, A. Way (Eds.), European Language Equality. A Strategic Agenda for Digital Language Equality, Springer, Cham, 2023, p. 1–9.
[41] P. Shu, J. Chen, Z. Liu, H. Wang, Z. Wu, T. Zhong, Y. Li, H. Zhao, H. Jiang, Y. Pan, Y. Zhou, C. Owl, X. Zhai, N. Liu, C. Saunt, T. Liu, Transcending language boundaries: Harnessing LLMs for low-resource language translation (2024). doi:10.48550/arXiv.2411.11295.
[42] T. O. Tafa, S. Z. M. Hashim, M. S. Othman, H. Alhussian, M. Nasser, S. J. Abdulkadir, S. H. Huspi, S. O. Adeyemo, Y. A. Bena, Machine translation performance for low-resource languages: A systematic literature review, IEEE Access 13 (2025) 72486–72505. doi:10.1109/ACCESS.2025.3562918.
[43] S. Dhuliawala, I. Kulikov, P. Yu, A. Celikyilmaz, J. Weston, S. Sukhbaatar, J. Lanchantin, Adaptive decoding via latent preference optimization, 2024. URL: https://arxiv.org/abs/2411.09661. arXiv:2411.09661.
[44] J. Juraska, D. Deutsch, M. Finkelstein, M. Freitag, MetricX-24: The Google submission to the WMT 2024 metrics shared task, in: B. Haddow, T. Kocmi, P. Koehn, C. Monz (Eds.), Proceedings of the Ninth Conference on Machine Translation, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 492–504. URL: https://aclanthology.org/2024.wmt-1.35/. doi:10.18653/v1/2024.wmt-1.35.
[45] P. Chen, N. Bogoychev, K. Heafield, F. Kirefu, Parallel sentence mining by constrained decoding, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1672–1678. URL: https://aclanthology.org/2020.acl-main.152/. doi:10.18653/v1/2020.acl-main.152.
[46] A. R. Lommel, A. Burchardt, H. Uszkoreit, Multidimensional quality metrics: a flexible system for assessing translation quality, in: Proceedings of Translating and the Computer 35, Aslib, London, UK, 2013. URL: https://aclanthology.org/2013.tc-1.6/.

A. Prompt structure for homonym disambiguation during legal translation

```
[{
  "role": "user",
  "content":
    "You are a German translator based in South Tyrol and this is a translation task. "
    "You are tasked to translate a legal sentence from Italian into South-Tyrolean
        German. "
    "South-Tyrolean German is a standard variety of German. "
    "There are terminological constraints you must adhere to: "
    "{term_it} can be translated with only one of these terms: "
    "{term_de_1, term_de_2}. "
    "You must output only the translated text without any explanation, "
    "enclosing it in '<>' symbols. "
    "This is the text to be translated into German: "
    "<{source_sentence}>"
}
```