The *LegISTyr* Test Set: Investigating Off-the-Shelf Instruction-Tuned LLMs for Terminology-Constrained Translation in a Low-Resource Language Variety

Paolo Di Natale[1] **Egon W. Stemle**[1,2] **Elena Chiocchetti**[1] **Marlies Alber**[1]
Natascia Ralli[1] **Isabella Stanizzi**[1] **Elena Benini**[1,3]

[1]Eurac Research, Bolzano/Bozen, Italy [2]Masaryk University, Brno, Czech Republic
[3]University of Bologna, Bologna, Italy

{paolo.dinatale, egon.stemle, elena.chiocchetti, marlies.alber, natascia.ralli, isabella.stanizzi}@eurac.edu
elena.benini99@gmail.com

Abstract

We investigate the effect of terminology injection for terminology-constrained translation in a low-resource language variety, with a particular focus on off-the-shelf instruction-tuned Large Language Models (LLMs). We compare a total of 9 models: 4 instruction-tuned LLMs from the Tower and EuroLLM suites, which have been specifically trained for translation-related tasks; 2 generic open-weight LLMs (LLaMA-8B and Mistral-7B); 3 Neural Machine Translation (NMT) systems (an adapted version of MarianMT and ModernMT with and without the glossary function). To this end, we release *LegISTyr*, a manually curated test set of 2,000 Italian sentences from the legal domain, paired with source Italian terms and target terms in the South Tyrolean standard variety of German. We select only real-world sources and design constraints on length, syntactic clarity, and referential coherence to ensure high quality. *LegISTyr* includes a homonym subset, which challenges systems on the selection of the correct homonym where sense disambiguation is deducible from the context. Results show that while generic LLMs achieve the highest raw term insertion rates (approximately 64%), translation-specialized LLMs deliver superior fluency (Δ COMET up to 0.04), reduce incorrect homonym selection by half, and generate more controllable output. We posit that models trained on translation-related data are better able to focus on source-side information, producing more coherent translations.

1 Introduction

While Neural Machine Translation (NMT) adaptation has demonstrated benefits from incorporating domain-specific terms (Farajian et al. 2018), it has yet to ensure consistent and unambiguous terminology enforcement. The delicate trade-off between the continuous expansion of parallel training corpora and the enforcement of lexical choices without hampering fluency further complicates cross-lingual alignment of terminologically relevant tokens (Alkhouli et al. 2018; Ferrando et al. 2022; Štefánik et al. 2023). This failure has even raised questions on the cost-effectiveness of maintaining termbases for MT purposes (Knowles et al. 2023). Terminology compliance has attracted notable interest and lead to the organization of shared MT tasks (Bawden et al. 2019; Alam et al. 2021b; Semenov et al. 2023).

Terminology compliance is a crucial quality aspect in high-stakes domains, such as the legal domain. Terminology mistakes in legal translation can have serious consequences (Mattila 2018), including legal disputes and infringement of basic linguistic and human rights (e.g., through incorrect use of critical terminology during interpretation in criminal trials). Furthermore, every legal system has its own specific set of rules and conceptual structures. Legal terminology expresses such specificities and is therefore always bound to a specific legal system (Gambaro and Sacco 2024). This system-boundness of legal terminology often results in conceptual incongruency between legal systems, even between legal systems sharing the same language. Consequently, correct terminology usage in languages with more than one recognized legal variety like Arabic, English, Spanish etc., pose notable translation challenges. In addition, the quality of MT in the legal domain hinges not so much on the language combination as on the legal subdomain (Quinci and Pontrandolfo 2023). Our experiments focus on South Tyrolean German, a minor standard variety of German spoken in Northern Italy that is used by the local public administration bodies and on legal terms from a range of different legal subdomains.

Researchers have addressed terminology enforcement using both NMT systems and Large Language Models (LLMs). In 2024, two new LLM suites specifically trained on translation-related tasks (TT LLMs) were released to the public (Tower and EuroLLM suites); yet to our knowledge their capabilities have not been adequately investigated thus far. To bridge the gap, we evaluate the instruction-following performances of these open models in terminology injection, comparing their term accuracy rate and overall translation quality against general-purpose LLMs as well as to adapted NMT models and their baseline. We also assess which models produce the most structured and clean outputs for easier downstream processing.

To this end, we curate *LegISTyr*[1], a test set comprising over 2,000 sentences from the legal domain in Italian. Each instance is annotated with both Italian source and South Tyrolean target terms, plus variants used in other German-speaking legal systems whenever available. We select the sentences from termbase contexts and other real-world sources while enforcing constraints on length, syntactic clarity, and referential coherence. The test set covers the usage of each term in multiple contexts. It also features a subset on the challenge of disambiguation between homonyms where the correct sense is deducible from the domain or context of use.

2 Background

2.1 Challenges of legal translation in South Tyrol

While the task of terminology injection presents an already serious challenge for highly resourced languages, difficulties grow when considering minor varieties of pluricentric languages (Clyne 1991). Such varieties often lack labeled datasets and resources in internationally standardized format (Lakew et al. 2020), and are under-represented in generic training corpora due to the sheer size of text produced by the respective speakers (Zampieri et al. 2020). This leads to dominant forms obfuscating diatopic variants (Koehn and Knowles 2017). The implication in statistical MT is that the stronger signal from numerically dominant varieties will tend to override terms from minority ones.

The situation of a minor variety used in the legal domain can be exemplified by South Tyrolean German. This is a standard variety of German (Ammon et al. 2016) used by about 300,000 members of the German language minority in Italy. In South Tyrol, German is a co-official language next to Italian. South Tyrolean German differs from other German varieties in various minor grammatical and lexical aspects, but most notably concerning food and legal vocabulary. Some of these differences are unlikely to affect terminology (e.g. the use of *sein* vs *haben* as an auxiliary for some verbs of position), others might (e.g. differences in grammatical gender like *Kataster*, cadastre, generally being masculine in South Tyrol and other Southern German areas but neuter in Germany).

In South Tyrol, there is a process of terminology standardization in place whereby the mandatory terminology for the local legal and administrative texts is being validated by a Terminology Commission (Chiocchetti 2021). The infringement of local terminology requirements may result in hindered clarity of government-citizen communications and in the denial of linguistic minority rights related to clear and consistent communication in the minority language (South Tyrolean German).

An analysis of machine-translated normative texts from Italian into South Tyrolean German has highlighted that terminology is the second most common type of mistake after mistranslations (De Camillis and Chiocchetti 2024). This can happen by effect of interference from most represented legal systems.

We also observe that Italian multiword terms (e.g., *decreto legislativo*, legislative decree; *decreto ingiuntivo*, payment order; *decreto di condanna*, penalty order, etc.) tend to be shortened to their headword within texts (i.e., to *decreto*) creating homonymous short forms. Context is needed for correct disambiguation when translating into German (i.e., choosing between *gesetzesvertretendes Dekret*, legislative decree; *Mahndekret*, payment order and *Strafbefehl*, penalty order), especially since there is a less pronounced tendency to reduce multiword terms or compounds to their headword in German. The notable presence of homonyms deriving from ellipsis of part of the term in Italian legal texts poses an issue of (domain) disambiguation in the correct selec-

[1] http://hdl.handle.net/20.500.12124/104

tion of the target term in German. Disambiguation is relevant even in cases where there is no shortening of a longer term (e.g. *procedura concorsuale* means 'bankruptcy proceedings' in insolvency law, in German *Insolvenzverfahren*, but 'open competitive employment procedure' in administrative law and should be translated with *Wettbewerbsverfahren*).

To evaluate translation quality into South Tyrolean German, we release *LegISTyr*, a highly curated test set (see 4.1). We use the test set to explore the success of different techniques aimed at terminology injection when translating from Italian into a minor standard variety of German, viz. South Tyrolean German.

2.2 Aligning LLMs to translation tasks

While LLMs have shown the capability to perform targeted translation tasks without task-specific training (Vilar et al. 2023; Hendy et al. 2023), top-notch performances are still bottlenecked to proprietary, large-scale models, which makes their adoption in low-resource environments excessively expensive. Another pitfall of generic LLMs is their tendency to exhibit undesired behavior, such as verbosity — the generation of explanatory text in addition to the proper translation (Briakou et al. 2024).

One line of research has sought to imbue text with the core features of translation tasks through in-context learning (ICL), prompting a generic model with exemplars in a predefined format (Brown et al. 2020). Despite being largely resource-efficient and effective in cross-lingual transfer for under-resourced languages (Zhu et al. 2024b; Zhu et al. 2024a), drawbacks still persist. In actuality, prompt design seems to be a hardly manageable strategy, as minimal permutations result in heavy performance volatility (Leidinger et al. 2023; Sclar et al. 2024; Weber et al. 2023), unreasonable templates give rise to acceptable outputs (Zhu et al. 2024b) and targeted exemplar selection for the prompt only has a limited impact (Zhang et al. 2023). In addition, ICL proves ever more effective as the scale of the model in use grows (Min et al. 2022). Overcoming these limitations might be possible by leveraging pre-trained, lightweight models specialized in translation-related tasks in order to attain comparable performances to larger models (Zeng et al. 2024).

Instruction tuning (IT) (Wei et al. 2022) consists of fine-tuning an existing model on instruction-output pairs, aligning next-token prediction with task-specific objectives to enhance adherence to user instructions (Zhang et al. 2024). As a matter of fact, (Zhou et al., 2023) (2023) highlight that a model's core knowledge is acquired during pretraining, while fine-tuning mainly influences interaction modalities and output format. This effect is further amplified when LLMs are pre-trained on multilingual corpora (Briakou et al. 2023), where alignment with downstream translation objectives begins at the initial training stage (Sia et al. 2024).

2.3 Terminology injection approaches

Researchers have explored diverse techniques to inject terminology and ensure exact term rendering.

In NMT models, relevant terminology can be marked at inference time by inserting inline term labels in the source, target, or both. This can be achieved either by replacing all terms in source and target with a placeholder (Dinu et al. 2019; Bergmanis and Pinnis 2021; Michon et al. 2020), which sacrifices semantic information, or by adding the target term to the source sentence (code-switching) (Song et al. 2019; Ailem et al. 2021). Variations of this approach include additional lemmatization or grammatical editing for increased fluency (Bergmanis and Pinnis 2021; Pham et al. 2021) as well as changes in the location of the label (Jon et al. 2021; Turcan et al. 2022). However, consistently predicting the term equivalent during training has proven challenging, particularly under more complex circumstances than those in experiment conditions (e.g., with more than one target term per sentence). This has resulted in the introduction of hard constraints – i.e. forced insertion of the term in the target sentence – that are in turn affected by fluency and grammatical issues (Post et al. 2019; Chen et al. 2020).

Other attempts directly act on the decoder behaviour. The decoder is the component that generates the target translation by sequentially predicting the next token, given the encoded representation of the source text and the incrementally generated output. One focus lies on the development of decoding algorithms which enforce the desired term (Molchanov et al. 2021; Hauhio and Friberg 2024) or exclude unsuitable terms from the search space (Bogoychev and Chen 2023). Despite their potential, these methods increase computational time and resource demands. Additionally, they fail to address issues such as incorrect morphological inflections and unintentional repetition of terminology (Dinu et al. 2019).

With the surge in the use of Large Language Models, prompt formulation has made it possible to exert greater control over features of the output, including "specific dialect" (Garcia and Firat 2022) and terminology injection. Initial experiments provided prompts augmented with glossary (Moslem et al. 2023a; Moslem et al. 2023b) and dictionary (Ghazvininejad et al. 2023) entries, also following an instruction tuning (Kim et al. 2024), where the retrieved target terms are injected in the prompt. Another strategy relies on post-editing existing translations (Bogoychev and Chen 2023; Chen et al. 2024; Liu et al. 2025, Sabo et al. 2024), which can be included within the wider concept of translation refinement (Feng et al. 2024; Koneru et al. 2024; Xu et al. 2024). This technique employs iterative prompting to adjust the generated translation until the terminological constraint is complied with.

3 Experimental target and limitations

In light of these insights, we evaluate off-the-shelf models that have undergone both pre-training (EuroLLM) and supervised fine-tuning (Tower) on translation tasks without additional adaptation to custom data. However, by deliberately excluding any modification to the model's hidden states representation, we restrict our investigation to the effect of terminology injection — particularly for terms likely under-represented during pre-training — on output fluency and decoding behavior under soft constraint conditions. This design choice may limit the performance upper bound for producing fully coherent South Tyrolean text, as some of the most effective approaches for handling language varieties often involve large-scale pre-training or continued training (Tejaswi et al. 2024; Nag et al. 2024), for machine translation (Kumar et al. 2021; Sousa et al. 2025) and evaluation tasks (Sun et al. 2023; Aepli et al. 2023) alike.

4 Methodology

4.1 Dataset curation — The *LegISTyr* Dataset

General principles

To make a fully comprehensive assessment of the models' term recognition, we impose stylistic and textual criteria in the collection of the test set sentences. We choose exemplars with a minimum of 8 and a maximum of 50 words, ignoring titles, truncated excerpts, captions, contents of tables and indexes, and bullet lists. All sentences are copied or adapted (e.g., shortened) from existing sources, including the contexts from *bistro*. The examples showcase the term of interest in different positions of the sentence with the possible variations in number but not in gender, as this would require a gender-specific equivalent in German and add a further layer of complexity. Subject and object are well defined, added manually when they are implicit, which is a common feature in Italian. Unresolvable co-reference relationships or ambiguous anaphoric references may appear in some exemplars but never affect the term. Parenthetical statements within brackets or dashes have been removed while maintaining the typical style of Italian legal language, which tends to use appositions and parenthetical material between commas.

Terms can be simple terms (e.g., *cittadinanza*, citizenship; *frode*, fraud) or complex terms (e.g., *decreto ministeriale*, Ministerial Decree; *capacità di intendere e di volere*, full possession of mental faculties). Most are nouns or noun phrases, with the exception of one collocation (*d'ufficio*, ex officio), which has a standardized German translation in South Tyrol. Almost all selected terms are available in *bistro* with their South Tyrolean variants and any terms used in other German-speaking legal systems.

The content of *LegISTyr* is largely based on the terminological data contained in the Information System for Legal Terminology *bistro*[2] (Ralli and Andreatta 2018). The latter collects the main legal concepts of the Italian legal system with their designations in Italian and in South Tyrolean German for use at the regional level, together with existing German language designations for any equivalent concepts from other legal systems that use German as an official language (i.e., Austria, Switzerland and Germany). *bistro* publishes over 13,000 fully-fledged term entries pertaining to a wide range of legal subdomains, for several legal systems (Italy, Austria, Germany, Switzerland, EU law, international law) and three languages (Italian, German, Ladin). However, many contexts in *bistro* are defining contexts, which are extremely useful to complement conceptual information given in the

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

definitions but not always ideal to showcase a term in its most frequent usages. In addition, we wanted more than just one example (i.e., context) for each selected term. We therefore complemented *bistro* data with examples from legal texts and websites. The test set also contains data from subdomains that haven't been fully published in *bistro* to date (e.g. subsidised housing).

Dataset structure The dataset is divided into different sections:

- Standardized terminology: 250 sentences covering different legal subdomains (partly including but not limited to the subdomains in the other subsets), with 5 instances for each term with a mandatory, standardized equivalent in South Tyrolean German.
- Main terminology: 1,000 sentences from 4 different legal subdomains (criminal and criminal procedure law, family law, subsidized housing, occupational health and safety) with 5 instances for each term.
- Homonyms: 250 sentences with 5 term instances for each of the two or three sub-domains where the term is used.

Each entry comprises a source sentence, a source term and a target term. Alongside the main South Tyrolean term, other variants expressing the same concept in South Tyrol or in other German-speaking legal systems are specified — when existent. This lets us measure to what extent the interference from more represented legal systems impacts term translation. In Appendix A, the reader can find an illustrative example of a test set entry. Each entry is designed to evaluate the insertion of its corresponding designated target term alone; occurrences of other terms from the test set within the same sentence are disregarded for evaluation.

Additionally, the dataset contains a subset for abbreviated forms (i.e., acronyms, initialisms, and abbrevations), a frequent source of mistakes in legal translation. There is also a subset with strategies for gender-inclusive writing (e.g. gender-neutral agentives, split forms, terms with symbols and neomorphemes) because local legislation demands that legal and administrative texts be possibly inclusive. Since these two sections have not been used for this paper, we do not give further details.

Consistency We recognize that a key requirement for efficient terminology management is that term rendering be consistent throughout the entire document (Semenov and Bojar 2022). While our evaluation is conducted at the segment level, we approximate terminology consistency by designing a test set that includes five same-domain usage contexts for each term. This choice allows to evaluate over longer stretches of running text and simulate (insofar as possible) multiple occurrences in a text. It also helps assess the behavior of the model in the presence of domain-relevant context information systematically.

Homonyms We analyze the homonyms subset to monitor the circumstances when the notion of unambiguity (i.e., avoiding one-to-many translations) proposed in similar works (Bogoychev and Chen 2023) has to be dismissed. While the statement that source terms should be associated with one correspondent only is generally valid, we put the lens on homonymy to measure performances on a linguistic phenomenon that has proven extremely challenging to address. In these cases, the variability on the target side of a single surface-form source term is crucial to refer to the correct concept.

4.2 Model selection

4.2.1 Generic LLMs

Llama we use the **Llama-8B**-Instruct model, part of Meta AI's suite of open-weight transformer-based language models (Touvron et al. 2023). The 8B variant offers a balance between model capacity and computational efficiency. It has also been chosen because it is one of the base models onto which Tower suite models have been fine-tuned.

Mistral An open-weight model, **Mistral-7B**-Instruct is another decoder-only transformer architecture comparable to Llama in size. It achieves efficient inference thanks to sliding window attention and grouped-query attention (Jiang et al. 2023). It also serves as one of the base models for one of the Tower-suite fine-tuned models.

4.2.2 Translation-tuned LLMs

Tower (Alves et al. 2024) is a suite of multilingual LLMs fine-tuned to translation-related tasks. We test the two available configurations: **Tower-7B**-Instruct and **Tower-13B**-Instruct. The development of TOWER involves a two-stage process. Initially, the base model, TOWERBASE, undergoes continued pretraining upon the LLaMA-2 architecture (Touvron et al. 2023) on a 20-billion-token dataset comprising both monolingual and parallel data. Subsequently, the model is fine-tuned using a curated dataset, TOWERBLOCKS, which specializes the LLM for translation-related tasks. We also test an updated version of the model named **Tower-Mistral-7B**-Instruct (Rei et al. 2024) and based on Mistral-7b (Jiang et al. 2023).

EuroLLM (Martins et al. 2024) is a suite of open-weight multilingual language models trained from scratch and featuring all official EU languages. The models are trained on a filtered corpus of assorted web data, code, parallel corpora, and domain-specific texts. A byte-pair encoding (BPE) tokenizer is created to handle linguistic diversity and efficient subword segmentation. Pre-training is conducted with mixed multilingual objectives, followed by instruction tuning to enhance zero-shot and few-shot task performance. We test the configuration **EuroLLM-9B**-Instruct.

4.2.3 Neural systems

ModernMT is accessed via its adaptive API for enterprises with a licensed account. We use the unidirectional glossary function uploading all term pairs gathered in the test set (indicated as **ModernMT-glossary** in Table 1) and compare it with its baseline performances (**ModernMT-baseline**).

MarianMT We fine-tune the Italian to German version of the Opus-MT model[3], using the MarianMT architecture (Junczys-Dowmunt et al. 2018) through Hugging Face's Transformers library. The model was trained on an in-house parallel corpus including LEXB (Contarino 2021), MT@BZ (De Camillis et al. 2023), CATEX (Gamper 1999) and other internal translation memories — for a total of 223,716 training instances. As parameters, we use a batch size of 64, 15 epochs, a learning rate of 3e-4, and 5,000 warm-up steps. Mixed-precision is used. In Table 1, it is indicated as **MARIANMT-adapted**. We do not report the results of the baseline model as its inadequacy for the South Tyrolean variety has already been exposed in Oliver et al. 2024.

4.3 Experimental setup

We interface with the models via the vLLM inference framework[4]. Working with instruction-tuned models, we utilize the *chat* method for text generation[5] to make the best of their instruction-following capabilities. The vLLM framework automatically retrieves and applies the model's predefined chat template when processing chat-formatted inputs.

Appendix B contains the prompt structure. In the system message, we explain the task together with the expected features of the output (language variety and terminology awareness). The user message consists solely of the source sentence, enclosed within <> symbols, following the approach of Zhang et al. 2024 and Cettolo et al. 2024. This implicitly signals that the translation should also be enclosed within these delimiters, facilitating the exclusion of extraneous commentary. Because Tower Suite models do not rely on system messages for instructions due to their structured text generation pipeline, we provide the instruction and the appended source sentence in the user message for this class of models only. For the homonym set, all possible homonym translations are provided as options, without suggesting the correct one.

As custom inference parameters, we set top-p to 0.95 and temperature to 0.2. Low temperature is meant to limit output variability and ensure high-confidence text coherence in a domain that is rich in formulaic expressions and sensitive to meaning corruption, while a relatively high top-p setting allows potentially under-represented terms to remain into the sampling space.

[3] https://huggingface.co/Helsinki-NLP/opus-mt-de-it

[4] https://docs.vllm.ai/en/latest/

[5] https://docs.vllm.ai/en/latest/models/generative_models.html

4.4 Evaluation

Existing methods for measuring the enforcement of terminology into the output vary depending on the structure of available termbase resources, though always relying on some form of exact-match algorithm. While we acknowledge the limitations of a naive matching approach, the lack of a reference translation prevents us from implementing the solutions suggested by Alam et al. 2021a.

Hence, we define our evaluation criteria as follows:

- **Accuracy**: The frequency with which the target term appears in the output. This metric assesses whether the system has incorporated the instructed term in any form.
- **Fluency**: The estimated extent to which the term is used in a coherent, well-formed sentence. This criterion evaluates the overall linguistic fluency of the machine translation output.

To measure **accuracy**, we design a custom pipeline with two pre-processing steps. Given a target term (TT) and a target sentence (TS), we first apply lemmatization to all tokens in both groups. We then use the SpaCy PhraseMatcher[6] — which allows to match terminology lists from provided text — to determine whether TT appears in TS, returning a positive match if found. However, due to the high density and complexity of inflected forms in German, we observe the lemmatizer may occasionally struggle to identify the uniform lemma of inflected variants of complex terms across TT and TS. Therefore, we apply the *CharSplit* tool from Tuggener, 2016 to decompose all tokens into their most probable subcomponents. Lemmatization is then re-applied to these decomposed units, and the matching procedure from the first step is repeated. We find this additional step particularly beneficial to detect the cases of internal inflection within complex nouns. Subsequently, we adopt the same procedure for: alternative acceptable terms belonging to the South-Tyrolean variety, variants used in other German-speaking legal systems, and incorrect homonyms.

To assess **fluency**, we compute the COMET-Kiwi-XL score (Rei et al., 2023), a reference-less automatic machine translation quality estimation metric. This allows us to evaluate the overall quality of the translated segment beyond mere terminology injection. Following the holistic approach proposed by Alam et al. 2021a, this evaluation ensures that the imposed terminological constraints do not compromise the meaning of the generated sentence.

5 Results

As it can be appreciated in Table 1, generic LLMs (LLama and Mistral) achieve the highest term success rate, outperforming other model paradigms by at least 10 percentage points. However, this should not be naively accepted at face value as superior overall suitability. The well-documented tension between terminological accuracy and output fluency remains tangible. As the Comet-Kiwi-XL scores highlight, TT LLMs achieve higher levels of fluency at the system level. This divergence may suggest that generic LLMs fall into the patterns of hard-constrained decoding techniques, where the insertion of low-probability tokens causes a cascading degradation of output fluency by redistributing a lower probability mass across the whole output. In contrast, TT LLMs tend to insert the desired term only when its inclusion aligns with a high-probability token prediction, as determined primarily by the model's pretrained hidden state representations, rather than external prompt conditioning.

In addition, the magnitude of the COMET score differential is non-trivial. According to Kocmi et al. (2024)'s tool[7], a 2-point delta in COMET-Kiwi corresponds heuristically to a 95% agreement rate with expert human judgments. This result reinforces the hypothesis that TT LLMs prioritize fluent output more effectively than other model paradigms.

Homonym set results point to this hypothesis too. The difficulty of detecting the correct domain lowers overall term enforcement rates, with the most pronounced decline observed in generic LLMs, hence narrowing their performance lead. Notably, generic LLMs exhibit a strong proneness to inserting any of the provided terms, even when contextual cues suggest otherwise. This results in more than twice as many incorrect homonym selections compared to TT LLMs — with critical consequences on meaning comprehension.

It could be argued that this divergence may be

[6] https://spacy.io/api/phrasematcher

[7] https://kocmitom.github.io/MT-Thresholds/

	Accuracy					Fluency
Models	Main + Standardized Terminology			Homonyms		
	Term Success Rate	Other South Tyrol	Other Legal System	Correct Homonym	Wrong Homonym	Comet Score
LLAMA 8B	**64.5**	1.68	3.10	**44.8**	21.2	0.6317
MISTRAL 7B	64.0	1.36	**2.55**	41.6	25.2	0.5983
TOWER 7B	39.2	1.84	10.0	37.2	10.4	0.6264
TOWER-MISTRAL 7B	43.6	2.40	7.64	34.0	11.6	0.6395
TOWER 13B	52.0	1.84	6.0	36.4	12.4	0.6534
EUROLLM 9B	46.5	3.28	8.72	36.8	**7.6**	**0.6717**
MMT-baseline	24.1	**5.84**	12.7	28.8	15.2	0.6693
MMT-glossary	51.3	1.70	4.80	32.4	26.4	0.6343
MARIANMT-adapted	49.5	4.53	3.7	34.8	14.0	0.5757

Table 1: Evaluation results according to accuracy and fluency criteria. The **Main + Standardized Terminology** section reports performance on the Standardized Terminology and Main Terminology subsets of the test set, as described in Section 4.1. *Term Success Rate* indicates the percentage of cases in which the exact instructed term was correctly injected into the translation output. *Other South Tyrol* denotes the proportion of translations — out of the sheer total — using an acceptable South Tyrolean variant instead of the specified term. *Other Legal System* reflects the relative frequency — computed only on applicable cases — of a term from another major legal variety of German being used in place of the target South Tyrolean form. The **Homonyms** section reports results on the homonym subset. The *Correct Homonym* column shows the percentage of cases the correct homonym has been inserted, while the *Wrong Homonym* column highlights the percentage of cases the erroneous homonym option has been used in place of the correct one. Finally, under **Fluency**, the *Comet Score* column reports the system-level evaluation scores performed with Comet-Kiwi-XL.

attributable to the different objectives of the respective tuning procedures. Specific training of TT LLMs on translation-specific tasks enables the models to learn source-target text alignment patterns, allowing its attention mechanisms to more effectively focus on source-side information. The higher learned attention to highly domain-relevant context in the source sentence reduces the likelihood of incorrect term selection. In contrast, instruction-tuning in generic LLMs often lacks explicit exposure to the parallel sentence structure typical of translation tasks. Therefore, contextual attention weights may end up excessively biased towards less important parts of the prompt.

Additional insights emerge from the generation of terms belonging to other major varieties of legal German (OLS) or other valid alternatives for South Tyrol (OST), in the cases where these were measurable. The higher selection rate of OLS variants by TT LLMs may signal that prompting embeddings cannot alone redress the token probability imbalance in the output distribution, which is presumably outmatched by major variant occurrences. However, we note that EuroLLM-9B achieves the highest OST selection rate among LLMs, despite these acceptable terms not being explicitly elicited in the prompt. Given its pre-training from scratch on European language corpora, this result suggests that including regionally-focused pre-training data may allow to better encode minor variant lexical forms.

Format-wise, Tower and EuroLLM have proven to be most stable models, with all translations enclosed in the desired delimiters and no signs of verbosity. While LLama 8B struggles to follow the requested format, spurious text remains a rare occurrence. Eventually, Mistral 7B shows a considerable amount of noise in the output, most notably in verbosity proneness, output in English language and repetition of parts of the prompt.

6 Discussion

The process of integrating LLMs into translation workflows is still at an early stage, with a major divide opening between the use of general-purpose models — leveraging scale-driven emergent capabilities — as opposed to models fine-tuned through task-specific, specialized training. While we register a higher overall success rate for generic LLMs, we also suggest that models trained specifically on translation data show greater promise to attend to the peculiar challenges of context-sensitive translation. On the technical side, future research should continue to explore strategies for conditioning more reliable terminology generation at decoding time.

Another promising direction involves exploiting the contextual metadata available in Linguistic Linked Data (LLD) resources (especially multilingual examples of use of term entries) for continued pre-training or task-adaptive fine-tuning. In this respect, while not having used terminology data as LLD for terminology injection in this first stage, we consider it a necessary follow-up step. Terminological data in machine-readable formats are likely to become crucial resources for terminology injection in future. We have shown that terminological data can be used to partly make up for the lack of training data in minor language varieties. Where training data from major varieties risks overriding language use in minor varieties and even leading to critical mistakes in high-stakes domains like the legal domain, available terminological resources could help to partially fill the gap.

The training data used for NMT systems and LLM models are skewed towards high-resource world languages and language combinations containing English. Further research is needed into non-English language combinations, which are a routine part of the work of many bodies with legislative, administrative and judicial powers that affect citizens daily lives. To name some examples in Europe, there is translation from German into a minor legal variety of Slovene in Austria for the Slovene-speaking minority in Carinthia, translation from Croatian into a minor legal variety of Italian in Croatia, translation from Finnish into a minor variety of Swedish in Finland. All these and other minority communities would profit from efficient strategies for adapting machine translation to their specific varieties and/or from terminology injection to fine-tune translation results.

7 Conclusion and limitations

We are aware of the limitations of our first experiments. Although proprietary LLMs set the current upper bound for performance, the opacity surrounding their training data precludes any assessment of whether results stem from explicit exposure to translation tasks. This constraint limits our ability to isolate and attribute observed advantages to translation-specific (pre-)training, thereby confounding the validation of the research hypothesis.

From a linguistic perspective, our data illustrates phenomena that apply to the South Tyrolean standard variety of German and to a non-English language combination. To assess whether similar results would apply, for example, to the standard variety used by the German-speaking community in Belgium and to translation from Dutch or French to another variety of German, we would need targeted studies. The same holds true for other minor (legal) varieties of major languages (e.g., Swiss French, Chilean Spanish etc.) and the language combinations that might be predominant in other minor or minority variety contexts. A further limitation consists in the lack of qualitative analyses that could shed better light on the results, which we are planning for the future.

8 Acknowledgements

Dataset creation was partly financed by the Autonomous Province of Bozen/Bolzano–South Tyrol within the project SSL-Laien “Optimising expert-lay communication. Case study: e-learning modules of the Autonomous Province of Bolzano – South Tyrol” awarded via the Programme Agreement for Eurac Research for 2022-2024. This financing only concerns the subset on occupational health and safety.

References

Noëmi Aepli, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. 2023. A benchmark for evaluating machine translation metrics on dialects without standard orthography. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1045–1065, Singapore. Association for Computational Linguistics.

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021. Encouraging neural machine translation to satisfy terminology constraints. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1450–1455, Online. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. 2021a. On the evaluation of machine translation for terminology consistency. *CoRR*, abs/2106.11891.

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021b. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Transla-*

tion, pages 652–663, Online. Association for Computational Linguistics.

Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney. 2018. On the alignment problem in multi-head attention-based neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 177–185, Brussels, Belgium. Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Ulrich Ammon, Hans Bickel, and Alexandra N. Lenz. 2016. *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*, 2 edition. De Gruyter, Berlin.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM‘s translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Eleftheria Briakou, Zhongtao Liu, Colin Cherry, and Markus Freitag. 2024. On the implications of verbose llm outputs: A case study in translation evaluation. *Preprint*, arXiv:2410.00863.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mauro Cettolo, Andrea Piergentili, Sara Papi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024. MAGNET - MAchines GeNErating translations: A CALAMITA challenge. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1089–1093, Pisa, Italy. CEUR Workshop Proceedings.

Pinzhen Chen, Nikolay Bogoychev, Kenneth Heafield, and Faheem Kirefu. 2020. Parallel sentence mining by constrained decoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1672–1678, Online. Association for Computational Linguistics.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).

Elena Chiocchetti. 2021. Effects of social evolution on terminology policy in south tyrol. *Terminology*, 27(1):110–139.

Michael Clyne, editor. 1991. *Pluricentric Languages. Differing Norms in Different Nations. Different Norms in Different Nations*. De Gruyter Mouton, Berlin, Boston.

Antonio Contarino. 2021. Neural machine translation adaptation and automatic terminology evaluation: A case study on italian and south tyrolean german legal texts. Master's thesis, Università di Bologna, Bologna, Italy.

Flavia De Camillis and Elena Chiocchetti. 2024. Machine-translating legal language: error analysis on an italian-german corpus of decrees. *Terminology science & research*, 27(1):1–27.

Flavia De Camillis, Egon W. Stemle, Elena Chiocchetti, and Francesco Fernicola. 2023. The MT@BZ corpus: machine translation & legal language. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 171–180, Tampere, Finland. European Association for Machine Translation.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

M. Amin Farajian, Nicola Bertoldi, Matteo Negri, Marco Turchi, and Marcello Federico. 2018. Evaluation of terminology translation in instance-based neural MT adaptation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 169–178, Alicante, Spain.

Zhaopeng Feng, Ruizhe Chen, Yan Zhang, Zijie Meng, and Zuozhu Liu. 2024. Ladder: A model-agnostic framework boosting LLM-based machine translation to the next level. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15377–15393, Miami, Florida, USA. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Antonio Gambaro and Rodolfo Sacco. 2024. *Sistemi giuridici comparati*, 4 edition. UTET, Milano.

Johann Gamper. 1999. Encoding a parallel corpus for automatic terminology extraction. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 275–276, Bergen, Norway. Association for Computational Linguistics.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *Preprint*, arXiv:2202.11822.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

Iikka Hauhio and Théo Friberg. 2024. Mitra: Improving terminologically constrained translation quality with backtranslations and flag diacritics. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 100–115, Sheffield, UK. European Association for Machine Translation (EAMT).

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *Preprint*, arXiv:2302.09210.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Josef Jon, Michal Novák, João Paulo Aires, Dusan Varis, and Ondřej Bojar. 2021. CUNI systems for WMT21: Terminology translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 828–834, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in c++. *Preprint*, arXiv:1804.00344.

Sejoon Kim, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim, and Jorge Gimenez Perez. 2024. Efficient terminology integration for LLM-based translation in specialized domains. In *Proceedings of the Ninth Conference on Machine Translation*, pages 636–642, Miami, Florida, USA. Association for Computational Linguistics.

Rebecca Knowles, Samuel Larkin, Marc Tessier, and Michel Simard. 2023. Terminology in neural machine translation: A case study of the Canadian Hansard. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 481–488, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual refinement of translations: Large language models for sentence and document-level post-editing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2711–2725, Mexico City, Mexico. Association for Computational Linguistics.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 110–121, Online. Association for Computational Linguistics.

Surafel M. Lakew, Matteo Negri, and Marco Turchi. 2020. Low resource neural machine translation: A benchmark for five african languages. *Preprint*, arXiv:2003.14402.

Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.

Jiarui Liu, Iman Ouzzani, Wenkai Li, Lechen Zhang, Tianyue Ou, Houda Bouamor, Zhijing Jin, and Mona Diab. 2025. Towards global ai inclusivity: A large-scale multilingual terminology dataset (gist). *Preprint*, arXiv:2412.18367.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Heikki E. S. Mattila. 2018. Legal language. In John Humbley, Gerhard Budin, and Christer Laurén, editors, *Languages for Special Purposes: An International Handbook*, pages 113–150. De Gruyter Mouton.

Elise Michon, Josep Crego, and Jean Senellart. 2020. Integrating domain terminology into neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexander Molchanov, Vladislav Kovalenko, and Fedor Bykov. 2021. PROMT systems for WMT21 terminology translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 835–841, Online. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023b. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.

Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee, and Niloy Ganguly. 2024. Efficient continual pre-training of llms for low-resource languages. *Preprint*, arXiv:2412.10244.

Antoni Oliver, Sergi Alvarez-Vidal, Egon Stemle, and Elena Chiocchetti. 2024. Training an NMT system for legal texts of a low-resource language variety south tyrolean German - Italian. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 573–579, Sheffield, UK. European Association for Machine Translation (EAMT).

Minh Quang Pham, Josep Crego, Antoine Senellart, Dan Berrebbi, and Jean Senellart. 2021. SYSTRAN @ WMT 2021: Terminology task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 842–850, Online. Association for Computational Linguistics.

Matt Post, Shuoyang Ding, Marianna Martindale, and Winston Wu. 2019. An exploration of placeholding in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 182–192, Dublin, Ireland. European Association for Machine Translation.

Carla Quinci and Gianluca Pontrandolfo. 2023. Testing neural machine translation against different levels of specialisation: An exploratory investigation across legal genres and languages. *Trans-Kom. Journal of Translation and Technical Communication Research*, 16(1):174–209. Special Issue on Communicative Efficiency.

Natascia Ralli and Norbert Andreatta. 2018. Bistro – ein tool für mehrsprachige rechtsterminologie. *Trans-Kom. Journal of Translation and Technical Communication Research*, 11(1):7–44.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Marek Sabo, Judith Klein, and Giorgio Bernardinello. 2024. Boosting machine translation with AI-powered terminology features. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 25–26, Sheffield, UK. European Association for Machine Translation (EAMT).

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *Preprint*, arXiv:2310.11324.

Kirill Semenov and Ondřej Bojar. 2022. Automated evaluation metric for terminology consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 450–457, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

Suzanna Sia, David Mueller, and Kevin Duh. 2024. Where does in-context translation happen in large language models. *Preprint*, arXiv:2403.04510.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

H. Sousa, S. Almasian, R. Campos, and A. Jorge. 2025. Tradutor: Building a variety specific translation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25183–25191.

Michal Štefánik, Marek Kadlcik, and Petr Sojka. 2023. Soft alignment objectives for robust adaptation of language generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8837–8853, Toronto, Canada. Association for Computational Linguistics.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. Dialect-robust evaluation of generated text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6028, Toronto, Canada. Association for Computational Linguistics.

Atula Tejaswi, Nilesh Gupta, and Eunsol Choi. 2024. Exploring design choices for building language-specific LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10485–10500, Miami, Florida, USA. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Don Tuggener. 2016. Incremental coreference resolution for german. Master's thesis, University of Zurich, Faculty of Arts.

Elsbeth Turcan, David Wan, Faisal Ladhak, Petra Galuscakova, Sukanta Sen, Svetlana Tchistiakova, Weijia Xu, Marine Carpuat, Kenneth Heafield, Douglas Oard, and Kathleen McKeown. 2022. Constrained regeneration for cross-lingual query-focused extractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2668–2680, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Lucas Weber, Elia Bruni, and Dieuwke Hupkes. 2023. The icl consistency test. *Preprint*, arXiv:2312.04945.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024. LLMRefine: Pinpointing and refining large language models via fine-grained actionable feedback. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1429–1445, Mexico City, Mexico. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. Teaching large language models to translate with comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19488–19496.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: less is more for alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024a. Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 388–409, Miami, Florida, USA. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Example of a test set entry

Source sentence	Per i videoterminalisti che non sono esposti a rischi aggiuntivi, è prevista una formazione di base di 4 ore più 4 ore di formazione per il rischio specifico.
Source term	videoterminalista
Target term	Bildschirmarbeiter
Other terms from South Tyrol	Bildschirmverwender
Terms from other legal systems	Bildschirmarbeitnehmer

Table 2: Example of an entry structure taken from the test set.

B Prompt Structure

```
[
    {
        "role": "system",
        "content": "This is a translation task. Translate the
           ↪following legal text from Italian into South-Tyrolean
           ↪German. There are terminological constraints you must
           ↪adhere to: {term_it} corresponds to {term_de}. You must
           ↪output only the translated text without any explanation.
           ↪ This is the text to be translated into South-Tyrolean
           ↪German:"
    },
    {
        "role": "user",
        "content": "<{source_sentence}>"
    }
]
```

Listing 1: Prompt structure for the Standard Terminology set. For the homonym set, both possible homonyms are provided as options, without suggesting the correct one.