Terminology Resources as (MT) Benchmarks: Evaluation of Legal Terminology Challenges in German Language Varieties by EURAC

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

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

Abstract

Language varieties and dialects remain an open challenge for machine translation (MT). Plus, existing benchmarks rarely combine the evaluation of specific varieties with domain-specific knowledge. We introduce a termbase-derived MT benchmark for legal terminology across four German varieties: Germany, Austria, Switzerland, and South Tyrol (a low-resource German variety spoken in Italy). In the *inter-variety* setting, models translate Italian source sentences by selecting the correct South Tyrolean German term among semantically equivalent terms from other German varieties. In the *intra-variety* settings, we test sense disambiguation and ontological relation capabilities using homographic and ontologically related distractors for all four language varieties. We find that distinguishing language varieties remains difficult. High performance on low-resourced South Tyrolean terminology is partly driven by indirect exposure to terms shared with better-resourced varieties; when this overlap is removed, even the best-performing models collapse toward small models. We also find mild evidence of dominant-variety interference, although stronger systems more often err toward the geographically and culturally closer Austrian variety. Generally, locale-oriented adaptation appears more challenging than linguistic and semantic reasoning, where the ontological relation task is largely saturated and no generalized bias is shown against a specific variety. We conclude that model scaling mainly reinforces dominant language varieties and that LLM-based MT shifts part of the performance bottleneck from individual term rarity to the underrepresentation of domain-variety intersections across training and alignment stages.

1 Introduction

Recent WMT shared tasks have foregrounded language varieties and dialects as open challenges for general MT (Kocmi et al., 2024, 2025b) and MT evaluation (Lavie et al., 2025; Knowles et al., 2024). This is reflected in the growing adoption of more fine-grained language codes (Finkelstein et al., 2026), in the Open Language Data Initiative (Dale et al., 2025) to support the collection of multilingual resources for underserved languages and varieties, and in the Multilingual Instruction Shared Task (Kocmi et al., 2025a) for assessing locale-oriented generation. These initiatives have addressed persistent weaknesses in LLMs with respect to a long-standing problem in cross-lingual NLP (Zampieri et al., 2020).

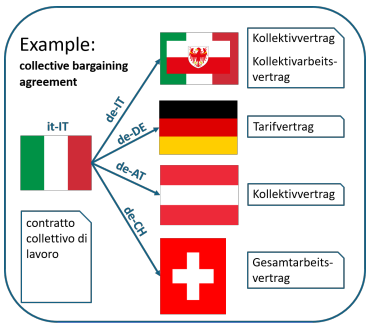


Figure 1: Example of term variation between varieties of German

The sparse and systematic signals of similar language varieties are notoriously complex to model with approaches other than n-gram-based algorithms (Jauhiainen et al., 2019) and even LLMs still lag behind smaller networks for language understanding tasks (Faisal and Anastasopoulos, 2025). Prior work suggests that such shortfalls originate in training data (Rystrøm et al., 2025), post-training alignment pipelines (Sahu et al., 2026), and evaluation protocols that are disproportionately English- and Western-centric (Singh et al., 2025b). Crucially, representing a language is not sufficient to represent its varieties. Benchmarks and datasets are often constructed at the language level, thereby collapsing dialectal, regional, and cultural varia-

*Corresponding author: pdinatale@unibz.it; pdinatale@eurac.edu

Proceedings of the Eleventh Conference on Machine Translation, pages 1506–1532
October 28-29, 2026 ©2026 Association for Computational Linguistics

tion into a single category (Plank, 2025; Gehrmann et al., 2023). This is particularly problematic for pluricentric languages that have several official standard varieties, where a shared language does not always equal a shared cultural or semantic ground (Chen et al., 2026). As a result, cultural misalignment is reflected in the loss of pragmatic markers and accessibility of culturally specific concepts (Kumar Navneet et al., 2026; Singh et al., 2025b). We posit that this loss may compromise conceptual accuracy to an ever greater extent in high-stakes domains like the legal one, that has notable social and economic relevance but where semantic misalignment can cause severe harm (Mattila, 2018).

We evaluate these limitations with a method based on structured terminological knowledge. We use *bistro*[1], an Italian-German legal terminology resource, as our knowledge base. It contains term equivalents, definitions, contexts of use, and other terminological information such as conceptual relations and usage notes. *bistro* covers the four German varieties used in Germany, Austria, Switzerland, and South Tyrol. The latter is a multilingual province in Italy where German is an officially recognized language with approximately 300,000 native speakers. This variety is low-resourced and lacks a dedicated ISO language code.

We process the termbase structure to derive a test set addressing three challenges. Each subset features an Italian source sentence and source term. Models are asked to produce a translation by selecting the appropriate term from challenge-motivated distractors. In the *inter-variety* setting, the LLM selects the correct South Tyrolean German equivalent from distractor terms belonging to other German varieties. Since these candidates designate equivalent concepts, they function as variety-specific distractors rather than semantic ones, the difference arising from distinct legal and institutional contexts (See Figure 1). In the two *intra-variety* settings, we assess model knowledge within each German variety through (1) sense disambiguation and (2) ontological relation tasks. In (1), the model chooses the German term corresponding to the intended sense of an ambiguous Italian term based on the sentential context. For example, *procedura concorsuale* should be rendered as *Insolvenzverfahren* (bankruptcy proceedings) in the context of commercial law and as *Wettbewerbsverfahren* (open competitive employment procedure) in administrative law. Whereas (2), tests whether models confuse the target term (e.g., *Strahlungsexposition* [exposure to radiation] in occupational health and safety) with ontologically related distractors such as hypernyms (e.g., *Exposition* [exposure]), co-hyponyms (e.g., *Lärmexposition* [exposure to noise]) or hyponyms (e.g., *UV-Exposition* [exposure to ultraviolet radiation]). This allows us to test whether variety-specific lexicalization affects the models' capability on linguistic and semantic challenges. Our findings are:

- Synthetic source sentences preserve term accuracy but inflate BLEU scores, meaning that they maintain the semantic cues needed for disambiguation while also making reference-based quality evaluation easier.
- Distinguishing South Tyrolean German from other German varieties remains challenging: even the strongest systems stay below saturation on the *inter-variety* subset.
- High performances on South Tyrolean terminology is partly driven by overlap with better-resourced German varieties: when terms are unique to South Tyrolean German, the advantage of model scaling is strongly reduced.
- *Intra-variety* linguistic challenges show no generalized bias against any specific variety, with the ontological relation task appearing largely saturated.
- For LLM-based MT, term rarity may no longer be one of the severest architectural deficiencies (Koehn and Knowles, 2017). Efforts should be directed towards tackling the socio-linguistic under-representation of specific domain-variety intersections over training and alignment stages.

2 Language varieties and legal translation

German—like Arabic, English, and Spanish, for example—is a pluricentric language with different codified standard forms that slightly diverge at the lexical, orthographic, and grammatical levels (Schuppler et al., 2024). South Tyrolean German is a non-dominant variety with co-official status in a part of Italy (Ammon et al., 2016). This German variety is underrepresented in NLP training corpora, as the stronger numeric signal from

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

dominant varieties tends to override non-dominant forms in statistical machine translation systems (Sousa et al., 2025). The challenge is substantially aggravated when such varieties are used in legal translation, where language is fundamentally system-bound: each German-speaking legal system (e.g., Germany, Austria, Switzerland, Italy, Belgium, the EU) has distinct legal concepts and standard terminology (Biel and Kockaert, 2023; Cao, 2007), with misalignments across systems being very frequent. Such misalignments manifest at multiple linguistic levels: source concepts may lack target counterparts, identical designations may refer to different concepts across legal systems, and incongruities may exist even across systems that share the same official language. For instance, a collective bargaining agreement is called *Kollektivvertrag* in Italy and Austria, *Gesamtarbeitsvertrag* in Switzerland, and *Tarifvertrag* in Germany (Figure 1). As the example shows, terms can differ but also coincide across legal systems.

3 Terminology

Terminology science studies terminologies (sets of designations and concepts belonging to one domain or subject), aspects of terminology work, the resulting terminology resources, and terminological data. Terminology work aims to identify and structure domain-specific concepts, their conceptual relations, and the terms used to designate the concepts (Costa, 2013). It serves to represent the conceptualizations of domain experts and facilitate the labeling and categorization of world knowledge in specialised communication. A fundamental aspect distinguishes terminology resources from lexicographic resources: while lexicography organizes information around words (lemmas) and their usage descriptions, terminology organizes information around concepts (ISO 704:2022). A concept-oriented approach takes the legal concept itself as its point of departure (e.g. the concept of a contract as a binding agreement) and associates it with the terms used to designate that concept in each language (e.g. contrat in French, Vertrag in German and contratto in Italian). However, apparent correspondence across languages does not necessarily entail conceptual equivalence across legal systems. This distinction becomes crucial when data need to be exchanged, modeled, or computationally processed, as a concept-oriented organization preserves sense distinctions between homonymous entries and groups synonyms and variants. Termbases structure information around persistent concept identifiers that anchor system-specific information (Di Nunzio and Vezzani, 2026), situate concepts within concept systems and ontological relations, and include cross-lingual mappings (Condello et al., 2026).

4 Termbases as Benchmarks

Factuality, truthfulness, grounding The quality standards for human curation of termbases align with criteria of NLG evaluation requiring outputs to be truthful, factual, and grounded in definitions and usage examples from authoritative sources (Tedeschi et al., 2023) based on verified world knowledge. Such expert validation has been found to be one of the strongest predictors of resistance to benchmark saturation (Akhtar et al., 2026).

Extensive coverage Termbases aim for extensive coverage of domain knowledge, ideally allowing to benchmark LLM performance across the full mapping of knowledge representation (Rajan et al., 2024). This has implications for experimental designs in machine translation. Validation and test partitions are often randomly sampled or designed to reproduce expected feature frequencies in realistic scenarios, though often without any certainty that relevant phenomena are equally and adequately represented (Søgaard et al., 2021) or that the selected samples are challenging enough (Proietti et al., 2025). This may neglect proper test-time supervision on long-tailed distributions and infrequent terms (Raunak et al., 2020, 2022; Tokarchuk et al., 2026; Saunders and DeNeefe, 2024). Moreover, singletons in training are a major cause of hallucinations (Kalai et al., 2025), so they would require targeted testing. By constructing test sets from termbase entries with informed data splitting approaches (Gehrmann et al., 2023), we can ensure that the test set contains at least one instance of each concept from the knowledge map.

Serving industry and community Benchmarks should be designed with the needs of the intended users in mind. For industry stakeholders, MT evaluation prioritizes post-editing effort, deployment costs, and client-specific terminology (Wang et al., 2026b). Language service providers commonly maintain in-house termbases for translation and quality-assurance workflows. Converting these resources into benchmarks enables evaluation on ter-

minology that is operationally relevant, customizable through export filtering, and cost-effective in repurposing existing resources.

Language communities—including language learners, lay MT users, and speakers of underrepresented varieties—are often concerned with ethical risks, social bias, and factual reliability (Savoldi et al., 2025; Wang et al., 2026b). From this perspective, targeted benchmarking from local (AI) watchdogs can play a role in exposing bias affecting speakers of non-dominant varieties and minority communities in multilingual countries (such as the South Tyrolean population). Decentralizing evaluation through local actors can also mitigate LLM leaderboard distortions caused by undisclosed private testing, selective score disclosure, and asymmetric access to evaluation data (Singh et al., 2025a; Kocyigit et al., 2025). Such local benchmarks can support the education of uninitiated audiences about the risks of conceptual imprecision when implementing AI output in official contexts and help identify systems that are less biased for deployment in real use cases even beyond MT (Biel, 2026).

Natively multilingual A common practice for multilingual LLM evaluation is to translate English benchmarks into other languages (Rajaee et al., 2025). However, this can limit cultural representativeness and reproduce English-centric assumptions, particularly when target-language concepts have no direct English counterpart (Singh et al., 2025b). In the legal domain, institutions in common-law systems may not have direct counterparts in civil-law systems and vice-versa, for example. Even countries sharing the same language may define and regulate legal concepts differently. Some benchmarks are originally developed in the target language (Zhang et al., 2024) or localized (Ong and Limkonchotiwat, 2023). Current legal benchmarks are predominantly English-centric (i.e., favoring US law) and largely quantitative in scope (Chalkidis et al., 2022). They also neglect comparative law issues and do not intersect legal aspects with language varieties of different countries. Termbases can address this gap because they organize and represent specialized knowledge natively within the legal and linguistic systems in which that knowledge is produced and used. They therefore allow us to identify where and why cross-lingual term rendering may fail.

MT as a world knowledge probe In standard knowledge benchmarking frameworks, knowledge is usually probed from a graph by verbalizing queries that LLMs must resolve. Query types can include a question-answer pattern (Talmor et al., 2019), predicting the answer using probability-based (Wiland et al., 2024) and confidence-based scores (Bexte et al., 2024), or selecting the correct statement from a multiple-choice item (Wang et al., 2026a), where distractors can be retrieved from node relations (Ren and Zhu, 2021) or top-k adversarial attacks (Luo et al., 2025). However, these approaches are often sensitive to prompt format (Leidinger et al., 2023), unreliable verbalizations (Xiong et al., 2024), difficulty of evaluating free-form answers (Bavaresco et al., 2025; Verga et al., 2024).

We propose to define MT as an extrinsic evaluation task for cross-lingual knowledge in specialized domains, leveraging structured knowledge encoded in termbases. The prompt and source text provide contextual information for concept selection. Accurate term realization shows whether the model can access, transfer, and express domain knowledge during text generation.

5 Method

5.1 Test set creation

We use *bistro*[2], which collects approximately 14,000 Italian legal concepts with their terms in Italian and South Tyrolean German and their equivalent terms in the other German-speaking legal systems (Ralli and Andreatta, 2018). For each concept c_i, we organize a language level l, comprising Italian l^{IT} = $\mathrm{IT}^{\mathrm{it}}$ and the German varieties l^{DE} = $\{\mathrm{DE}^{\mathrm{it}}, \mathrm{DE}^{\mathrm{de}}, \mathrm{DE}^{\mathrm{at}}, \mathrm{DE}^{\mathrm{ch}}\}$ (Ralli and Tamás, 2025). At a lower level, each language or variety contains a set of terms $t_k = \{t_1, \ldots, t_n\}$. Each term variant is associated with a real-world usage context $ref(t_n)$, selected to represent a prototypical use of the term. The structure of the knowledge base can therefore be summarized as follows:

$$c_i \mapsto l^{\{IT,DE\}} \mapsto \{\, t_n, ref(t_n) \,\}$$

5.2 Processing of subsets

For each subset, we extract tuples x, including a source term, target terms, term distractors, and a reference sentence.

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

Varieties subset (VAR) We focus on South Tyrolean German terms. For each concept c_i, we select the target terms $t_{i.k}^{\mathrm{DE}_{it}}$ and only one of the Italian source terms as $t_{i,n}^{\mathrm{IT}_{\mathrm{it}}}$. Distractors are selected among the non-identical terms attested in the other German varieties such that $d_{ik} = \{t_{i,k}^{\mathrm{DE}_{\{\mathrm{de,at,ch}\}}}\} \neq t_{i,k}^{\mathrm{DE}_{\mathrm{it}}}$. The reference sentence is the usage context associated with the selected South Tyrolean German target terms $ref(t_{i,k}^{\mathrm{DE}_{\mathrm{it}}})$.

$$x_{ik} = \left(t_i^{\mathrm{IT}_{\mathrm{it}}}, t_{i,k}^{\mathrm{DE}_{\mathrm{it}}}, d_{ik}, ref\left(t_{i,k}^{\mathrm{DE}_{\mathrm{it}}}\right)\right)$$

Homonyms subset (HOM) We identify Italian terms that are shared across multiple concept identifiers, so that t_{ij}^{IT} lexicalizes more than one concept: $t_{i,n}^{\mathrm{IT}} = t_{j,n}^{\mathrm{IT}}$. For each homonym item, we use the shared Italian term as the source term and pair it with its German target terms and reference contexts, $ref(t_{i,k}^{\mathrm{DE}})$ and $ref(t_{j,k}^{\mathrm{DE}})$. For each German variety *var*, we extract the reference $ref(t_{i,k}^{\mathrm{DE}_{var}})$. Then we choose distractors from all possible German terms in the other concept identifier $d = t_{j,k}^{\mathrm{DE}}$. We end up with:

$$x_{ik} = \left(t_{ij}^{\mathrm{IT}}, t_{i,k}^{\mathrm{DE}_{var}}, d_{j,k}^{\mathrm{DE}}, ref(t_{i,k}^{\mathrm{DE}_{var}})\right)$$

Ontological relations subset (REL) We select concept identifiers that participate in at least one ontological relation $r(c_i, c_j)$. For each concept c_i, we select target terms $\mathrm{t}_{\mathrm{i,n}}^{\mathrm{DE}_{\mathrm{var}}}$ that belong to only one German variety *var* and are not attested in other German varieties. The Italian term $t_{i,n}^{\mathrm{IT}_{\mathrm{it}}}$ is used as the source term, while distractors are obtained from the related concepts by extracting terms in the same German variety $\mathrm{t}_{\mathrm{j,k}}^{\mathrm{DE}_{\mathrm{var}}}$.

$$x_{ik} = \left(t_{i,k}^{\mathrm{IT}}, t_{i,k}^{\mathrm{DE}_{var}}, d_{j,k}^{\mathrm{DE}_{var}}, ref(t_{i,k}^{\mathrm{DE}_{var}})\right)$$

5.3 Data filtering

We first apply a check on reference sentences to ensure that target term matching is recoverable (and eliminate any evaluation error intrinsic in the metric). Using the same matcher later used for scoring, we retain only reference sentences in which the target term has been identified by our evaluation pipeline. This removes cases with irretrievable matching, e.g., due to the use of gender-fair strategies, incorrect lemmatization or discontinuous term realizations. We also discard instances where any distractor is matched. The final test set contains 1,662 VAR instances, 1,212 HOM instances and 527 REL instances.

5.4 Creation of synthetic sources

Since contexts of use across languages are not necessarily parallel in termbases, we create source sentences by machine-translating the extracted German references into Italian. We choose not to create synthetic references by translating authentic source sentences because synthetic references are found to heavily affect evaluation robustness (Freitag et al., 2022). For source generation, we identify a sample of 314 authentic, human-written sentence pairs from the VAR subset and 89 from the HOM subset, which we use to compare candidate translation models against authentic Italian references. Gemini 2.5 Flash achieves the highest BLEU score from the models tested in Di Natale et al. (2026) and is therefore selected to generate the synthetic Italian sources.

5.5 Evaluation metrics

Term accuracy is computed with a custom pipeline[3] (Di Natale et al., 2026) that sequentially applies four matching strategies to catch the morphological richness of German: surface-form, lemmatization, splitting compound nouns, and hand-coded inflectional rules. For fluency metrics, we report BLEU through SacreBLEU[4], given the limitations of neural metrics on this language combination (Di Natale et al., 2025).

6 Results

6.1 Effect of synthetic sources

We assess whether using synthetic Italian sources affects model performance by comparing them with the human-authored parallel sources already presented in Section 5.4. Across both subsets, term accuracy remains largely stable between authentic and synthetic sources (see Appendix B), including in the homonym setting where sentential context is crucial for sense disambiguation (although the small sample weakens statistical significance). For our specific purposes, synthetic sources do not substantially affect the models' ability to identify the intended concept.

However, fluency metrics reveal a systematic skew: BLEU is consistently higher for synthetic sources, with a constant advantage of at least five

[3] **Evaluation code**: https://github.com/PaoloDiNatale/LegISTyr-evaluation **Version hash**: e65f284

[4] **Signature**: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.6.0

points. We attribute this gap to translationese effects. Since the synthetic Italian sources are produced from the German reference using MT, they tend to preserve clause structure and word-by-word correspondences that are easier to back translate into the structure of the reference (Dutta Chowdhury et al., 2020). Another reason for divergence may be context. Authentic sources originate from document-level drafting and may rely on surrounding context, whereas synthetic sources are generated from isolated sentences. This can induce a different degree of explicitation and word selection (Castilho et al., 2021).

We therefore treat synthetic sources as adequate for probing terminological knowledge, but not as neutral inputs for overall quality estimation, as they provide a structural scaffold that can inflate BLEU scores. We recommend considering BLEU only for model comparisons within the same subset.

6.2 General findings

Overall results for term accuracy are shown in Table 1. The best-performing systems are Gemini 3.1 Pro, GPT 5.5, DeepSeek V4 Pro and TRIVE, with Dubformer and UvA-MT also performing competitively on some challenges. The VAR subset is the most difficult setting, especially for larger models, whereas homonym disambiguation and especially ontological relations yield higher term accuracy. Notably, model scale is less predictive on the VAR subset: medium and smaller systems occasionally narrow the gap with top-tier models compared with the other challenges. By contrast, the homonym setting discriminates more clearly between models, including within the same model family.

BLEU scores largely follow the expected pattern, mostly within the same model family, of increasing with model size and term accuracy (Table 4). We calculate a Pearson's r correlation between BLEU scores and Term accuracy rate for each variety and subset intersection (Table 4). Correlations are strongly positive and consistent across subsets. Models better able to access and realize variety-specific knowledge also tend to replicate the wording of that variety.

6.3 Variety subset

We evaluate South Tyrolean German as the target variety, using terms from the other German varieties as distractors. Even the strongest models remain below 80% term accuracy, while smaller models cluster around 50-60% (Table 1). This suggests that performance on this challenge is not yet saturated. No system inserts distractor terms more often than the target term, including models that do not receive the target terms in the prompt (GoogleTranslate). This suggests that interference from other German varieties is not predominant and that South Tyrolean contexts are sufficiently discriminative for term selection.

Distractor analysis When the South Tyrolean target term is not retrieved, we trace the variety provenance of the inserted term (Table 5). Since some terms are shared across varieties, attributions are not mutually exclusive and therefore do not sum to 100%. They indicate the proportion of distractor instances from that variety that were selected. Overall, terms from Germany are selected most often, while Swiss terms the least, suggesting a slight bias toward the dominant German variety. Interestingly, the strongest models more frequently err toward Austrian terms, presumably by virtue of the cultural and geographical proximity between Austria and South Tyrol. This suggests that stronger models may fall back on neighboring German locales when the South Tyrolean term is missing from their knowledge.

Unique South Tyrolean terms Many terms used in South Tyrol are also attested in other German varieties. We therefore isolate terms that are unique to South Tyrolean German (363 instances) to test whether model performance partly depends on exposure through better-resourced varieties. On this subset, performance drops sharply, and the decline is most pronounced for larger models (Table 7). Gemini 3.1 Pro and TRIVE remain the strongest systems, whereas Dubformer and DeepSeek V4 Pro drop to the level of much smaller models. Within model families, including GPT, Qwen, and Gemma, smaller models even outperform, or nearly do, medium-sized ones.

These results suggest that model scaling does not yield proportional gains for genuinely low-resource, variety-specific legal concepts. This glass ceiling for low-resource concepts may reflect data shortage in the specific domain-variety intersection. By contrast, larger models can continue to improve on higher-resource varieties because additional data reinforces support for those concepts. The use of synthetic data augmentation may further amplify this imbalance by reproducing dominant-variety forms.

Table 1: Term accuracy rates for all subsets. Mean is not indicated for the VAR subset because it would not be comparable to the other subsets.

	VAR	HOM				REL			
Model									
Cohere CAT+	53.1	62.4	63.8	68.0	65.2	82.0	86.1	82.4	84.4
Command A+	56.6	62.1	49.4	64.8	58.2	81.3	55.7	64.7	76.6
CUNI-EdUKate	27.7	22.2	27.1	21.6	31.6	28.9	18.0	23.5	17.2
CUNI-MH-v3	30.6	28.0	28.4	26.0	34.8	30.4	27.0	29.4	25.0
DeepSeek V4 Pro	64.0	75.7	82.0	82.8	81.6	90.1	**95.1**	89.7	93.8
Dubformer	63.7	65.9	74.6	77.2	71.5	89.0	92.6	92.6	89.1
Gemini 3.1 Pro	**76.4**	77.8	86.5	87.2	**88.3**	**94.9**	94.3	**95.6**	**96.9**
Gemma 4 - 4B	54.6	50.3	53.0	50.0	51.6	78.0	77.9	63.2	68.8
Gemma 4 - 31B	61.0	64.7	72.9	69.2	65.2	85.0	85.2	80.9	87.5
GLM 4 - 9B	45.5	38.2	39.5	34.0	33.6	62.6	49.2	47.1	51.6
GoogleTranslate	36.0	29.8	32.0	27.2	37.9	33.3	35.2	33.8	28.1
GPT 5.5	72.3	**80.1**	**86.7**	**88.8**	87.5	**94.9**	94.3	**95.6**	**96.9**
GPT OSS 20B	54.7	57.5	53.9	56.4	54.7	72.5	73.8	73.5	68.8
GPT OSS 120B	56.9	64.2	66.3	68.4	63.7	76.9	71.3	77.9	78.1
Ministral 3 - 14B	54.5	48.6	52.2	51.2	49.2	73.6	74.6	82.4	76.6
Mistral Medium 3.5	61.9	57.5	59.4	63.2	58.6	84.6	85.2	86.8	81.2
Qingqiu-MT-9B	51.7	57.5	62.2	60.0	59.0	78.8	70.5	63.2	67.2
Qwen 3.5 - 9b	48.0	48.3	51.7	48.0	52.0	70.3	69.7	60.3	70.3
Qwen 3.6 - 27b	55.8	54.0	61.3	61.2	53.1	78.8	80.3	76.5	82.8
SalamandraTA	51.7	54.9	60.2	54.8	54.7	67.4	68.8	63.2	68.8
Tiny Aya Global	42.2	35.8	40.1	36.8	38.3	51.3	50.0	44.1	43.8
Tower 9B	49.1	49.4	51.7	48.8	52.7	74.4	77.9	58.8	71.9
TRIVE	72.2	78.0	**86.7**	86.8	85.6	94.5	94.3	94.1	95.3
UvA-MT	59.9	67.9	74.0	79.2	72.7	91.6	86.9	86.8	89.1
Mean	N.A.	55.4	59.0	58.8	58.4	73.6	71.4	69.4	71.2
Instances	1,662	346	362	250	256	273	122	68	64

Term frequency correlation Rare terms have long been a known source of MT errors, typically linked to their scarce representation in training data. For LLMs, however, this relation is harder to measure: terms may appear across multiple domains in much larger generic corpora and closed models do not expose their training data. As a rough proxy, we use term frequency in an internally compiled corpus of South Tyrolean legal texts and correlate it with correct term insertions using point-biserial correlation, a variation of Person's correlation for binary variables.

We find no meaningful correlation between corpus frequency and term accuracy (Table 8). This suggests that LLM term selection failures cannot be explained by the statistical representation of terms in the target domain. Instead, performance appears to depend more strongly on whether the relevant concept belongs to a relatively low-resourced domain-variety intersection, as observed in the previous paragraph.

6.4 Homonym subset

Performance on homonym disambiguation is broadly stable across varieties. South Tyrolean terms remain slightly behind the others, and the pattern observed in the variety subset reappears: the strongest models show a larger relative drop in performance on South Tyrolean terms than on terms from the other varieties, whereas smaller models from the same families do not exhibit the same drop. Command A+ is an exception, showing a drop in accuracy for Germany's variety. We notice that models specifically trained for MT on parallel sentences (CUNI, GoogleTranslate, Tower)

tend to have a higher accuracy rate for the Swiss variety. We justify this with the fact that both Italian and German are official languages in Switzerland, thus a relatively larger size of parallel training data is available for this language combination. Parallel data appears to improve sense disambiguation.

6.5 Ontological relations subset

Performance is broadly comparable across varieties, with a small South Tyrolean advantage. Unlike in other subsets, larger models do not underperform on South Tyrolean terms, suggesting that this task depends less on variety-specific legal knowledge and more on general semantic reasoning. Models below 9B parameters tend to underperform on the Austrian variety, while Command A+ again shows an unexpected performance drop on Germany's variety.

Distractor analysis We analyse the type of ontologically related concept that is inserted when the target term is not identified (Table 6). Hyponyms are the most challenging distractors, suggesting that models struggle when the incorrect option is more specific than the target concept. Smaller and transformer-based systems tend to insert hypernyms more often. This supports a tentative interpretation that, when uncertain, models may prefer more generic rather than more specific concepts.

6.6 Legal subdomain

We also dissect results by legal subdomain for the VAR and HOM subsets (Tables 9 and 10). Domain-level differences are present but not systematic. The ranking of legal areas changes across subsets, suggesting that legal subdomain is not an independent predictor of performance. If anything, we notice that in the VAR subset, lower accuracy in procedural areas may reflect the difficulty of representing terminology from specific legal systems, whereas higher scores in public and constitutional law may benefit from stronger inter-institutional standardization. In the HOM subset, domain effects are flatter, indicating that once sentential context supports sense disambiguation, performance depends less on the legal area than on the separability of the competing concepts.

6.7 Number of Distractors

We evaluate whether performance varies with the number of distractors (See Appendix H), considering at least 50 instances for each distractor class. Unsurprisingly, term accuracy tends to decrease as the number of distractors increases for all subsets. The decline is less pronounced in the REL subset, likely because this task is less challenging than inter-variety term selection. Overall, this pattern suggests that the distractors are sufficiently competitive in the other subsets and that models struggle to handle a growing number of plausible options. We find no clear tipping point, but the largest degradation occurs around three and five distractors.

7 Discussion

Broader implications for benchmark design and LLM-based MT can be drawn. First, synthetic data may be useful for reconstructing source sentences when only target-side references are available, but the intuitive informativeness of the Bleu score should be questioned. Second, locale-oriented adaptation appears more challenging than linguistic and semantic reasoning tasks. Finally, model scaling mainly reinforces dominant language varieties: the overperformance of very large LLMs is less marked and plateaus on terms that are not shared with other language varieties. This trend is not correlated with word frequency, long held as the primary cause of ineffective learning of rare terminology (Koehn and Knowles, 2017). Instead, the capability of LLMs to learn from large monolingual data (Pang et al., 2025) reframes the challenge as one of data composition rather than data volume: progress will likely depend on how well linguistic variety and domain-aware representation are balanced in the data mix used for pretraining, augmentation, and alignment.

8 Future outlooks

As recent research has stated that we are approaching near-saturation of terminology injection at the sentence level (Semenov et al., 2025), terminology resources such as termbases can trace new research directions addressed to the design of shared tasks and future studies on terminology-aware MT.

Harder test sets: Following the trend of Proietti et al. (2025), this work can inspire the creation of more challenging test sets. Many of the shortcomings described in Section 4 also affect some of the studies in the field. Retrieval of gold terms for previous WMT shared tasks has been based on initial LLM-based extraction (Semenov et al., 2025), which likely favors high-confidence, easier

terminology (Cirillo et al., 2026). Random splits can further overrepresent frequent—and therefore easier-to-model—terms (Oncevay et al., 2025). The extensive coverage of termbases helps address these shortfalls by cataloging terms across all frequency bands. Additionally, while Semenov et al. (2025) has relied on randomly selected distractor lists for terminology dictionaries, other criteria—such as cosine distance, locale, or homonymy relations—could yield more discriminative distractor sets.

Keeping our tails: (Vanmassenhove, 2026) Long-tailed terminology remains problematic in zero-shot inference and during fine-tuning. To identify which fine-tuning methods best support rare terminology in shared tasks, we recommend releasing training corpora and characterizing how term frequency in the corpora relates to term accuracy across participants' training regimes.

New frontiers: As the current terminology setup approaches saturation, the field should move towards knowledge-based training. Rather than relying only on bilingual term lists or parallel corpora, participants could be asked to exploit monolingual definitions, usage notes, knowledge graphs, and domain metadata to train systems that aim to be aware of term meaning and not only term form. This would make terminology shared tasks a test bed for knowledge-aware MT, especially in domains and varieties where parallel data is scarce.

References

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, Jennifer Mickel, Leshem Choshen, Marek Šuppa, Jan Batzner, Jenny Chim, Jeba Sania, Yanan Long, Hossein A. Rahmani, Christina Knight, and 18 others. 2026. When ai benchmarks plateau: A systematic study of benchmark saturation.

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.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Marie Bexte, Andrea Horbach, Lena Schützler, Oliver Christ, and Torsten Zesch. 2024. Scoring with confidence? – exploring high-confidence scoring for saving manual grading effort. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 119–124, Mexico City, Mexico. Association for Computational Linguistics.

Łucja Biel and Hendrik J. Kockaert. 2023. Introduction: Legal terminology. In Łucja Biel and Hendrik J. Kockaert, editors, *Handbook of Terminology: Volume 3. Legal Terminology*, volume 3, pages 1–14. John Benjamins, Amsterdam and Philadelphia.

Łucja Biel. 2026. Ai literacy for legal translation: Developing digital resilience. *Preprint*, arXiv:2608.04641.

Deborah Cao. 2007. *Translating Law*, volume 33 of *Topics in Translation*. Multilingual Matters, Clevedon.

Sheila Castilho, João Lucas Cavalheiro Camargo, Miguel Menezes, and Andy Way. 2021. DELA corpus - a document-level corpus annotated with context-related issues. In *Proceedings of the Sixth Conference on Machine Translation*, pages 566–577, Online. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Pinzhen Chen, Koel Dutta Chowdhury, Xiaoya Xu, David Tan, Doreen Osmelak, Ona de Gibert, Ariun-Erdene Tumurchuluun, Ashok Urlana, Fedor Sizov, Hale Sirin, Jesujoba Alabi, Karrar Talib Abed, Mateusz Klimaszewski, Nikolay Bogoychev, Niyati Bafna, Patricia Schmidtova, Preksha Manjunath Shanbhag, Sherrie Shen, Vilem Zouhar, and 4 others. 2026. Cultivar: A contrastive and locale-oriented translation benchmark for investigating contamination and localisation robustness. *Preprint*, arXiv:2608.09766.

Nicola Cirillo, Giorgio Maria Di Nunzio, and Federica Vezzani. 2026. ATE-IT at EVALITA 2026: Overview of the automatic term extraction italian testbed task. In *Proceedings of the 9th Evaluation Campaign of Natural Language Processing and*

Speech Tools for Italian: Final Workshop (EVALITA 2026), volume 4195 of *CEUR Workshop Proceedings*, Bari, Italy. CEUR-WS.org.

Angela Condello, Giorgio Maria Di Nunzio, and Federica Vezzani. 2026. Law, language, and terminology: Challenges of generative AI. *International Journal for the Semiotics of Law – Revue internationale de Sémiotique juridique*, pages 1–15.

Rute Costa. 2013. Terminology and specialised lexicography: two complementary domains. *Lexicographica*, 29(2013):29–42.

David Dale, Laurie Burchell, Jean Maillard, Idris Abdulmumin, Antonios Anastasopoulos, Isaac Caswell, and Philipp Koehn. 2025. Findings of the WMT 2025 shared task of the open language data initiative. In *Proceedings of the Tenth Conference on Machine Translation*, pages 495–502, Suzhou, China. Association for Computational Linguistics.

Paolo Di Natale, Elena Chiocchetti, Marlies Alber, and Egon Waldemar Stemle. 2026. Beyond simple term injection: Reasoning models for legal translation in a non-dominant language variety. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation*, Tilburg, the Netherlands. European Association for Machine Translation.

Paolo Di Natale, Elena Chiocchetti, and Egon Waldemar Stemle. 2025. Meta-evaluation of automatic machine translation metrics between Italian and a minor language variety of German. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 371–383, Cagliari, Italy. CEUR Workshop Proceedings.

Giorgio Maria Di Nunzio and Federica Vezzani. 2026. When fuzziness is in the mind, not in the concept: Terminology, prototypes, and ontological modeling. SSRN preprint. Accepted for presentation at TOTH 2026.

Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2020. Understanding translationese in multi-view embedding spaces. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6056–6062, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fahim Faisal and Antonios Anastasopoulos. 2025. Testing the boundaries of LLMs: Dialectal and language-variety tasks. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 68–92, Abu Dhabi, UAE. Association for Computational Linguistics.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. Translategemma technical report. *Preprint*, arXiv:2601.09012.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, 77.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: a survey. *J. Artif. Int. Res.*, 65(1):675–682.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *Preprint*, arXiv:2509.04664.

Rebecca Knowles, Samuel Larkin, and Chi-Kiu Lo. 2024. MSLC24: Further challenges for metrics on a wide landscape of translation quality. In *Proceedings of the Ninth Conference on Machine Translation*, pages 475–491, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. 2025a. Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 414–435, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025b. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow,

Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. 2025. Overestimation in llm evaluation: a controlled large-scale study on data contamination's impact on machine translation. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *Preprint*, arXiv:1706.03872.

Satyam Kumar Navneet, Joydeep Chandra, and Yong Zhang. 2026. When ai writes, whose voice remains? quantifying cultural marker erasure across world english varieties in large language models. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26, New York, NY, USA. Association for Computing Machinery.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

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.

Chu Fei Luo, Ahmad Ghawanmeh, Bharat Bhimshetty, Kashyap Murali, Murli Jadhav, Xiaodan Zhu, and Faiza Khan Khattak. 2025. Red-teaming for inducing societal bias in large language models. *Preprint*, arXiv:2405.04756.

Heikki 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, Berlin and Boston.

Arturo Oncevay, Charese Smiley, and Xiaomo Liu. 2025. The impact of domain-specific terminology on machine translation for finance in European languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2758–2775, Albuquerque, New Mexico. Association for Computational Linguistics.

David Ong and Peerat Limkonchotiwat. 2023. SEA-LION (Southeast Asian languages in one network): A family of Southeast Asian language models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 245–245, Singapore. Association for Computational Linguistics.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Barbara Plank. 2025. Human-centered llms for inclusive language technology: The need to embrace variation holistically in nlp. *Annals of Computer Science and Information Systems*, 43:43–45.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24261–24285, Suzhou, China. Association for Computational Linguistics.

Sara Rajaee, Rochelle Choenni, Ekaterina Shutova, and Christof Monz. 2025. An empirical analysis of machine translation for expanding multilingual benchmarks. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1–30, Suzhou, China. Association for Computational Linguistics.

Sai Sathiesh Rajan, Ezekiel Soremekun, and Sudipta Chattopadhyay. 2024. Knowledge-based consistency testing of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10185–10196, Miami, Florida, USA. Association for Computational Linguistics.

Natascia Ralli and Norbert Andreatta. 2018. bistro: Ein tool für mehrsprachige rechtsterminologie. *transkom: Journal of Translation and Technical Communication Research*, 11(1):7–44.

Natascia Ralli and Dóra Mária Tamás. 2025. How to create and manage terminology resources: A practical guide from two termbases. *Journal of Digital Terminology and Lexicography*, 1(1):61–82.

Vikas Raunak, Siddharth Dalmia, Vivek Gupta, and Florian Metze. 2020. On long-tailed phenomena in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3088–3095, Online. Association for Computational Linguistics.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. SALTED: A framework for SAlient long-tail translation error detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5163–5179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Siyu Ren and Kenny Q. Zhu. 2021. Knowledge-Driven Distractor Generation for Cloze-Style Multiple Choice Questions. In *AAAI Conference on Artificial Intelligence*, pages 4339–4347.

Jonathan Hvithamar Rystrøm, Hannah Rose Kirk, and Scott Hale. 2025. Multilingual != multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pages 74–85, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ananya Sahu, Mehrnaz Mofakhami, Daniel D'Souza, Thomas Euyang, Julia Kreutzer, and Marzieh Fadaee. 2026. The culture funnel: You can't align what isn't in the data. *Preprint*, arXiv:2606.13808.

Danielle Saunders and Steve DeNeefe. 2024. Domain adapted machine translation: What does catastrophic forgetting forget and why? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12660–12671, Miami, Florida, USA. Association for Computational Linguistics.

Beatrice Savoldi, Alan Ramponi, Matteo Negri, and Luisa Bentivogli. 2025. Translation in the hands of many: Centering lay users in machine translation interactions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13876–13889, Suzhou, China. Association for Computational Linguistics.

Barbara Schuppler, Martine Adda-Decker, Catia Cucchiarini, and Rudolf Muhr. 2024. An introduction to pluricentric languages in speech science and technology. *Speech Communication*, 156:103007.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 terminology translation task: Terminology is useful especially for good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, pages 554–576, Suzhou, China. Association for Computational Linguistics.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel Dsouza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. 2025a. The leaderboard illusion. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025b. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Hugo Sousa, Satya Almasian, Ricardo Campos, and Alípio Jorge. 2025. Tradutor: Building a variety-specific translation model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25183–25191.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491, Toronto, Canada. Association for Computational Linguistics.

Evgeniia Tokarchuk, Maya K. Nachesa, Sergey Troshin, and Vlad Niculae. 2026. Representation collapse in machine translation through the lens of angular dispersion. *Preprint*, arXiv:2602.17287.

Eva Vanmassenhove. 2026. Losing our tail, again: (un)natural selection & multilingual LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11589–11600, San Diego, California, United States. Association for Computational Linguistics.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *Preprint*, arXiv:2404.18796.

Wenxuan Wang, Yuk-Kit Chan, Zixuan Ling, Juluan Shi, Youliang Yuan, Jen tse Huang, Yifei Zhang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2026a. Identifying the achilles' heel: An iterative method for dynamically uncovering factual errors in large language models. *Preprint*, arXiv:2401.00761.

Yujun Wang, Ehud Reiter, Shimei Pan, Steffen Eger, and Wei Zhao. 2026b. Beyond accuracy: Community perspectives on machine translation. *Preprint*, arXiv:2606.09655.

Jacek Wiland, Max Ploner, and Alan Akbik. 2024. BEAR: A unified framework for evaluating relational knowledge in causal and masked language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2393–2411, Mexico City, Mexico. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *Preprint*, arXiv:2306.13063.

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.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. *Preprint*, arXiv:2407.19672.

A Prompt

```
You are a professional legal translator specializing in Italian-to-
   German translation for the legal system of {country}.

Your task is to translate a legal sentence from Italian into the
   standard variety of German used in {country}, without dialectal
   expressions. The translation must be fully aligned with the legal
    terminology, conventions, and drafting style of that
   jurisdiction. It must preserve the exact legal meaning and
   reflect how such concepts are expressed within the legal system
   of {country}, so that the result can function as an authoritative
    legal text.

There are terminological constraints you must adhere to:
{term_it} can be translated with only one of these terms:
{term_de_list}.

You must output only the translated text, without any explanation,
   enclosed in angle brackets "<>".

This is the text to be translated into {country} German:

IT: <{source_sentence}>
de-{country_code}:
```

B Effect of synthetic sources

B.1 Term accuracy

Table 2: Term accuracy for synthetic and human source sentences in the VAR subset.

Model	Synthetic Source	Human Source
Cohere CAT+	54.8	53.8
Command A+	54.5	56.4
CUNI-EdUKate	23.6	23.9
CUNI-MH-v3	28.3	28.3
DeepSeek V4 Pro	60.5	58.6
Dubformer	58.6	58.9
Gemini 3.1 Pro	72.6	72.6
Gemma 4 - 4B	53.2	48.4
Gemma 4 - 31B	59.6	56.4
GLM 4 - 9B	43.9	46.2
GoogleTranslate	33.4	34.4
GPT 5.5	69.7	67.8
GPT OSS 20B	55.7	54.1
GPT OSS 120B	55.7	56.4
Ministral 3 - 14B	53.2	53.5
Mistral Medium 3.5	55.7	59.2
Qingqiu-MT-9B	49.0	46.8
Qwen 3.5 - 9b	52.2	48.4
Qwen 3.6 - 27b	53.5	53.5
SalamandraTA	51.6	46.8
Tiny Aya Global	35.0	34.7
Tower 9B	45.9	46.8
TRIVE	70.7	68.5
UvA-MT	60.5	64.0

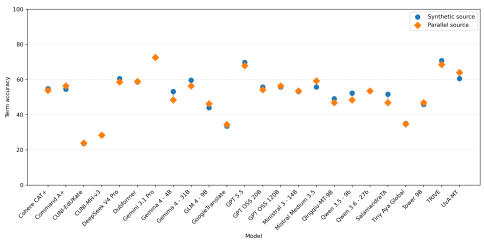


Figure 2: Plot of Term accuracy for synthetic and human source sentences in the VAR subset.

B.2 BLEU scores

Table 3: BLEU score for synthetic and human source sentences in the VAR subset.

Model	Synthetic Source	Human Source
Cohere CAT+	35.8	27.2
Command A+	30.8	25.2
CUNI-EdUKate	25.8	18.6
CUNI-MH-v3	27.6	20.5
DeepSeek V4 Pro	45.7	33.7
Dubformer	43.4	31.1
Gemini 3.1 Pro	50.6	36.5
Gemma 4 - 31B	38.2	28.9
Gemma 4 - 4B	31.2	23.1
GLM 4 - 9B	24.8	18.1
GoogleTranslate	26.9	20.4
GPT 5.5	49.1	36.0
GPT OSS 120B	31.1	23.6
GPT OSS 20B	24.7	17.7
Ministral 3 - 14B	27.3	20.0
Mistral Medium 3.5	38.1	28.9
Qingqiu-MT-9B	33.8	24.7
Qwen 3.5 - 9b	27.1	17.6
Qwen 3.6 - 27b	35.5	26.5
SalamandraTA	30.9	23.7
Tiny Aya Global	24.4	17.2
Tower 9B	32.6	24.2
TRIVE	50.2	35.0
UvA-MT	41.2	30.4

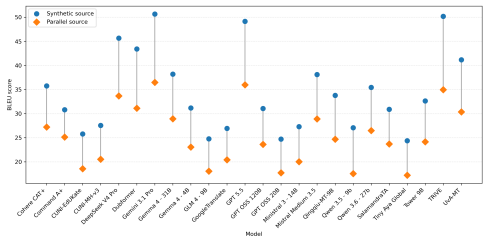


Figure 3: Plot of BLEU score for synthetic and human source sentences in the VAR subset.

C Main results

C.1 Term accuracy

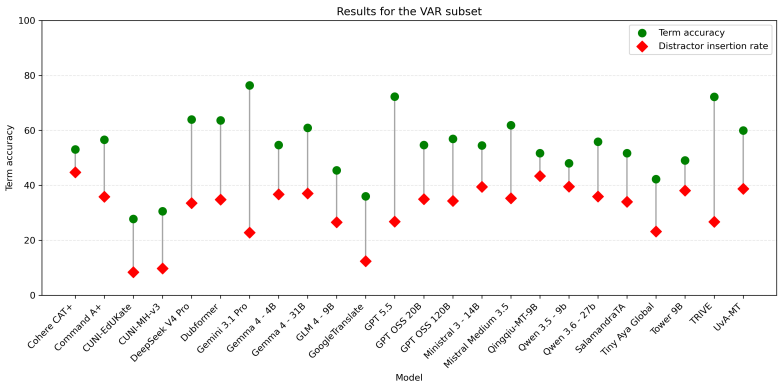


Figure 4: Plot of Term Accuracy for the VAR subset

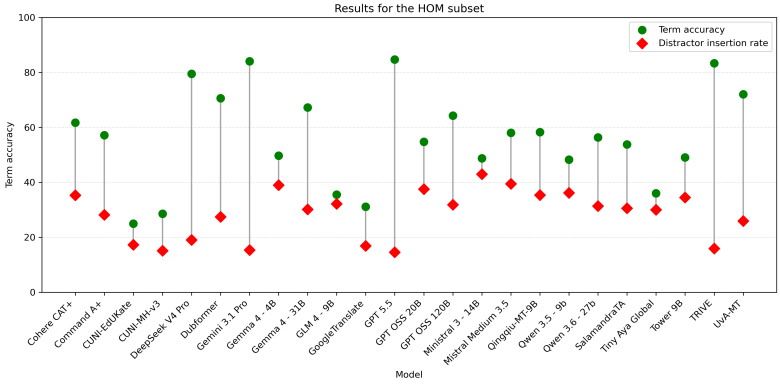


Figure 5: Plot of Term Accuracy for the HOM subset

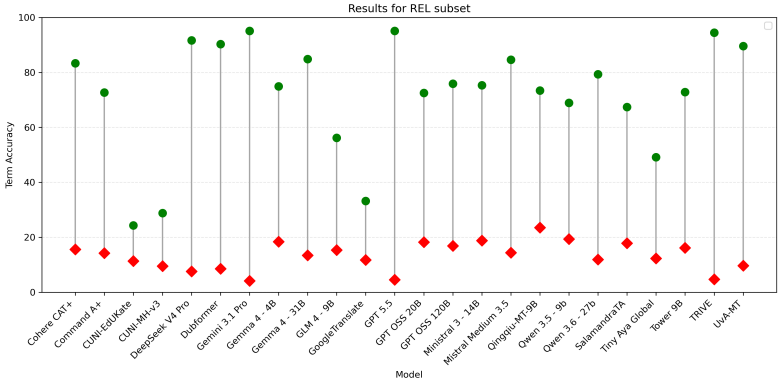


Figure 6: Plot of Term Accuracy for the REL subset

C.2 BLEU scores

Table 4: BLEU scores for all subsets. Correlation with Term accuracy indicates the Pearson's r between subset-level BLEU score and Term accuracy for each language variety.

Model	BLEU SCORE								
	VAR	HOM				REL			
Cohere CAT+	35.2	36.7	39.8	34.9	37.4	39.1	36.7	40.2	41.0
Command A+	32.7	36.6	31.4	31.9	33.8	38.8	30.6	36.5	34.5
CUNI-EdUKate	26.6	27.9	26.0	21.7	25.4	30.8	27.2	28.3	29.6
CUNI-MH-v3	29.0	29.6	29.8	25.7	28.4	31.7	30.2	29.0	30.5
DeepSeek V4 Pro	44.3	45.3	52.9	44.0	49.1	47.0	44.9	48.5	47.6
Dubformer	42.6	43.2	48.9	41.0	45.8	46.2	44.0	46.1	46.1
Gemini 3.1 Pro	**48.4**	50.2	**63.4**	50.1	60.1	**53.2**	**51.8**	**54.2**	54.1
Gemma 4 - 31B	38.0	39.8	40.5	35.6	38.3	42.8	39.4	41.7	43.6
Gemma 4 - 4B	31.4	32.0	30.7	27.0	28.5	35.2	30.6	32.1	34.9
GLM 4 - 9B	24.5	25.4	25.4	21.1	23.2	28.2	23.9	25.4	28.3
GoogleTranslate	28.1	27.4	28.3	23.4	25.9	29.7	28.8	31.8	30.3
GPT 5.5	47.4	50.1	60.1	**50.6**	**60.7**	50.8	49.9	53.2	**57.2**
GPT OSS 120B	31.8	33.8	35.2	28.9	29.3	36.3	31.0	33.8	33.5
GPT OSS 20B	25.1	26.3	25.9	20.8	21.1	29.6	25.4	24.1	25.4
Ministral 3 - 14B	26.9	27.3	27.5	23.8	24.5	29.4	25.3	29.5	22.9
Mistral Medium 3.5	38.1	40.1	45.2	37.4	39.4	40.8	39.9	42.1	43.9
Qingqiu-MT-9B	33.0	35.2	34.7	29.1	30.6	37.3	31.2	34.6	32.4
Qwen 3.5 - 9b	27.1	28.2	29.4	22.6	25.4	28.9	27.8	26.9	26.3
Qwen 3.6 - 27b	34.1	35.0	36.5	30.9	32.5	37.2	33.3	35.8	33.9
SalamandraTA	30.8	31.8	35.4	27.9	30.3	34.0	32.1	35.0	33.0
Tiny Aya Global	21.8	14.3	25.7	20.8	22.4	29.5	26.0	25.6	27.6
Tower 9B	33.3	34.0	36.4	30.4	32.6	36.5	34.4	37.1	38.6
TRIVE	46.8	**50.4**	57.5	47.0	58.3	51.2	49.9	52.7	52.6
UvA-MT	40.6	42.6	49.8	43.9	49.5	44.5	42.8	47.7	49.4
Correlation with Term accuracy	0.79	0.79	0.82	0.82	0.85	0.73	0.70	0.71	0.63

D Distractor analysis

Table 5: Variety provenance of inserted distractor terms when the South Tyrolean target term is not retrieved for the VAR subset. Since some terms are shared across varieties, attributions are not mutually exclusive and therefore do not sum to 100%. They indicate the share of matches out of the total featured distractor instances of the variety.

Model			
Cohere CAT+	45.0	47.1	36.8
Command A+	35.1	36.9	30.3
CUNI-EdUKate	7.4	8.4	7.7
CUNI-MH-v3	8.4	10.6	7.6
DeepSeek V4 Pro	36.1	33.9	26.3
Dubformer	35.0	36.0	28.6
Gemini 3.1 Pro	25.0	24.0	20.0
Gemma 4 - 4B	33.5	36.5	33.0
Gemma 4 - 31B	35.3	36.2	31.0
GLM 4 - 9B	23.4	24.6	23.2
GoogleTranslate	12.0	13.0	11.6
GPT 5.5	29.5	28.5	23.8
GPT OSS 20B	32.3	34.4	29.6
GPT OSS 120B	32.1	36.0	26.5
Ministral 3 - 14B	37.1	38.4	31.8
Mistral Medium 3.5	32.4	36.6	26.5
Qingqiu-MT-9B	41.4	43.4	37.0
Qwen 3.5 - 9b	37.1	40.6	33.0
Qwen 3.6 - 27b	34.3	36.9	32.2
SalamandraTA	30.3	35.0	27.2
Tiny Aya Global	20.8	24.2	20.7
Tower 9B	37.4	38.6	32.5
TRIVE	29.1	27.6	21.5
UvA-MT	39.2	41.1	32.1
Mean	30.4	32.0	26.3

Table 6: Type of inserted distractor terms when the target term is not retrieved for the REL subset.

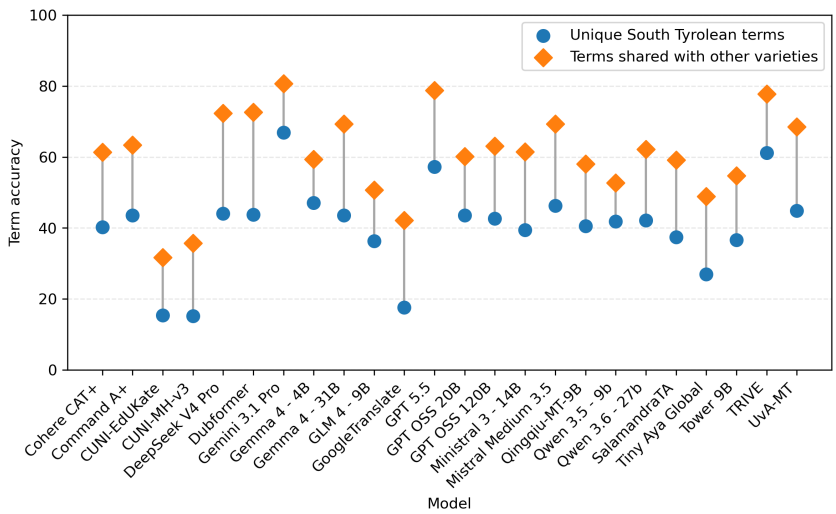
Model	**Hyponym**	**Cohyponym**	**Hypernym**
Cohere CAT+	23.0	8.2	9.5
Command A+	15.1	6.8	11.0
CUNI-EdUKate	3.3	2.7	14.4
CUNI-MH-v3	2.6	2.7	11.5
DeepSeek V4 Pro	12.5	2.7	4.9
Dubformer	13.2	4.6	4.6
Gemini 3.1 Pro	5.3	2.7	2.9
Gemma 4 - 4B	17.1	10.0	14.7
Gemma 4 - 31B	15.8	7.8	9.5
GLM 4 - 9B	8.6	7.8	15.0
GoogleTranslate	3.9	5.5	13.0
GPT 5.5	7.9	1.8	2.6
GPT OSS 20B	19.7	10.0	13.5
GPT OSS 120B	18.4	6.4	13.5
Ministral 3 - 14B	12.5	11.0	16.7
Mistral Medium 3.5	17.1	8.7	8.9
Qingqiu-MT-9B	25.0	14.2	17.0
Qwen 3.5 - 9b	18.4	11.9	14.1
Qwen 3.6 - 27b	13.2	4.6	9.5
SalamandraTA	21.1	14.6	10.7
Tiny Aya Global	4.6	6.8	12.7
Tower 9B	11.2	10.0	13.8
TRIVE	5.9	2.7	3.5
UvA-MT	11.8	5.0	6.6
Mean	12.8	7.1	10.6

E Analysis of unique South Tyrolean terms

Table 7: Term accuracy on South Tyrolean terms that are unique to South Tyrolean German versus terms also attested in other German varieties.

Model	Unique South Tyrolean Terms	Terms Shared with Other Varieties
Cohere CAT+	40.2	61.4
Command A+	43.5	63.4
CUNI-EdUKate	15.4	31.6
CUNI-MH-v3	15.2	35.7
DeepSeek V4 Pro	44.1	72.4
Dubformer	43.8	72.7
Gemini 3.1 Pro	**66.9**	**80.7**
Gemma 4 - 4B	47.1	59.4
Gemma 4 - 31B	43.5	69.3
GLM 4 - 9B	36.4	50.7
GoogleTranslate	17.6	42.1
GPT 5.5	57.3	78.8
GPT OSS 20B	43.5	60.2
GPT OSS 120B	42.7	63.0
Ministral 3 - 14B	39.4	61.5
Mistral Medium 3.5	46.3	69.4
Qingqiu-MT-9B	40.5	58.0
Qwen 3.5 - 9b	41.9	52.7
Qwen 3.6 - 27b	42.1	62.2
SalamandraTA	37.5	59.1
Tiny Aya Global	27.0	48.9
Tower 9B	36.6	54.7
TRIVE	61.2	77.8
UvA-MT	44.9	68.5
Mean	40.6	60.6

Figure 7: Comparison between unique South Tyrolean terms and South Tyrolean terms attested in other varieties



F Corpus frequency correlation

Table 8: Point-biserial correlation between expected term frequency in a legal corpus and term accuracy in the VAR subset.

Model	r
Cohere CAT+	0.061
Command A+	0.065
CUNI-EdUKate	0.159
CUNI-MH-v3	0.141
DeepSeek V4 Pro	0.082
Dubformer	0.067
Gemini 3.1 Pro	0.061
Gemma 4 - 4B	0.040
Gemma 4 - 31B	0.069
GLM 4 - 9B	0.070
GoogleTranslate	0.117
GPT 5.5	0.078
GPT OSS 20B	0.089
GPT OSS 120B	0.076
Ministral 3 - 14B	0.071
Mistral Medium 3.5	0.069
Qingqiu-MT-9B	0.060
Qwen 3.5 - 9b	0.070
Qwen 3.6 - 27b	0.062
SalamandraTA	0.054
Tiny Aya Global	0.106
Tower 9B	0.071
TRIVE	0.065
UvA-MT	0.068

G Legal subdomain

Table 9: Term accuracy by legal subdomain for the VAR subset

Model	Civil Law	Occ. Health & Safety Law	Criminal Law	Criminal Procedure Law	Administrative Law	Road Traffic Law	Civil Procedure Law	Company Law	Commercial Law	Law of Obligations	Family Law	Employment Law	Law of Obl. – Special	Constitutional Law	Public Law	Law of Succession	juris	Health Law	Insurance Law	Other
Cohere CAT+	57.8	50.8	65.5	49.6	60.8	54.5	48.2	52.2	51.7	59.1	57.9	48.0	65.5	63.8	60.8	55.6	64.7	53.1	42.9	54.6
Command A+	58.3	60.3	65.5	58.0	56.9	57.1	56.4	57.6	56.2	58.0	55.3	44.0	67.2	69.0	62.8	51.1	64.7	59.4	46.4	62.3
CUNI-EdUKate	28.6	30.2	25.9	20.6	34.6	22.3	16.4	28.3	25.8	30.7	30.3	29.3	31.0	41.4	41.2	15.6	35.3	43.8	28.6	20.8
CUNI-MH-v3	35.7	32.8	29.9	20.6	30.8	25.9	12.7	37.0	29.2	27.3	36.8	30.7	39.7	44.8	47.1	28.9	35.3	46.9	28.6	35.1
DeepSeek V4 Pro	69.4	56.6	70.7	60.3	73.1	56.3	60.0	65.2	68.5	77.3	57.9	62.7	77.6	65.5	70.6	73.3	73.5	56.3	50.0	67.5
Dubformer	69.9	64.0	73.0	61.1	72.3	58.0	62.7	62.0	62.9	70.5	61.8	62.7	72.4	69.0	64.7	62.2	73.5	53.1	67.9	67.5
Gemini 3.1 Pro	73.9	74.1	82.8	67.2	80.8	75.0	73.6	80.4	73.0	83.0	80.3	80.0	86.2	84.5	78.4	82.2	82.4	75.0	75.0	76.6
Gemma 4 - 4B	53.3	57.7	63.8	51.2	58.5	50.9	56.4	44.6	52.8	62.5	55.3	48.0	60.3	65.5	68.6	60.0	61.8	59.4	53.6	53.3
Gemma 4 - 31B	65.8	61.4	70.7	56.5	66.2	52.7	56.4	65.2	64.0	69.3	63.2	60.0	72.4	69.0	62.8	66.7	73.5	50.0	57.1	64.9
GLM 4 - 9B	55.3	45.5	48.9	45.0	56.2	42.9	38.2	50.0	40.5	53.4	48.7	40.0	51.7	50.0	58.8	40.0	38.2	40.6	32.1	44.2
GoogleTranslate	38.2	45.0	33.9	25.2	36.2	33.0	27.3	30.4	33.7	33.0	50.0	37.3	37.9	50.0	58.8	40.0	44.1	40.6	32.1	36.4
GPT 5.5	75.9	69.8	78.2	67.9	80.0	64.3	68.2	71.7	67.4	81.8	68.4	76.0	82.8	81.0	80.4	77.8	76.5	65.6	64.3	75.3
GPT OSS 20B	56.8	47.6	58.1	49.6	58.5	54.5	54.5	58.7	50.6	60.2	59.2	53.3	62.1	77.6	66.7	66.7	61.8	56.2	35.7	53.3
GPT OSS 120B	62.8	48.2	65.5	51.9	57.7	56.2	57.3	58.7	55.1	67.0	50.0	50.7	58.6	72.4	72.5	62.2	64.7	53.1	53.6	64.9
Ministral 3 - 14B	56.8	52.4	64.4	46.6	54.6	50.9	51.8	67.4	55.1	72.7	48.7	60.0	62.1	70.7	70.6	48.9	61.8	53.1	42.9	46.8
Mistral Medium 3.5	60.8	60.9	69.5	58.0	65.4	61.6	65.5	65.2	62.9	75.0	61.8	58.7	69.0	65.5	64.7	66.7	67.6	59.4	57.1	66.2
Qingqiu-MT-9B	51.8	58.2	54.6	52.7	60.8	48.2	42.7	44.6	44.9	59.1	50.0	56.0	56.9	62.1	64.7	62.2	67.6	53.1	39.3	48.0
Qwen 3.5 - 9b	51.3	47.1	54.0	50.4	57.7	42.0	43.6	45.7	41.6	61.4	50.0	52.0	50.0	56.9	56.9	57.8	47.1	34.4	28.6	49.4
Qwen 3.6 - 27b	59.3	54.0	65.5	51.9	56.2	48.2	56.4	56.5	59.6	65.9	51.3	53.3	63.8	69.0	70.6	62.2	61.8	56.2	35.7	52.0
SalamandraTA	56.3	55.6	56.9	51.2	56.2	47.3	50.9	53.3	50.6	68.2	46.1	48.0	62.1	53.5	54.9	48.9	58.8	59.4	46.4	46.8
Tiny Aya Global	46.7	45.5	47.7	30.5	50.0	40.2	32.7	40.2	40.5	42.1	36.8	46.7	53.5	56.9	64.7	33.3	44.1	65.6	35.7	40.3
Tower 9B	52.8	46.0	60.3	43.5	58.5	44.6	45.5	53.3	46.1	55.7	44.7	41.3	53.5	60.3	58.8	51.1	44.1	53.1	39.3	45.5
TRIVE	77.9	68.8	81.6	67.2	80.0	65.2	71.8	73.9	71.9	81.8	67.1	70.7	86.2	82.8	82.4	75.6	67.6	65.6	67.9	70.1
UvA-MT	64.8	54.0	67.2	54.2	72.3	62.5	58.2	64.1	59.6	70.5	57.9	56.0	63.8	70.7	70.6	66.7	73.5	59.4	67.9	66.2
Mean	57.5	53.6	60.6	49.6	59.7	50.6	50.3	55.3	52.7	61.9	53.7	52.7	61.9	64.7	64.7	56.5	60.2	54.7	47.0	54.5

Table 10: Term accuracy by legal subdomain for the HOM subset

Model	Civil Law	Administrative Law	Occ. Health & Safety Law	Civil Procedure Law	Criminal Law	juris	Criminal Procedure Law	Commercial Law	Law of Obligations	Company Law	Employment Law	Law of Obl. – Special	Public Law	Family Law	Environmental Law	Road Traffic Law	Constitutional Law	Other
Cohere CAT+	65.8	56.7	56.7	68.4	77.9	73.7	74.3	65.5	61.5	56.0	67.4	70.2	59.1	62.5	65.8	69.7	73.1	68.9
Command A+	61.5	50.8	55.0	57.9	54.5	77.6	55.7	56.4	55.8	64.0	65.3	61.7	54.5	55.0	60.5	57.6	61.5	49.2
CUNI-EdUKate	27.2	16.6	26.7	17.9	32.5	30.3	15.7	21.8	13.5	26.0	28.6	29.8	36.4	37.5	36.8	24.2	46.2	19.7
CUNI-MH-v3	33.9	19.8	27.5	20.0	35.1	42.1	21.4	21.8	19.2	32.0	32.7	29.8	31.8	37.5	36.8	36.4	34.6	26.2
DeepSeek V4 Pro	85.2	69.5	73.3	76.8	88.3	88.2	72.9	85.5	86.5	82.0	81.6	91.5	86.4	70.0	86.8	84.9	88.5	80.3
Dubformer	73.9	58.8	65.8	76.8	74.0	80.3	71.4	83.6	75.0	76.0	77.6	74.5	61.4	70.0	65.8	87.9	69.2	73.8
Gemini 3.1 Pro	88.3	73.3	75.8	83.2	90.9	88.2	82.9	92.7	88.5	88.0	87.8	93.6	77.3	95.0	92.1	97.0	76.9	83.6
Gemma 4 - 4B	52.9	45.5	45.8	55.8	57.1	63.2	54.3	52.7	44.2	50.0	44.9	51.1	50.0	52.5	60.5	45.5	46.2	49.2
Gemma 4 - 31B	71.6	59.9	60.0	67.4	63.6	71.1	70.0	78.2	75.0	72.0	69.4	66.0	72.7	57.5	78.9	72.7	80.8	73.8
GLM 4 - 9B	41.6	31.6	30.0	37.9	40.3	43.4	34.3	36.4	30.8	36.0	40.8	40.4	43.2	40.0	34.2	27.3	42.3	32.8
GoogleTranslate	35.4	24.6	34.2	28.4	37.7	39.5	30.0	27.3	25.0	24.0	22.5	27.7	47.7	42.5	44.7	24.2	38.5	24.6
GPT 5.5	88.3	72.7	81.7	85.3	93.5	90.8	80.0	85.5	90.4	86.0	91.8	93.6	84.1	95.0	92.1	93.9	80.8	83.6
GPT OSS 20B	56.0	52.9	55.0	52.6	52.0	67.1	55.7	72.7	57.7	52.0	46.9	59.6	56.8	42.5	57.9	66.7	65.4	59.0
GPT OSS 120B	69.3	55.6	58.3	68.4	64.9	72.4	67.1	81.8	57.7	74.0	65.3	66.0	59.1	62.5	76.3	75.8	73.1	70.5
Ministral 3 - 14B	53.7	49.2	40.8	57.9	53.3	63.2	48.6	49.1	63.5	60.0	38.8	42.6	54.5	42.5	47.4	57.6	61.5	47.5
Mistral Medium 3.5	60.7	55.1	47.5	64.2	64.9	72.4	57.1	54.5	63.5	72.0	59.2	63.8	52.3	45.0	73.7	72.7	50.0	59.0
Qingqiu-MT-9B	60.3	48.1	59.2	65.3	67.5	69.7	52.9	58.2	69.2	60.0	59.2	57.5	47.7	72.5	68.4	48.5	46.2	62.3
Qwen 3.5 - 9b	51.4	37.4	44.2	53.7	61.0	67.1	50.0	43.6	46.2	58.0	55.1	46.8	40.9	55.0	65.8	45.5	38.5	50.8
Qwen 3.6 - 27b	58.4	50.3	54.2	64.2	52.0	64.5	60.0	63.6	67.3	60.0	61.2	61.7	52.3	52.5	65.8	42.4	53.9	63.9
SalamandraTA	57.6	50.3	50.8	59.0	62.3	57.9	62.9	63.6	50.0	70.0	40.8	63.8	59.1	62.5	60.5	42.4	76.9	59.0
Tiny Aya Global	41.3	30.5	34.2	37.9	48.1	43.4	38.6	36.4	21.2	42.0	44.9	44.7	38.6	45.0	47.4	30.3	30.8	31.2
Tower 9B	53.3	44.4	43.3	56.8	53.3	57.9	62.9	50.9	50.0	50.0	46.9	48.9	52.3	57.5	63.2	30.3	50.0	47.5
TRIVE	86.8	71.7	76.7	83.2	88.3	88.2	80.0	90.9	88.5	88.0	89.8	93.6	84.1	92.5	89.5	93.9	80.8	83.6
UvA-MT	77.0	61.5	63.3	77.9	80.5	76.3	71.4	83.6	80.8	76.0	69.4	78.7	75.0	70.0	76.3	78.8	76.9	77.1
Mean	60.5	49.4	52.5	59.0	62.2	66.2	57.1	60.7	57.5	60.6	57.8	60.7	57.4	59.0	64.5	58.6	60.1	57.4

H Distractor number analysis

Table 11: Term accuracy for the VAR subset by number of distractors provided to the model.

Model	1	2	3	4	5	6	7	8+
Cohere CAT+	75.2	58.8	41.8	42.9	39.7	30.6	12.5	4.4
Command A+	75.7	58.8	48.1	45.9	43.8	40.8	16.7	13.0
CUNI-EdUKate	32.7	27.4	26.1	24.8	21.9	28.6	16.7	4.4
CUNI-MH-v3	36.5	31.7	25.8	24.1	32.9	32.7	25.0	8.7
DeepSeek V4 Pro	81.6	66.1	53.7	51.9	57.5	51.0	29.2	43.5
Dubformer	82.3	64.7	55.8	51.1	53.4	53.1	45.8	30.4
Gemini 3.1 Pro	**90.0**	**77.2**	**66.2**	**68.4**	**75.3**	65.3	**62.5**	47.8
Gemma 4 - 4B	69.0	57.7	49.0	42.9	38.4	46.9	41.7	13.0
Gemma 4 - 31B	76.9	64.5	53.7	51.1	48.0	49.0	37.5	39.1
GLM 4 - 9B	58.4	46.2	40.4	44.4	41.1	30.6	20.8	13.0
GoogleTranslate	43.4	35.8	33.5	27.8	34.3	38.8	29.2	0.0
GPT 5.5	89.0	73.5	60.8	60.2	68.5	**67.3**	41.7	**52.2**
GPT OSS 20B	69.4	61.5	46.9	36.8	32.9	42.9	37.5	17.4
GPT OSS 120B	73.3	59.7	49.9	44.4	45.2	28.6	29.2	13.0
Ministral 3 - 14B	71.1	57.9	49.3	36.8	37.0	36.7	29.2	26.1
Mistral Medium 3.5	76.6	67.0	58.8	48.1	50.7	34.7	33.3	17.4
Qingqiu-MT-9B	71.1	52.9	43.3	44.4	41.1	30.6	8.3	8.7
Qwen 3.5 - 9b	64.7	52.5	40.7	32.3	39.7	28.6	16.7	8.7
Qwen 3.6 - 27b	73.8	58.4	46.9	43.6	37.0	34.7	29.2	30.4
SalamandraTA	67.0	55.2	43.9	45.1	43.8	36.7	33.3	21.7
Tiny Aya Global	54.0	42.5	37.1	39.1	38.4	38.8	20.8	8.7
Tower 9B	69.2	49.6	35.3	39.9	35.6	34.7	20.8	13.0
TRIVE	87.3	73.8	62.9	63.2	68.5	63.3	54.2	43.5
UvA-MT	79.7	64.9	49.0	51.9	50.7	42.9	29.2	17.4
Mean	69.5	56.6	46.6	44.2	44.8	41.2	30.0	20.7

Table 12: Term accuracy for the HOM subset by number of distractors provided to the model.

Model	1	2	3	4+
Cohere CAT+	69.5	57.4	57.4	46.3
Command A+	62.2	56.1	42.6	36.6
CUNI-EdUKate	24.5	25.6	30.6	31.7
CUNI-MH-v3	28.4	30.2	33.3	24.4
DeepSeek V4 Pro	82.6	76.7	73.2	82.9
Dubformer	76.7	67.5	58.3	53.7
Gemini 3.1 Pro	86.2	82.6	**78.7**	82.9
Gemma 4 - 4B	56.6	44.3	40.7	34.2
Gemma 4 - 31B	71.8	63.3	61.1	56.1
GLM 4 - 9B	39.5	33.4	32.4	22.0
GoogleTranslate	33.2	28.5	29.6	31.7
GPT 5.5	**87.2**	83.3	77.8	**87.8**
GPT OSS 20B	61.2	47.9	43.5	41.5
GPT OSS 120B	72.0	57.7	45.4	58.5
Ministral 3 - 14B	56.7	42.3	35.2	31.7
Mistral Medium 3.5	64.5	53.8	44.4	48.8
Qingqiu-MT-9B	63.8	55.7	50.9	36.6
Qwen 3.5 - 9b	54.9	44.9	37.0	31.7
Qwen 3.6 - 27b	61.5	53.4	50.0	34.2
SalamandraTA	60.3	50.5	50.0	46.3
Tiny Aya Global	42.9	29.5	35.2	12.2
Tower 9B	53.3	48.9	43.5	34.2
TRIVE	86.3	**83.6**	74.1	70.7
UvA-MT	76.1	70.5	59.3	73.2
Mean	61.3	53.7	49.3	46.2

Table 13: Term accuracy for the REL subset by number of distractors provided to the model.

Model	1	2	3	4	5+
Cohere CAT+	88.6	79.1	86.8	78.0	79.1
Command A+	72.2	72.4	82.9	80.0	61.5
CUNI-EdUKate	23.3	27.6	26.3	18.0	23.1
CUNI-MH-v3	29.6	29.1	32.9	26.0	25.3
DeepSeek V4 Pro	92.1	92.5	94.7	92.0	86.8
Dubformer	90.9	91.0	89.5	96.0	85.7
Gemini 3.1 Pro	**94.9**	96.3	**96.1**	92.0	**94.5**
Gemma 4 - 4B	76.7	70.9	84.2	72.0	71.4
Gemma 4 - 31B	86.4	84.3	89.5	82.0	80.2
GLM 4 - 9B	64.8	57.5	54.0	44.0	46.2
GoogleTranslate	35.8	32.8	32.9	32.0	29.7
GPT 5.5	94.3	**97.0**	93.4	**98.0**	93.4
GPT OSS 20B	77.8	75.4	71.1	66.0	62.6
GPT OSS 120B	78.4	75.4	79.0	74.0	70.3
Ministral 3 - 14B	81.8	79.9	64.5	76.0	64.8
Mistral Medium 3.5	88.6	85.1	82.9	88.0	75.8
Qingqiu-MT-9B	77.8	70.9	69.7	76.0	70.3
Qwen 3.5 - 9b	72.2	69.4	71.1	62.0	63.7
Qwen 3.6 - 27b	84.1	76.9	85.5	84.0	65.9
SalamandraTA	76.1	64.2	65.8	64.0	58.2
Tiny Aya Global	50.6	53.7	56.6	42.0	37.4
Tower 9B	76.7	74.6	69.7	72.0	65.9
TRIVE	94.3	95.5	93.4	94.0	**94.5**
UvA-MT	90.9	89.6	89.5	88.0	87.9
Mean	75.0	72.5	73.4	70.7	66.4