Evaluating LLMs on Explicit Knowledge of Italian Legal Concepts

Anonymous ACL submission

Abstract

Despite their growing usage in downstream applications, LLMs remain under-evaluated in specialized non-English domains such as healthcare, finance, and law. This creates gaps not only in linguistic assessment, but also in the evaluation of language-specific technical knowledge, calling for high-quality resources tailored to narrow domains. We address the evaluation gap of LLMs' **Explicit Knowledge of Italian Legal Concepts**, introducing a set of evaluation resources derived from institutional and expert-crafted sources, and specifically aimed at assessing Italian Legal knowledge across various subdomains. We organize these resources around three tasks targeting both *generative* and *discriminative* LLMs abilities and evaluate diverse frontier and small open models, both Italian-specific and general-purpose, across multiple metrics and settings.

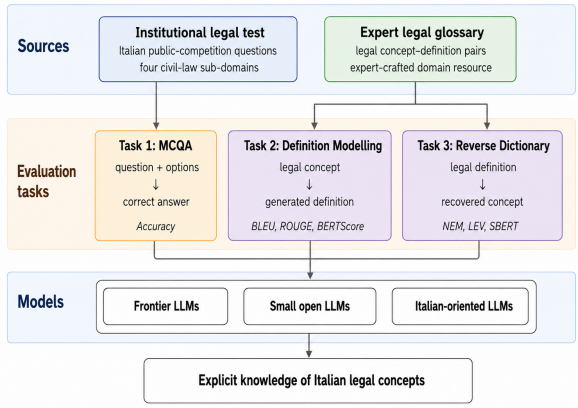

Figure 1: Overview of the evaluation framework for explicit knowledge of Italian legal concepts. Two legal-domain sources support three complementary tasks, evaluated across different LLM families.

1 Introduction

Large Language Models (LLMs) are currently being developed and adopted in a growing variety of languages, tasks and domains. However, careful evaluation of their performance in specific languages and domains is still relatively incomplete due to lack of dedicated evaluation resources. This causes **evaluation gaps** in highly specialized domains, meaning that LLMs are deployed in many more scenarios than those they are specifically evaluated for in standard benchmarks. These gaps are even sharper when technical knowledge, such as financial, healthcare and legal has to be evaluated in languages other than English, which is still the primary idiom for LLMs development. In this work, we address the problem of evaluating **LLMs' knowledge of Italian Legal concepts**, sensibly underrepresented in current benchmarks, albeit strategically crucial for the reference community.

Evaluation resources for iterative testing are central to Machine Learning practice, with MCQA commonly used as an efficient and easy-to-parse format. For Italian, recent benchmarks include *ItaEval* (Attanasio et al., 2024), covering multiple domains and capabilities; *ItaBench* (Moroni et al., 2024), based on translated, adapted, and manually crafted data; and *Italic* (Seveso et al., 2025), targeting Italian cultural and linguistic evaluation. Yet, none of these resources, to the best of our knowledge, specifically targets the Italian legal domain. We fill this gap with a benchmark derived from a public test and covering four subfields of Italian Civil Law. However, MCQA alone may misrepresent LLM capabilities in generative settings, where models express lexical and conceptual knowledge more freely. Prior work has evaluated such knowledge in Transformer models (Vaswani et al., 2017) through dictionary-definition tasks: *Definition Modelling* (DM) (Noraset et al., 2017), which generates a gloss from a term, and *Reverse Dictionary* (RD) (Mane et al., 2022), which recovers the term from a definition. These tasks have been explored for English (Chen and Zhao, 2022; Bevilacqua et al., 2020), Italian (Ciaccio

et al., 2025), and technical domains such as finance (Jhirad et al., 2023), but not yet for Italian legal concepts. We address the evaluation gap in the **explicit knowledge of Italian legal concepts** and contribute by: i) introducing a set of domain-specific, high quality, evaluation resources grounded in institutional and expert-built sources; ii) organizing these resources around different tasks comprising both discriminative and generative evaluations, and iii) evaluating a range of models including top tier frontier LLMs and small open-source ones, both Italian-optimized and not.

Our experimental setup is designed to assess explicit knowledge of Italian legal concepts along complementary dimensions. First, we test whether models can recognize the correct legal answer in a discriminative MCQA setting, while also measuring how performance varies across different civil-law sub-domains. Second, we move to a generative setting and evaluate whether models can produce definitions of legal concepts that are not only semantically adequate, but also close to expert-written formulations. Finally, we test the inverse ability, namely recovering the correct legal concept from its definition, thus probing lexical-conceptual knowledge independently from multiple-choice cues. Overall, the three tasks offer a multifaceted evaluation of Italian legal-domain knowledge.[1]

2 Experimental Setup

Task 1 - MCQA. We adapted our evaluation dataset from a preliminary assessment for an institutional Italian Public Competition available on the web. [2] This source includes four Italian civil law sub-fields (*Private Law, Administrative Law, Civil Procedure and Administrative Procedure*), from each of which we derived ≈ 800 questions, for a total of $3,631$ items, as shown in Table 1.[3] We evaluated each sub-domain separately using four randomized shuffles of questions and answer options to improve robustness and detect potential choice-position bias. Performance was measured with accuracy.

Task 2 - DM. We extracted concept-definition pairs from an expert-written glossary available online.[4] We reduced the initial number of definitions (4k) to approximately 1,240 items (Table 1). We retained definitions with near-average length, discarding overly short or long instances, and we cleaned the texts by removing redundant examples and digressive material. Performance was evaluated with both n-gram metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), and a semantic metric like BERTScore (Zhang et al., 2019).[5]

Task	# Items	Evaluation Metrics
MCQA	3,631	Accuracy
Definition Modelling	1,240	BLEU, ROUGE, BERTScore
Reverse Dictionary	455	NEM, LEV, SBERT

Table 1: Tasks, Items and Metrics overview.

Task 3 - RD. In this task, LLMs have to generate the correct legal concept associated with the provided definition. We reused the Task 2 dataset, removing entries whose definitions contained the target concept to avoid label leakage, leaving 455 valid items (Table 1). Performance was assessed with increasingly abstract metrics: *Normalized Exact Match* (NEM), *Levenshtein similarity* (LEV), and *semantic similarity* (SBERT) (Reimers and Gurevych, 2019). All metrics were computed on normalized texts, after removing punctuation, extra spaces, and non-alphanumeric characters. We also devised a *discriminative* version of this task in the form of MCQA where LLMs have to choose the correct concept in a pool of answers comprising distractors. We derived distractors by sampling 3 terms among the top 20 having the most similar definitions with respect to the correct answer, ensuring morphological contiguity between the choices to avoid that LLMs would have relied on shallow heuristics to select the correct choice.[7]

Models. We evaluated six frontier models (Claude-sonnet-4.5, Gemini-3-flash, Gemini-2.5-flash, GPT-5.1-chat, GPT-4o and DeepSeek-v3-chat) and four small open-source models: two English-first models (Llama-3-8b-it and Mistral-7b-it), one Italian-tuned with Direct Preference Optimization (Rafailov et al., 2023) (Llamantino-3-8b-it), and one pretrained on Italian (Minerva-7b-it). All models were queried in a zero-shot setting with temperature set to 0.[8]

3 Results

Task 1. Figure 2 reports task's results by sub-domain (bars) and overall performance (star), with

[1] Code and data will be released upon acceptance.

[2] https://www.giustizia-amministrativa.it

[3] Detailed breakdown across subdomains is reported in Table 4 in Appendix A

[4] https://www.brocardi.it/dizionario/

[5] We used `paraphrase-multilingual-mpnet-base-v2`[6] (Reimers and Gurevych, 2019) and applied baseline rescaling.

[7] Details about datasets are reported in Appendix C.

[8] Models and Querying details are reported in Appendix B.

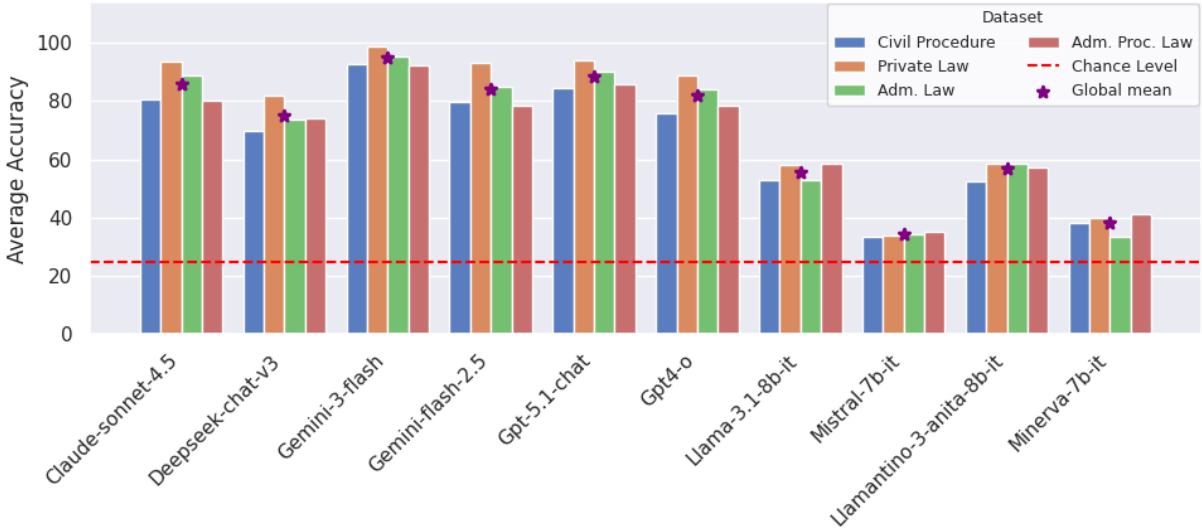

Figure 2: Average accuracy (across random shuffles) of the tested LLMs on each law sub-field for MCQA Task 1.

detailed scores and standard deviations in Appendix B. As expected, frontier models largely outperform smaller open models, including Italian-native and Italian-adapted ones. Gemini-3-flash achieves the best overall performance, with an average accuracy of 94.71 and the highest score on *Private Law* (98.66 $\pm$ 0.36). Gpt-5.1-chat, Claude-sonnet-4.5, and Gemini-flash-2.5 also perform strongly, with average accuracies of 88.78, 85.80, and 84.01, respectively, while Deepseek is the weakest frontier model (74.76). Smaller models remain above chance but lag substantially behind. Notably, Minerva, the only Italian-native model, is the second weakest model overall (38.21), outperforming only Mistral (34.20), whereas Llama (55.57) and Llamantino (56.75) perform similarly, suggesting only marginal gains from Italian adaptation in this setting. Across models, *Private Law* is consistently the easiest sub-domain, while *Civil Procedure* is the most challenging. Overall, these results show that frontier models perform strongly on MCQA legal knowledge, but also that domain difficulty persists across model families, highlighting the need for specialized evaluation.

Task 2 - Moving on to DM results, we can see the scores obtained by the LLMs on different text generation metrics in Table 2. The results show a clear gap between the semantic metric, *BERTScore*, and structural n-gram metrics, especially BLEU, which yields the lowest scores across models. This suggests that models often produce definitions that differ formally from the gold standard while still partially preserving their meaning, possibly through simpler or less expert-like formulations. Compared to Task 1, model differences are much less pronounced: n-gram metrics are particularly homogeneous, with ROUGE-2 and BLEU providing the weakest signal, while *BERTScore* captures slightly more variation. Gemini-3 remains the best overall model, reaching a *BERTScore* of 0.50, but the advantage of frontier models is far smaller than in MCQA. Among smaller models, Minerva shows a notable improvement over Task 1, ranking fifth with a *BERTScore* of 0.43, whereas Llamantino performs worse than its non-Italian-optimized counterpart, Llama-3-it.

Task 3. - Table 3 reports the results for the RD Task 3 across the different metrics. The results show that Claude-sonnet-4.5 and Gemini-3-flash perform best, both reaching an SBERT score of 0.82 and high NEM scores of 0.47 and 0.46, respectively. Gpt-5.1-chat follows closely, with competitive scores across metrics. Like in previous tasks, frontier models clearly outperform smaller open-weight models: Mistral-7B-it and Minerva-7B-it obtain particularly low exact-match scores (NEM $\leq$ 0.1). Across all models, SBERT is consistently higher than NEM and LEV, suggesting that models often produce semantically related terms even when they miss the exact legal concept. Among 7-8B models, Llama and Llamantino show a slight advantage in semantic similarity (SBERT 0.64–0.65).

A different picture is offered by the *discriminative* version of this task, as shown in Figure 3. Frontier models seem to perform at ceiling, all obtaining an accuracy above 95%, partially replicating the hierarchy of Task 1, with Gemini-3 on top (98%). Llama (87%) and Llamantino (%88) shows competitive scores above 85%. The worst model is again Mistral (42%) surpassed by Minerva (62%),

Model	BERT-S	R-1	R-2	R-L	BLEU
Claude-sonnet-4.5	0.49	0.27	0.08	0.19	0.04
Deepseek-chat-v3	0.41	0.25	0.06	0.17	0.04
Gemini-2.5-flash	0.47	0.27	0.08	0.20	0.05
Gemini-3-flash	0.50	0.27	0.08	0.19	0.05
Gpt-4o	0.44	0.26	0.07	0.19	0.04
Gpt-5.1-chat	0.48	0.25	0.07	0.18	0.04
Llama-3-8b-it	0.42	0.22	0.04	0.16	0.03
Mistral-7b-it	0.35	0.20	0.03	0.14	0.02
Llamantino-3-anita-8b-it	0.39	0.22	0.04	0.15	0.03
Minerva-7b-it	0.43	0.22	0.04	0.14	0.02

Table 2: Definition Modelling (DM) Task 2 results across several metrics (F1 averages).

Model	NEM	LEV	SBERT
Claude-sonnet-4.5	0.47	0.67	0.82
Deepseek-chat-v3	0.21	0.55	0.75
Gemini-2.5-flash	0.37	0.65	0.8
Gemini-3-flash	0.46	0.67	0.82
Gpt-4o	0.38	0.61	0.78
Gpt-5.1-chat	0.45	0.67	0.81
Llama-3.1-8B-it	0.16	0.41	0.64
Mistral-7B-it	0.07	0.33	0.6
LLamantino-3-anita-8B-it	0.15	0.39	0.65
Minerva-7B-it	0.1	0.33	0.61

Table 3: Reverse Dictionary (RD) Task 3 results across several metrics.

which is nonetheless substantially behind its open rivals. These results suggest that the discriminative MCQA version of Task 3 is inherently easier than both Task 1 and its generative counterpart.

4 Discussion

We performed three tasks designed to assess LLMs' explicit knowledge of Italian legal concepts across both *discriminative* and *generative* settings, using resources grounded in institutional and expert-crafted sources. The results show that frontier models can achieve strong performance, especially in MCQA, but also that there is substantial room for improvement, especially across sub-domains.

In **Task 1**, even the strongest models showed uneven performance across areas of law, suggesting that broad legal competence should not be assumed to transfer uniformly to more specialized sub-domains. This task also produced the clearest gap between frontier and smaller open models. However, this difference may reflect not only stronger knowledge encoding, but also better instruction following in a discriminative setting.

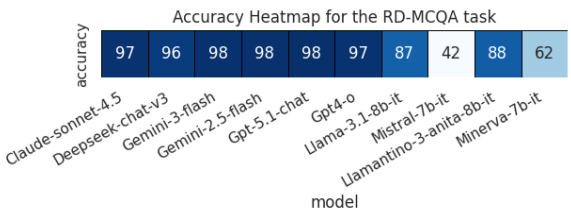

Figure 3: Accuracy heatmap of models on the RD-MCQA Task 3

By contrast, **Task 2**, focused on free-form definition generation, produced a more homogeneous picture. Differences among models were smaller across both n-gram and semantic metrics, while the generally low n-gram scores suggest that models often capture the core meaning of legal concepts without matching expert-level formulations. This highlights the importance of formal precision in legal-domain generation, where semantic similarity as measured by *BERTScore* may not be sufficient.

Task 3 further confirms this pattern: Models often retrieved semantically related concepts even when they failed to produce the exact legal term. However, the *discriminative* version of this Task shows that, when provided a pool of terms to choose the correct on from based on the definition, all models perform much better. In general, we found no clear advantage for Italian-optimized models across tasks, whether natively trained or adapted. Overall, our results indicate that the tested LLMs can handle Italian legal concepts to some extent, but their performance remains fragmented across tasks, sub-domains, model size, and model type. This confirms the need for dedicated evaluation resources in specialized non-English domains, particularly for smaller open models, which appear to suffer most from the current evaluation gap in this area.

5 Conclusions

In this work we introduced a set of evaluation resources, adapted from institutional and expert made sources to test LLMs explicit knowledge of Italian legal concepts and alleviate the evaluation gap affecting such a specialized and technical domain. We evaluated a group of LLMs on both *discriminative* and *generative* tasks and provide a nuanced picture of LLMs abilities through several metrics. Future work will expand such resources to other areas of the Italian law and the formalize additional tasks, addressing features going beyond knowledge that may be crucial in legal scenarios.

Limitations

This work has a number of limitations. First of all, while we varied across legal sub-domains, we did not cover *all* possible areas of Italian Law, as this is a much more challenging endeavor to be pursued incrementally in future work. Secondly, we tested several models but many more exists and a more comprehensive evaluation including other size scales would be beneficial. Thirdly, we tested the models in 0-shot and non reasoning setting, avoiding Chain-of-Thought prompting as well as any knowledge intensive query framework like RAG. Further tests may be led on the same dataset with different configurations. We leave all the limitations as open points for future reasearch.

References

Giuseppe Attanasio, Moreno La Quatra, Andrea Santilli, and Beatrice Savoldi. 2024. ItaEval: A CALAMITA challenge. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1064–1073, Pisa, Italy. CEUR Workshop Proceedings.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

Pinzhen Chen and Zheng Zhao. 2022. A unified model for reverse dictionary and definition modelling. In *AACL*.

Cristiano Ciaccio, Alessio Miaschi, and Felice Dell'Orletta. 2025. Evaluating lexical proficiency in neural language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1267–1286, Vienna, Austria. Association for Computational Linguistics.

James Jhirad, Edison Marrese-Taylor, and Yutaka Matsuo. 2023. Evaluating large language models' understanding of financial terminology via definition modeling. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 93–100, Nusa Dua, Bali. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Sunil B Mane, Harshal Navneet Patil, Kanhaiya Balaji Madaswar, and Pranav Nitin Sadavarte. 2022. Wordalchemy: a transformer-based reverse dictionary. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pages 1–5. IEEE.

Luca Moroni, Simone Conia, Federico Martelli, and Roberto Navigli. 2024. Towards a more comprehensive evaluation for Italian LLMs. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 584–599, Pisa, Italy. CEUR Workshop Proceedings.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Andrea Seveso, Daniele Potertì, Edoardo Federici, Mario Mezzanzanica, and Fabio Mercorio. 2025. ITALIC: An Italian culture-aware natural language benchmark. 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 1469–1478, Albuquerque, New Mexico. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Datasets Construction

A.1 Task 1

Dataset building for Task 1, MCQA, involved downloading source data from an institutional website, (Giustizia-Amministrativa.it). These data where stored in a tabular format in PDF files, which required careful conversion and parsing to properly extract questions and answers, which where shuffled and stored in JSON format afterwards. Detailed distribution of items across legal subdomains is reported in Table 4

Legal Domain	# items
Private Law	800
Civil Procedure	800
Adm. Law	825
Adm. Procedure	806
Total	3631

Table 4: Detailed reporting of number of items per legal domain for Task 1.

A.2 Task 2

For this task, we collected definitions of Italian legal concepts from an expert-built and managed source available in the web. To scope the source and collect definitions we applied the following main criteria:

- Availability, in terms of quantity of concept - definition pairs
- Easiness of extraction
- Domain- specificity
- Search Likelihood, in terms of probability for a user to encounter the source with a simple web search querying the meaning of a given concept

To assess the Search-Likelihood of web sources we started from a pool of 500 seed terms derived by the Italian Wiktionary having at least one sense marked as law-related (#diritto). We performed multiple web searches using a dedicated google search api[9] and recorded how many times a source came out for a given query in the top-k positions, with k = 10. We found the site Brocardi.it to be the most retrieved source across our queries. We chose that as our concepts and definitions source, as it also turned out to be a great one with respect to our selection criteria of i) availability, as it provides lots of concept-definition pairs, ii) easiness of extraction, as the data is formatted in a consistently structured way. Additionally, it satisfies our domain-relatedness criterion for data selection, as the site is built, managed and maintained by legal professionals as lawyers, making this data selection process highly similar to asking experts to generate the desired definitions, but keeping the process much more efficient and less costly. A further prove of the quality of the derived definitions is given by the fact that some of them are selected by expert lexicologists and included in the *Interactive Terminology for Europe* (IATE)[10] database, a primary terminological source built and used by lexicologists and translators working in technical domains across European Languages[11] within the European Union.

A.3 Task 3

The same dataset used in Task 2, Definition Modelling, is also used in Task 3, Reverse Dictionary, in both its *generative* and *discriminative* versions. The data were filtered to cut out definitions containing the target concept, thus leaving 455 entries. For the *discriminative* version of this task, formatted as an MCQA test, we automatically built output choices comprising the correct answer along with distractors To derive challenging distractors we used the same embedding model as in Task 2 to compute a similarity matrix among definitions. We then randomly sampled 3 items from the top 20 most similar ones. We also tried to ensure morphological contiguity between the correct answer and the distractors to avoid that LLMs would have relied on shallow heuristics to select the correct choice. We did so through a combination of heuristics and classic NLP parsing using the SpaCy Python library [12].

A.4 Data Provenance and Responsible Use

The benchmark is intended only for evaluating language models' explicit knowledge of Italian legal

[9] https://serper.dev/

[10] https://new.iate.europa.eu

[11] Consider the example entry: *negozio giuridico simulato*: https://iate.europa.eu/search/result/1772815409169/1.

[12] https://spacy.io/

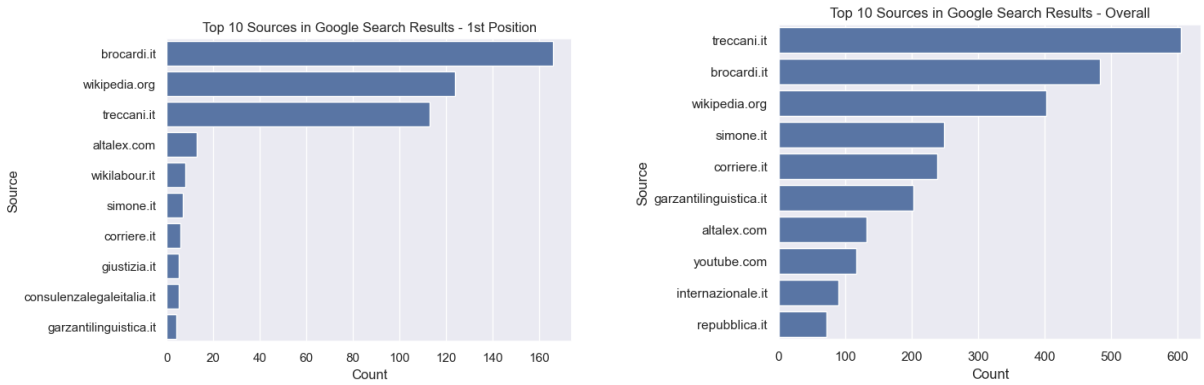

Figure 4: Search results of our seed concepts, both in first position and overall.

concepts. It should not be interpreted as a substitute for professional legal advice, nor as a system for supporting legal decisions. Model outputs are evaluated against reference answers and definitions, but these outputs may still be incomplete, imprecise, or legally unreliable in real-world settings.

The evaluation resources used in this work are derived from publicly accessible web sources. Task 1 is based on material made available through an institutional website, while Tasks 2 and 3 are based on an expert-curated online legal glossary. We used these sources exclusively to construct evaluation items for research purposes, and not to train or fine-tune any model.

During dataset construction, we retained only the information needed to instantiate the evaluation tasks. The data were converted into structured formats, cleaned, and filtered to remove redundant, excessively long, or unsuitable entries. For the reverse-dictionary task, we additionally removed cases in which the target concept appeared in the definition, in order to avoid label leakage.

The resulting resources contain legal questions, answer options, legal terms, and definitions. To the best of our knowledge, they do not contain private personal information, user-generated social-media content, or sensitive personal data. Since the original sources are public and concern legal concepts rather than individuals, the main ethical and legal risks concern source attribution, redistribution, and potential copyright restrictions rather than privacy.

For this reason, we will release only the derived experimental resources and code required to reproduce the evaluation, subject to the licenses and terms of use of the original sources. Where full redistribution is not appropriate, we will release scripts, metadata, and processing instructions that allow the resources to be reconstructed from the original public sources. We will also provide source attribution and document the preprocessing steps applied to each dataset.

B Evaluation Details

All open-source local models models have been queried on a single NVIDIA 48GB A100 GPU. As for the bigger and proprietary models, which are closed or wouldn't run on our hardware, we used the API provider *OpenRouter*[13], which offer the possibility to query several models through a unified, OpenAI compatible, API interface in a pay-per-use manner. API calls across experiment needed a budget of around 100$ including trials, errors and clean experiment executions. Experiments were built with the help of common coding assistants to write redundant and boilerplate code (I/O management, file reading and writing, comments etc.)

B.1 Querying Strategies for Task 1

While we used the same prompt for both proprietary/big models and open source ones, we employ two different strategies to access and query them. As already pinpointed in the previous subsection, local models run on a single NVIDIA GPU. Experiments for these models where implemented using the *HuggingFace Transformers* library (Wolf et al., 2019) and *Outlines*[14](Willard and Louf, 2023), a structured data generation library, allowing to restrict the possible search space of the models to target outputs. This has been done to make the comparison with bigger and most instruction-following proficient models, as well as to make easier the answer parsing in the post-processing phase. Closed models were queried through *OpenRouter*.

[13]https://openrouter.ai/

[14]https://dottxt-ai.github.io/outlines/latest/

B.2 Prompts

In this section we report the prompt for the three tasks.

Task 1 – Figure 5 shows the prompt for the first Task, a classic MCQA format made of basic instructions, question text and possible answers to choose from.

Task2 – We devised our prompt for Task 2 to be as neutral and simple as possible, while keeping a real-world usage flavor, mirroring a question that an actual user may ask. In this sense, our queries are inspired by the most widespread LLMs usages by common users, which is finding information and explaining concepts. We used several metrics to evaluate models' performances in this task and we briefly review them here.

Task 3 – We follow similar simplicity-based guidelines to compose the prompt for Task 3. The prompt is show in 7. The prompt used for the *discriminative* version of this task follows the same template of Task 1.

C Further analysis

C.1 Detailed scores for Task 1

Table 5 shows the complete accuracy scores obtained by each model on each law sub domain. Results are averaged across random shuffles and reported along with standard deviation values. Last column reports the global average for each model across sub domains, while last row indicates the global score of each dataset averaged across models. On the model side, it can be seen that Gemini-3 is the best overall model, while Mistral is the worst one. On the dataset side, we see how *Private Law* is the easieast sub-domain across models, while seems to be the hardest one.

C.2 Cross-Model analysis of Task 2

We compared the definitions generated by the models across each other along all the metrics used in Task 2. Two interesting things emerged from this additional analysis. First, models generated definitions are more similar among each others than to the gold standard. Second, this is particularly true for a set of models, namely the frontier ones. This result can be seen in the heatmap shown in Figure 8, and in particular in the n-gram based metrics, where such models form a darker squared region in the upper left of the plot. The same results is summarized in Table 6, where it is shown how scores for each metrics, averaged across models, are higher when obtained from the comparisons across model definitions rather than when those are compared to the gold standard. This results show that models' generated definitions, especially in frontier models, are aligned one to the other both semantically and formally, implying reduced linguistic and conceptual variation.

C.3 Data Examples

Examples of the data used to build experimental items are given in Table 7 for Task 1 and Table 8 for Task 2 and 3. English translations are literal and automatically generated with Google translate.

(IT)

ISTRUZIONI

Sei un esperto in materia di giurisprudenza italiana e diritto italiano.
Di seguito ti verrà sottoposto un quesito e delle possibili risposte. Il tuo compito è selezionare la risposta corretta in base al quesito.
Restituisci in output solo il label corrispondente alla risposta corretta selezionando tra [A, B, C, D] e null'altro.
*Non devi** assolutamente aggiungere altro testo o spiegazioni alla risposta, ma solo il label della risposta corretta.

QUESITO {quesito}

RISPOSTE POSSIBILI {risposte_possibili}

Risposta:

(ENG)

INSTRUCTIONS

You are an expert in Italian jurisprudence and Italian law.
Below, you will be presented with a question and possible answers. Your task is to select the correct answer based on the question.
Return only the label corresponding to the correct answer by selecting between [A, B, C, D] and nothing else.
*You must not** add any additional text or explanations to the answer, just the label of the correct answer.

QUESTION {question}

POSSIBLE ANSWERS {possible answers}

Answer:

Figure 5: Prompt template used for Task 1 in Italian and its English translation.

(IT)

Che cosa significa "*enfiteusi*" in diritto italiano?
Fornisci una breve ma esaustiva definizione del concetto.
Se possibile la definizione deve consistere in una frase.
Fornisci esclusivamente la definizione e non aggiungere altro testo.

(ENG)

What does "*emphyteusis*" mean in Italian law?
Provide a brief but comprehensive definition of the concept.
If possible, the definition should consist of one sentence.
Provide only the definition and do not add any additional text.

Figure 6: Prompts for the Definition Modelling task, with an exemplar term, *enfiteusi* (eng. *enphyteusis*

(IT)

ISTRUZIONI
Sei un esperto di Giurisprudenza Italiana.

Fornisci il concetto più adatto che corrisponde alla seguente definizione.

Il concetto può essere costituito da una o più parole, ma non deve contenere spiegazioni o commenti.
Devi restituire solo il concetto, senza spiegazioni o commenti.

Definizione:
Risposta:

(ENG)

INSTRUCTIONS
You are an expert in Italian Law.
Provide the most suitable concept that corresponds to the following definition.

The concept may consist of one or more words, but must not contain explanations or comments.

You must return only the concept, without explanations or comments.

Definition:
Answer:

Figure 7: Prompts for the Reverse Dictionary task in *generative* form

Model	Adm. Proc. Law	Adm. Law	Priv. Law	Civ. Proc.	Glob. (M)
Claude-sonnet-4.5	80.21 (± 1.24)	88.82 (± 0.84)	93.69 (± 0.36)	80.50 (± 0.66)	85.80
Deepseek-v3	73.88 (± 0.98)	73.66 (± 2.45)	81.81 (± 1.07)	69.69 (± 0.86)	74.76
Gemini-3-flash	92.37 (± 0.60)	95.36 (± 0.72)	98.66 (± 0.36)	92.44 (± 0.16)	**94.71**
Gemini-flash-2.5	78.60 (± 0.74)	84.76 (± 0.29)	93.22 (± 0.21)	79.47 (± 0.82)	84.01
Gpt-5	85.67 (± 0.33)	90.00 (± 0.87)	93.75 (± 0.18)	84.62 (± 0.39)	88.51
Gpt4-o	78.48 (± 0.82)	84.09 (± 0.20)	88.78 (± 1.14)	75.75 (± 0.73)	81.78
Llama-3.1-8b-instruct	58.47 (± 1.22)	53.00 (± 0.56)	57.94 (± 0.63)	52.87 (± 1.02)	55.57
Llamantino-3-anita-8b-inst-dpo-ita	57.23 (± 1.76)	58.52 (± 0.77)	58.69 (± 1.43)	52.56 (± 1.25)	56.75
Minerva-7b-instruct-v1.0	41.22 (± 1.61)	33.45 (± 0.60)	39.79 (± 0.89)	38.38 (± 1.08)	38.21
Mistral-7b-instruct	35.11 (± 0.43)	34.27 (± 2.50)	34.03 (± 1.76)	33.38 (± 2.02)	34.20
Glob. (D)	68.12	69.59	**74.03**	65.96	

Table 5: Average accuracy and standard deviation across four shuffles for each law sub-domain. Global accuracy with respect to each model averaged across law-sub field (M) and for each dataset averaged across models (D).

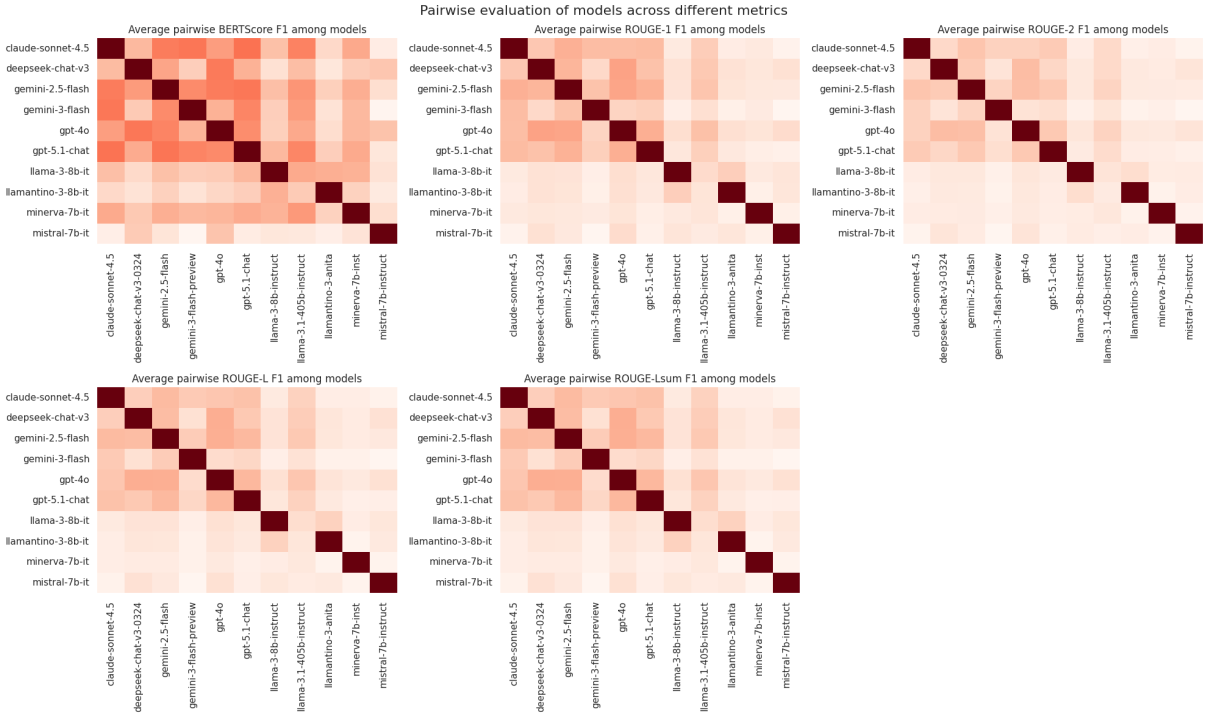

Figure 8: Cross models analysis for Task 2. Models' generated definitions are compared across each other.

Metric	**Average**	
	Model@Gold	Model@Model
BERTScore	0.44	0.64
R-1	0.24	0.43
R-2	0.06	0.24
R-L	0.17	0.36
BLEU	0.03	0.18

Table 6: Comparison between average values obtained for each metric comparing models' definitions and the gold standard (left) and across models' generations.

Question	Choices
Se una delle parti che conclude il contratto è legalmente incapace di contrattare, il contratto è: (eng: *If one of the parties entering into the contract is legally incapable of contracting, the contract is:*)	**Annullabile**, Nullo, Inefficace, Rescindibile (eng: ***Voidable**, Null, Ineffective, Rescindable*)
Come sono considerate dall'art. 812 del codice civile le costruzioni unite al suolo a scopo transitorio? (eng: *How are buildings attached to the land for temporary purposes treated under Article 812 of the Civil Code?*)	**Beni immobili**, Beni mobili, Pertinenza di beni immobili, Pertinenza di beni mobili (eng : ***Immovable property**, Movable property, Appurtenances of immovable property, Appurtenances of movable property*)
A chi appartengono i beni pubblici demaniali? (eng: *A chi appartengono i beni pubblici demaniali?*)	**Allo Stato o ad altri enti territoriali**, Solo allo Stato, Ai privati che li acquistano, Ai comuni (eng: ***To the State or other territorial bodies**, Only to the State, To private individuals who purchase them, To municipalities*)
Il diritto dell'enfiteuta si prescrive per effetto dell'uso non protratto per: (eng: *The emphyteuta's right expires as a result of non-prolonged use for:*)	**Venti anni**, dieci anni, quindici anni, cinque anni (eng: ***Twenty years**, ten years, fifteen years, five years*)
Il contratto col quale una parte consegna all'altra una determinata quantità di denaro o di altre cose fungibili, e l'altra si obbliga a restituire altrettante cose della stessa specie e qualità, è un contratto di: (eng: *The contract by which one party delivers to the other a certain quantity of money or other fungible things, and the other undertakes to return an equal quantity of things of the same kind and quality, is a contract of:*)	**Mutuo**, Vendita, Deposito, Commissione (eng: ***Mortgage**, Sale, Deposit, Commission*)

Table 7: Pre-shuffling examples questions for Task 1. Correct answers are in boldface.

Term	Definition
Abbuono (eng: *Rebate*)	Con il termine "abbuono" si indicano i componenti positivi o negativi di reddito che rettificano i costi di acquisto o i ricavi di vendita di merci che presentano caratteristiche difformi da quelle richieste. (eng: *The term "rebate" refers to the positive or negative income components that correct the purchase costs or sales revenues of goods whose characteristics differ from those required.*)
Concorrenza (eng: *Competition*	Termine che si usa per indicare la competizione tra persone o enti sul mercato (in un determinato settore), finalizzata all'acquisizione e alla conservazione della clientela. (eng: *A term used to indicate contest between people or entities on the market (in a given sector), aimed at acquiring and retaining customers.*)
Erede (eng: *Heir*)	è il chiamato alla successione, per legge o per testamento, che abbia accettato l'eredità, subentrando così nei rapporti giuridici trasmissibili di cui faceva parte il de cuius. (eng: *it is the person called to the succession, by law or by will, who has accepted the inheritance, thus taking over the transmissible legal relationships of which the deceased was a part.*)
Giudice istruttore (eng: *Investigating judge*)	è il giudice-persona fisica (appartenente alla Magistratura) cui é demandata la direzione della fase c.d. istruttoria del procedimento, ossia quella che precede la definizione della controversia. (eng: *he is the judge-natural person (belonging to the Judiciary) who is entrusted with the direction of the so-called investigative phase of the proceedings, that is, the one preceding the resolution of the dispute.*)
Obbligazione naturale (eng: *Natural obligation*)	Con questa locuzione ci si riferisce a tutte quelle prestazioni effettuate spontaneamente, in esecuzione di doveri morali o sociali, che non sono ripetibili da parte di colui che le ha effettuate. (eng: *This expression refers to all those services performed spontaneously, in fulfillment of moral or social duties, which cannot be repeated by the person who performed them.*)

Table 8: Examples of the raw data used for Task 2 and 3, made of term-definition pairs.