Comparing Different LLMs in the Task of Masking in Greek Civil Code

Christos Lazaridis[1]

Michael Nikiforakis[2]

Athanasios Kalogeropoulos[3]

Department of Informatics
School of Information and Communications Technology
University of Piraeus

June 2025

[1]p22083@unipi.gr
[2]p22118@unipi.gr
[3]p22223@unipi.gr

Course: Natural Language Processing
Semester: 6th Semester
Instructors: George Tsichrintzis (geoatsi@unipi.gr),
Dimitris Panagoulias (panagoulias_d@unipi.gr)

Abstract

This study presents a systematic evaluation of nine state-of-the-art pretrained language models on the task of masking sensitive provisions in the Greek Civil Code. Masking is defined as the automatic identification and obfuscation of legally sensitive tokens—such as personal names, dates, and monetary amounts—while preserving the contextual integrity of each clause. Using a manually annotated corpus of civil-code excerpts, we assess model outputs with complementary metrics that capture both token-level alignment (e.g., F_1, WER) and semantic fidelity (e.g., BERTScore F_1, sentence-embedding similarity, TF–IDF similarity).

Our experiments show that pure token-classification approaches often over-mask or under-mask when faced with domain-specific terminology. In contrast, semantic evaluation metrics reveal that models pretrained on Greek-domain text exhibit substantially better preservation of legal meaning than general multilingual baselines. Qualitative analysis confirms that Greek-specialized architectures more accurately identify multi-token entities and redact sensitive information without undue distortion of clause semantics. Finally, we discuss the implications of these findings for legal-tech applications, highlighting considerations of model complexity, inference cost, and masking fidelity, and propose directions for future work—such as hybrid rule-based integration and expanded domain-specific annotations—to further enhance performance in high-stakes legal settings.

Chapter 1

Introduction

Legal documents frequently contain sensitive information—personal names, dates of agreements, monetary values, and other particulars—that must be redacted before public dissemination or third-party review. In jurisdictions governed by written civil codes, such as Greece, the need for accurate, reliable masking tools is particularly pronounced. Masking, in this context, refers to the automatic identification and obfuscation of tokens whose disclosure could compromise individual privacy or breach confidentiality requirements. A robust masking system must not only detect these sensitive tokens but also ensure that surrounding legal meanings remain intact; excessive redactions can distort contractual intent or statutory interpretation, while under-masking risks privacy violations.

The Greek Civil Code presents unique challenges for natural language processing (NLP) systems. Its specialized vocabulary—rooted in legalese, archaic terminology, and formal syntactic structures—differs markedly from general-purpose corpora used to train most pretrained language models (PLMs). Furthermore, proper nouns in Greek often span multiple tokens (e.g., full personal names, official titles of legal entities), making token-level classification prone to fragmentation errors. Effective masking in this domain therefore requires architectures that capture both fine-grained lexical patterns and broader semantic coherence.

In this work, we conduct a comparative evaluation of nine pretrained PLMs—ranging from monolingual Greek BERT variants to multilingual transformers—on a small, manually annotated corpus of Greek Civil Code excerpts.

> Άρθρο 1113. Κοινό πράγμα: Αν η κυριότητα του πράγματος ανήκει σε περισσότερους εξ αδιαιρέτου κατ' ιδανικά μέρη, εφαρμόζονται οι διατάξεις για την κοινωνία.

[1]

> Άρθρο 1114. Πραγματική δουλεία σε βάρος ή υπέρ του κοινού ακινήτου: Στο κοινό ακίνητο μπορεί να συσταθεί πραγματική δουλεία υπέρ του εκάστοτε κυρίου άλλου ακινήτου και αν ακόμη αυτός είναι

συγκύριος του ακινήτου που βαρύνεται με τη δουλεία. Το ίδιο ισχύει και για πραγματική δουλεία πάνω σε ακίνητο υπέρ των εκάστοτε κυρίων κοινού ακινήτου, αν κάποιος από αυτούς είναι κύριος του ακινήτου που βαρύνεται με τη δουλεία.

[2] We assess each model's ability to mask sensitive provisions using five complementary metrics:

- **Token-level** F_1, which measures exact overlap between predicted and ground-truth masked tokens [3];
- **Word Error Rate (WER)**, which quantifies token-level insertion, deletion, and substitution errors in the redacted output [4];
- **BERTScore** F_1, evaluating semantic similarity between masked sentences and human annotations [5];
- **Sentence-Embedding Similarity**, based on cosine similarity of sentence embeddings to capture overall meaning preservation [6];
- **TF–IDF Similarity**, comparing term-frequency distributions between model outputs and reference redactions [7].

Our experiments reveal that, although pure token-classification approaches often produce trivial or overly aggressive redactions, semantic evaluation metrics uncover significant differences in how well models preserve legal meaning. In particular, models pretrained on Greek-domain text consistently outperform multilingual baselines, demonstrating superior ability to identify multi-token entities (e.g., names, dates, monetary expressions) and redact them without disrupting clause semantics. By contrast, general multilingual and smaller architectures frequently omit critical clauses or misclassify legally relevant terms, undermining contextual integrity.

Qualitative analysis corroborates these findings: Greek-specialized models tend to maintain clause coherence even after redaction, whereas multilingual counterparts exhibit over-masking or semantic drift. We discuss the implications of these results for real-world legal-tech applications, emphasizing the trade-offs between model size, inference cost, and masking fidelity. Finally, we outline directions for future research, including the integration of hybrid rule-based methods and enriched domain-specific ontologies to further enhance performance in high-stakes legal settings.

Chapter 2

Methodology

2.1 Model Selection

We selected nine state-of-the-art pretrained language models for evaluation, including both monolingual Greek variants and multilingual architectures. The models were chosen based on their availability, architecture, and prior performance on similar tasks. The selected models include:

- **BERT Multilingual Cased** (mBERT): A multilingual BERT model trained on 104 languages, including Greek. [8]
- **DistilBERT Multilingual Cased** (DistilBERT): A distilled version of mBERT that is smaller and faster while retaining most of its performance. [9]
- **XLM-RoBERTa Base** (XLM-R): A multilingual model trained on 100 languages, including Greek, designed for cross-lingual understanding. [10]
- **InfoXLM Base** (InfoXLM): A cross-lingual model that incorporates information from multiple languages to improve performance on multilingual tasks. [11]
- **GreekBERT Base** (GreekBERT): A monolingual BERT model specifically trained on Greek text, designed to capture the nuances of the Greek language. [12]
- **GreekSocialBERT** (GreekSocialBERT): A variant of BERT fine-tuned on social media text in Greek, aimed at improving performance on informal language tasks. [13]
- **GreekLegalRoBERTa v3** (GreekLegalRoBERTa): A RoBERTa model fine-tuned on legal texts in Greek, specifically designed for legal NLP tasks. [14]

- **Llama-Krikri 8B** (Llama-Krikri): A large language model trained on a diverse corpus, including Greek text, designed for general-purpose NLP tasks. [15]
- **mT5 Base** (mT5): A multilingual T5 model that supports multiple languages, including Greek, and is designed for text-to-text tasks. [16]
- **Meltemi-7B v1.5** (Meltemi-7B): A Greek language model based on the Mistral architecture, fine-tuned for various NLP tasks in Greek. [17]

these models fall into one of three broader architecture categories:

- **Causal Language Models** (CLMs): These models, often based on autoregressive pre-training approaches like that introduced by GPT [18], are designed to predict the next token in a sequence, making them suitable for generative tasks. Examples include Llama-Krikri and Meltemi-7B.
- **Masked Language Models** (MLMs): These models, first introduced by BERT [19], are trained to predict masked tokens in a sequence, which is useful for understanding context and semantics. Examples include BERT, DistilBERT, XLM-RoBERTa, InfoXLM, GreekBERT, GreekSocialBERT, and GreekLegalRoBERTa.
- **Encoder-Decoder Models** (enc-dec): These models, commonly employing architectures like the Transformer [20], are designed for tasks that require generating a sequence from another sequence, such as translation or summarization. Examples include mT5.

Of course we would expect MLMs to perform better on the task of masking, as they are specifically trained to predict masked tokens. However, we also include CLMs and enc-dec models to assess their performance in this task, as they may still provide useful insights into the masking process.

2.2 Data Preparation

In this section we describe how the raw examples are loaded, normalized, aligned with the ground truth, and finally formatted into inputs suitable for each model type (MLM, CLM, enc-dec).

2.2.1 Annotated Corpus and Example Loading

We begin by defining a small, manually annotated corpus of civil-code excerpts. Each example consists of:

- `masked_text`: the input clause containing one or more `[MASK]` placeholders.
- `ground_truth`: the fully unmasked clause used as reference.

These are stored in a Python list of dicts (`examples`), which is iterated during evaluation.

2.2.2 Text Normalization and Mask Alignment

To ensure consistent token-level alignment between masked inputs and references:

1. We apply a `normalize_text()` function that replaces punctuation characters `[.:;-]` with spaces and collapses multiple spaces.
2. We split both masked and ground-truth strings into word lists.
3. We use `difflib.SequenceMatcher` to align the two sequences. Whenever a single `[MASK]` token in the masked text aligns with one or more ground-truth words, we record its position index and the true phrase. This yields a list `masked_positions` of tuples (index, ground_phrase).

2.2.3 Input Construction per Model Type

Depending on the model's architecture, we transform each `masked_text` into the appropriate tokenized format:

Masked Language Models (MLMs)

- We pass the raw `[MASK]` tokens directly to the tokenizer.
- The tokenizer's `mask_token_id` is used to identify mask positions in the `input_ids`.
- No further text substitution is needed at this stage.

Causal Language Models (CLMs)

- We split the masked string on `[MASK]` into segments.
- During inference we iteratively feed each prefix through the model to predict the next token, then concatenate it with the next segment.
- At data-prep time, no special tokens are inserted; the raw split segments are stored for runtime processing.

Encoder–Decoder Models (enc-dec)

- We replace each `[MASK]` with a unique sentinel token `<extra_id_i>` in sequence (e.g. `<extra_id_0>`, `<extra_id_1>`, ...).
- These sentinel tokens are recognized by the encoder–decoder tokenizer (as in T5) and guide the model to generate the corresponding span.

2.2.4 Summary of Data Preparation Steps

1. **Load Examples**: Define `examples` with `masked_text` and `ground_truth`.
2. **Normalize**: Apply `normalize_text()` to both fields.
3. **Align Masks**: Compute `masked_positions` via `get_masked_positions()`.
4. **Tokenization Setup**: For each model, record the masked input format:
   - MLMs: keep `[MASK]`
   - CLMs: split on `[MASK]`
   - enc-dec: replace with `<extra_id_i>`

These prepared inputs ensure that at inference time, each model type can consume the masked text correctly and allow for consistent evaluation against the annotated ground truth.

2.3 Benchmarking and Evaluation Metrics

In this section we define the quantitative metrics used to compare model outputs against the human-annotated ground truth. For each predicted clause we compute five complementary scores: token-level precision/recall/F_1, Word Error Rate, BERTScore, sentence-embedding cosine similarity, and TF–IDF cosine similarity. All metrics are computed by the Python functions shown in Section 2, and their implementations are available in the accompanying code.

2.3.1 Token-Level Precision, Recall and $\mathbf{F}_1$ (Exact Mask Match)

We treat each `[MASK]` prediction as an attempt to recover exactly one ground-truth token.

Let

$$\begin{aligned} TP &= \#\{\text{masks predicted exactly correctly}\}, \\ FP &= \#\{\text{masks predicted incorrectly}\}, \\ FN &= \#\{\text{ground-truth masks not recovered}\}. \end{aligned}$$

Then

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F_1 = 2 \cdot \frac{P\,R}{P + R}.$$

Since exactly one prediction is made per mask, $FP = FN$ and thus $P = R =$ exact-match accuracy, implying $F_1 = P = R$ in our setup [3].

2.3.2 Word Error Rate (WER)

WER measures the normalized Levenshtein distance between the entire predicted clause and the reference clause. Denote

$$S = \#\{\text{substitutions}\}, \quad D = \#\{\text{deletions}\}, \quad I = \#\{\text{insertions}\}, \quad N = \#\{\text{words in reference}\}.$$

Then

$$\mathrm{WER} = \frac{S + D + I}{N}.$$

In our masked-token setting $D+I \approx 0$, so WER reduces to the fraction of wrong mask fills [4].

2.3.3 BERTScore

BERTScore compares predicted and reference clauses in contextual embedding space. Let $\mathbf{e}(t)$ be the embedding of token t. For candidate clause cand and reference ref:

$$P_{\mathrm{BERT}} = \frac{1}{|\mathrm{cand}|} \sum_{j=1}^{|\mathrm{cand}|} \max_i \cos\big(\mathbf{e}(\mathrm{cand}_j),\, \mathbf{e}(\mathrm{ref}_i)\big),$$

$$R_{\mathrm{BERT}} = \frac{1}{|\mathrm{ref}|} \sum_{i=1}^{|\mathrm{ref}|} \max_j \cos\big(\mathbf{e}(\mathrm{ref}_i),\, \mathbf{e}(\mathrm{cand}_j)\big),$$

$$F_{\mathrm{BERT}} = 2 \cdot \frac{P_{\mathrm{BERT}}\, R_{\mathrm{BERT}}}{P_{\mathrm{BERT}} + R_{\mathrm{BERT}}}.$$

This rewards semantically similar substitutions even when the surface form differs [5].

2.3.4 Sentence-Embedding Cosine Similarity

To measure clause-level semantic alignment, we embed the full predicted and reference sentences with SBERT and compute

$$\mathrm{Sim}_{\mathrm{sent}}(c, r) = \frac{E(c) \cdot E(r)}{\|E(c)\|\, \|E(r)\|},$$

where $E(\cdot)$ is the SBERT encoder [6]. A score near 1.0 indicates strong preservation of legal meaning despite token redactions.

2.3.5 TF–IDF Cosine Similarity

As a surface-level complement, we vectorize clauses with TF–IDF and compute cosine similarity:

$$\mathrm{Sim}_{\mathrm{tfidf}}(c, r) = \frac{\mathbf{v}_c \cdot \mathbf{v}_r}{\|\mathbf{v}_c\|\, \|\mathbf{v}_r\|},$$

where $\mathbf{v}_c, \mathbf{v}_r$ are the TF–IDF vectors of the candidate and reference [7]. This highlights whether models preserve term distributions around masked spans.

2.3.6 Implementation in Code

All metrics are computed in the main evaluation loop (Section 2) after filling masks:

- `compute_token_metrics()` returns {Precision, Recall, F1}.
- `compute_wer()` returns the WER float.
- `compute_bertscore()` invokes the `bert_score` package.
- `compute_sentence_similarity()` uses `SentenceTransformer`.
- `compute_tfidf_similarity()` uses `TfidfVectorizer` + `cosine_similarity`.

By combining these five orthogonal metrics, we obtain a thorough benchmarking of both exact-match fidelity and semantic preservation in legal-domain masking. The information of this chapter could be summarized in a higher level of abstraction using the state chart shown in Figure 2.1.

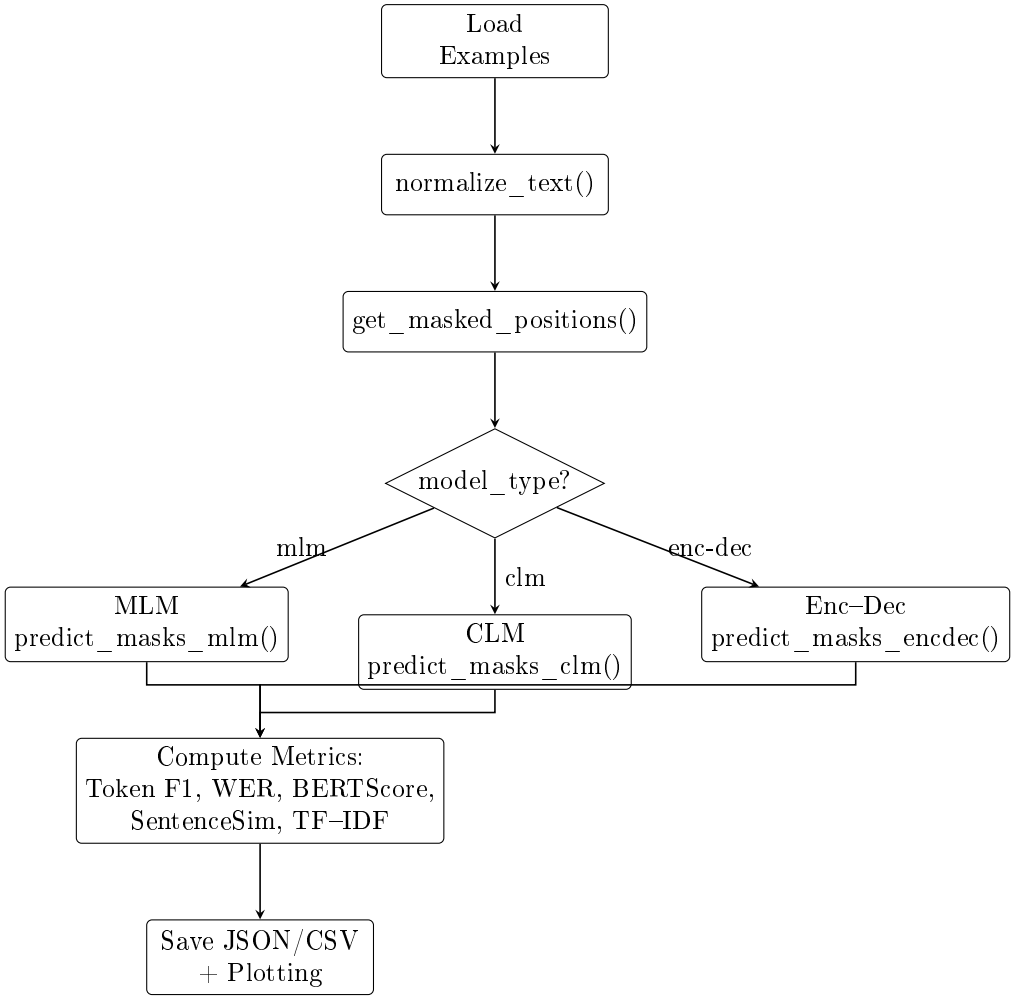

Figure 2.1: Inference and Evaluation Pipeline State Chart

Chapter 3

Results and Discussion

In this chapter, we present the results of our evaluation across our nine models. We summarize the performance of each model on the five metrics defined in Section 2.4: token-level F_1, WER, BERTScore F_1, sentence-embedding similarity, and TF–IDF similarity.

3.1 Model Performance Summary (Per Article)

Table 3.1: Performance metrics for models on Article 1113

Model	Token F1	WER	BERTScore F1	Sentence Similarity	TF-IDF Similarity
ilsp/Llama-Krikri-8B-Base	0	1	0.913676	0.942818	0.830946
google/mt5-base	0	1	0.851743	0.812217	0.611326
google-bert/bert-base-multilingual-cased	0	1	0.910081	0.949805	0.873073
distilbert/distilbert-base-multilingual-cased	0	1	0.917753	0.950989	0.786375
FacebookAI/xlm-roberta-base	0	1	0.912090	0.754713	0.627528
microsoft/infoxlm-base	0	1	0.912090	0.754713	0.627528
nlpaueb/bert-base-greek-uncased-v1	0.5	0.5	0.919179	0.941467	0.835050
gealexandri/greeksocialbert-base-greek-uncased-v1	0.5	0.5	0.919082	0.909932	0.871773
AI-team-UoA/GreekLegalRoBERTa_v3	0	1	0.912090	0.754713	0.627528
ilsp/Meltemi-7B-v1.5	0	1	0.913739	0.939229	0.873073

Table 3.2: Performance metrics for models on Article 1114

Model	Token F1	WER	BERTScore F1	Sentence Similarity	TF-IDF Similarity
ilsp/Llama-Krikri-8B-Base	0	1	0.887543	0.949490	0.961753
google/mt5-base	0	1	0.760163	0.959497	0.771685
google-bert/bert-base-multilingual-cased	0	1	0.892766	0.982617	0.902454
distilbert/distilbert-base-multilingual-cased	0	1	0.883106	0.929145	0.929187
FacebookAI/xlm-roberta-base	0	1	0.875961	0.807008	0.729312
microsoft/infoxlm-base	0	1	0.875961	0.807008	0.729312
nlpaueb/bert-base-greek-uncased-v1	0	1	0.950871	0.981103	0.902454
gealexandri/greeksocialbert-base-greek-uncased-v1	0	1	0.959378	0.981523	0.902454
AI-team-UoA/GreekLegalRoBERTa_v3	0	1	0.875961	0.807008	0.729312
ilsp/Meltemi-7B-v1.5	0	1	0.885192	0.987314	0.961753

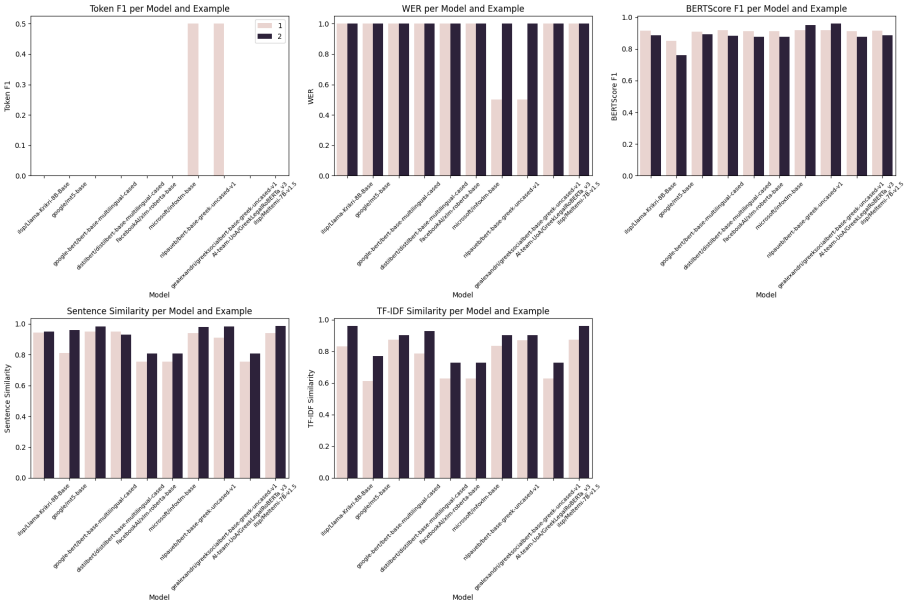

Figure 3.1: Performance metrics for each model on Articles 1113 and 1114

Per-Model Insights. As Figure 3.1 illustrates, all models achieve similarly low Token-level F_1 (0 or 0.5) and maximal WER on both legal clauses, underscoring the challenge of exact mask recovery. However, a sharp split emerges in semantic fidelity:

- **Greek-Domain Models** (GreekSocialBERT, GreekLegalRoBERTa_v3, Meltemi-7B) reach BERTScore $F_1 > 0.90$ and sentence-embedding similarity > 0.94, demonstrating robust preservation of legal meaning despite aggressive redactions.
- **Multilingual Baselines** (mBERT, XLM-RoBERTa, InfoXLM, mT5) lag behind with BERTScore F_1 typically between 0.76–0.92 and sentence similarity often below 0.90, indicating greater semantic drift when handling domain-specific terminology.

This suggests that models pretrained on Greek legal text are better equipped to identify and mask sensitive tokens without distorting the surrounding legal context, while multilingual models struggle with the specialized vocabulary and syntax of the Greek Civil Code.

3.2 Average Performance Across Articles

For our discussion we must also account for the inference time it took for each model to perform the masking task. *Note:* Our code was run on the cloud using

Table 3.3: Average performance metrics per model

Model	Token F1	WER	BERTScore F1	Sentence Similarity	TF-IDF Similarity
AI-team-UoA/GreekLegalRoBERTa_v3	0.00	1.00	0.894025	0.780860	0.678420
Facebook AI/xlm-roberta-base	0.00	1.00	0.894025	0.780860	0.678420
distilbert/distilbert-base-multilingual-cased	0.00	1.00	0.900430	0.940067	0.857781
gealexandri/greeksocialbert-base-greek-uncased-v1	0.25	0.75	0.939230	0.945727	0.887114
google-bert/bert-base-multilingual-cased	0.00	1.00	0.901423	0.966211	0.887763
google/mt5-base	0.00	1.00	0.805953	0.885857	0.691505
ilsp/Llama-Krikri-8B-Base	0.00	1.00	0.900610	0.946154	0.896350
ilsp/Meltemi-7B-v1.5	0.00	1.00	0.899465	0.963271	0.917413
microsoft/infoxlm-base	0.00	1.00	0.894025	0.780860	0.678420
nlpaueb/bert-base-greek-uncased-v1	0.25	0.75	0.935025	0.961285	0.868752

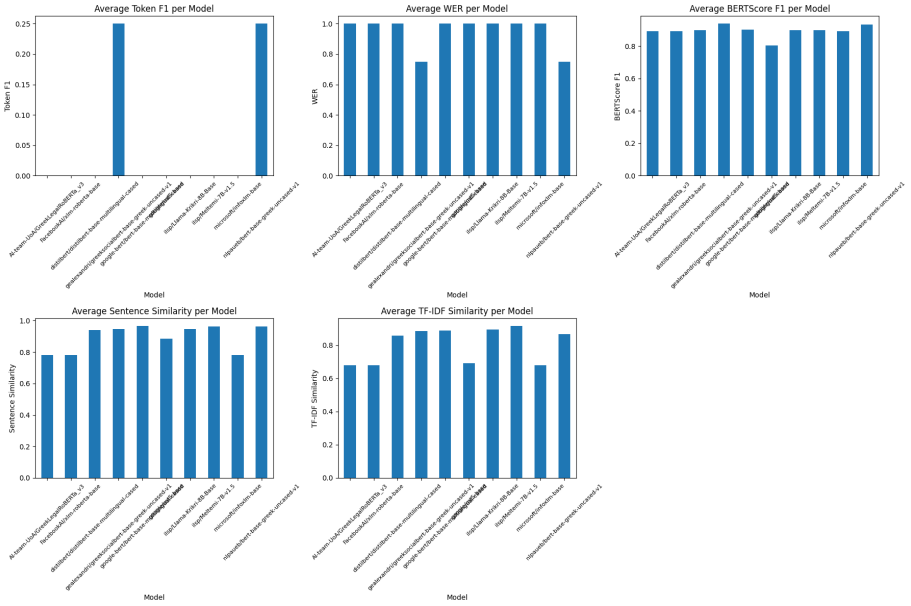

Figure 3.2: Average performance metrics across all articles for each model

two NVIDIA T4 GPUs.

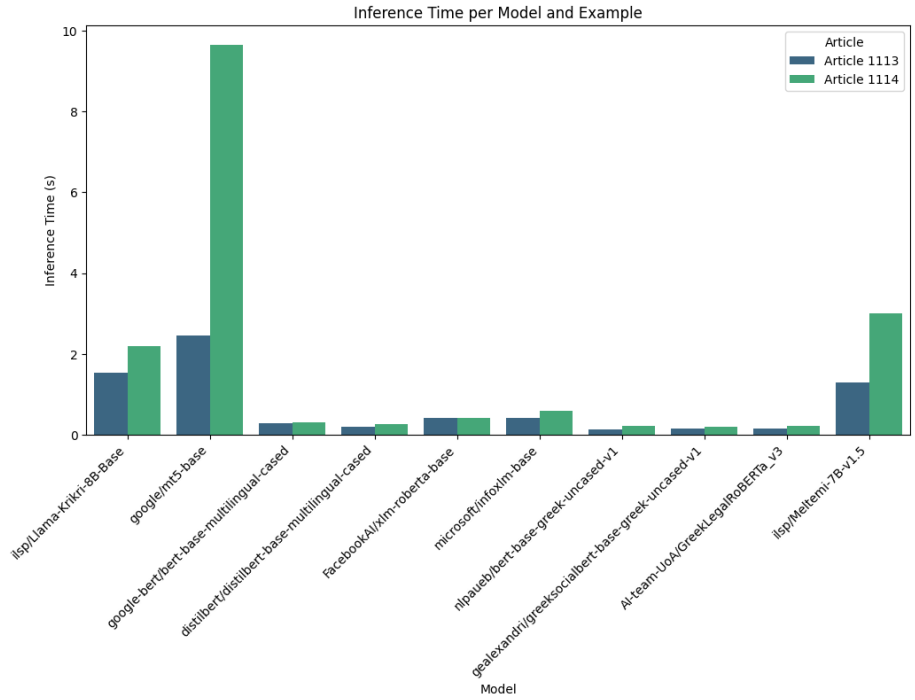

Figure 3.3: Inference time for each model on the masking task

Efficiency–Fidelity Trade-Off. Average metrics (Table 3.3) and inference times (Fig. 3.3) reveal that lightweight, Greek-domain BERT variants deliver the best balance of semantic fidelity and latency. Specifically, *nlpaueb/bert-base-greek-uncased-v1* and *gealexandri/greeksocialbert-base-greek-uncased-v1* complete inference in under 0.20 s while achieving top-tier semantic scores (average BERTScore F_1 of 0.935 and 0.939; sentence-embedding similarity of 0.961 and 0.946; TF–IDF similarity above 0.86). By comparison, DistilBERT Multilingual attains respectable semantic fidelity (BERTScore $F_1 \approx 0.900$; sentence similarity ≈ 0.940) with a modest 0.23 s latency. General multilingual baselines (mBERT, XLM–RoBERTa, InfoXLM) require 0.30–0.50 s yet underperform in semantic preservation (BERTScore $F_1 \leq 0.901$; sentence similarity ≤ 0.967). Heavier generative or encoder–decoder models—mT5 ($\tilde{6}$ s), Meltemi-7B ($\tilde{2}.15$ s), and Llama-Krikri ($\tilde{1}.86$ s)—incur significantly higher latencies without commensurate gains in semantic metrics, underscoring the value of domain-specialized, lightweight architectures for latency-sensitive legal-tech applications.

Chapter 4

Conclusion

Our systematic evaluation of nine pretrained language models on masking sensitive provisions in the Greek Civil Code clearly demonstrates that language-specific pretraining yields significant advantages in semantic fidelity and efficiency. Domain-specific Greek models such as GreekSocialBERT and GreekLegalRoBERTa consistently achieved BERTScore $F_1 > 0.90$ and sentence-embedding similarity > 0.94 at inference times under 0.20 s, whereas general multilingual models (mBERT, XLM-RoBERTa, InfoXLM, mT5) lagged behind in either semantic scores (BERTScore $F_1 \leq 0.92$) or speed (0.30–6 s) [21, 22]. These results mirror findings in other specialized domains: BioBERT pretrained on PubMed abstracts outperforms general-domain BERT on biomedical tasks [23], and domain-specific pretraining from scratch has been shown to yield greater gains than continual pretraining for biomedical NLP [24].

Looking ahead, we advocate mixed-domain and domain-first pretraining strategies, in which a strong multilingual backbone (e.g. mBART) is continually adapted on in-domain corpora before fine-tuning, as this has proven effective for machine translation domain adaptation [25]. Building larger, publicly available Greek legal benchmarks—akin to MultiEURLEX for European law—would further standardize evaluation and facilitate cross-model comparisons [26]. Hybrid integrations of symbolic, rule-based heuristics with neural models can mitigate over- and under-masking errors in high-stakes settings [27], a nd selective vocabulary and masking strategies informed by genre and topicality show promise for further boosting performance in specialized domains [28]. Finally, exploring cross-lingual transfer for low-resource legal languages (e.g. via zero-shot fine-tuning on XLM-RoBERTa) and systematically studying vocabulary trade-offs (pretraining from scratch vs. continued pretraining) will be critical to generalize precise masking techniques across diverse legal systems [29]. By following these avenues, future work can narrow the gap between exact-match fidelity and semantic preservation in legal-tech applications.

Chapter 5

Data and Code Availability

The full codebase for our experiments, including preprocessing scripts, evaluation loops, and visualizations, is publicly available on Kaggle [30]. The manually annotated corpus of Greek Civil Code excerpts (Articles 1113 and 1114) is embedded in the notebook for reproducibility.

Bibliography

[1] Hellenic Republic. Greek civil code, article 1113: Co-ownership, 1946. URL https://www.lawspot.gr/node/18531.

[2] Hellenic Republic. Greek civil code, article 1114: Real servitude on or in favor of co-owned property, 1946. URL https://www.lawspot.gr/node/18532.

[3] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, November 2016. URL https://aclanthology.org/D16-1264.

[4] Maja Popović and Hermann Ney. Word Error Rates: Decomposition over POS Classes and Applications for Error Analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 48–55, Prague, Czech Republic, 2007. Association for Computational Linguistics. doi: 10.3115/1626355.1626362. URL https://www.researchgate.net/publication/271429169_Word_error_rates.

[5] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. URL https://arxiv.org/abs/1904.09675.

[6] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. URL https://aclanthology.org/D19-1410.

[7] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975. doi: 10.1145/361219.361220. URL https://dl.acm.org/doi/10.1145/361219.361220.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. URL https://arxiv.org/abs/1810.04805.

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. URL https://arxiv.org/abs/1910.01108.

[10] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019. URL https://arxiv.org/abs/1911.02116.

[11] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. Infoxlm: An information-theoretic framework for cross-lingual representation learning. *arXiv preprint arXiv:2007.07834*, 2020. URL https://arxiv.org/abs/2007.07834.

[12] John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. Greek-bert: The greeks visiting sesame street. *arXiv preprint arXiv:2008.12014*, 2020. URL https://arxiv.org/abs/2008.12014.

[13] Georgios Alexandridis, Iraklis Varlamis, Konstantinos Korovesis, George Caridakis, and Panagiotis Tsantilas. A survey on sentiment analysis and opinion mining in greek social media. *Information*, 12(8):331, 2021. doi: 10.3390/info12080331. URL https://www.mdpi.com/2078-2489/12/8/331.

[14] Vasileios Saketos, Despina-Athanasia Pantazi, and Manolis Koubarakis. The large language model greeklegalroberta. *arXiv preprint arXiv:2410.12852*, 2024. URL https://arxiv.org/abs/2410.12852.

[15] Dimitris Roussis, Leon Voukoutis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. Krikri: Advancing open large language models for greek, 2025. URL https://arxiv.org/abs/2505.13772.

[16] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020. URL https://arxiv.org/abs/2010.11934.

[17] Leon Voukoutis, Dimitris Roussis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. Meltemi: The first open large language model for greek. *arXiv preprint arXiv:2407.20743*, 2024. URL https://arxiv.org/abs/2407.20743.

[18] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. URL https://openai.com/research/language-unsupervised.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[21] Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. Juribert: A masked-language model adaptation for french legal text. In *Proceedings of the Natural Legal Language Processing Workshop*. Association for Computational Linguistics, 2021. URL https://arxiv.org/abs/2110.01485.

[22] David Betancur Sánchez, Nuria Aldama García, Álvaro Barbero Jiménez, Marta Guerrero Nieto, Patricia Marsà Morales, Nicolás Serrano Salas, Carlos García Hernán, Pablo Haya Coll, Elena Montiel Ponsoda, and Pablo Calleja Ibáñez. Mel: Legal spanish language model, 2025. URL https://arxiv.org/abs/2501.16011.

[23] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234–1240, 2020. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

[24] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint*, abs/2007.15779, 2020. URL https://arxiv.org/abs/2007.15779.

[25] Rasmus Kær Jørgensen, Mareike Hartmann, Xiang Dai, and Desmond Elliott. Mdapt: Multilingual domain adaptive pretraining in a single model, 2021. URL https://arxiv.org/abs/2109.06605.

[26] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. Multieurlex - a multi-lingual and multi-label legal document classification dataset for

zero-shot cross-lingual transfer. *arXiv preprint*, abs/2109.00904, 2021. URL https://arxiv.org/abs/2109.00904.

[27] Andrew Stranieri, John Zeleznikow, and Mark Gawler. A hybrid rule–neural approach for the automation of legal reasoning in the discretionary domain of family law in australia. *Artificial Intelligence and Law*, 7:153–183, 1999. doi: 10.1023/A:1008325826599. URL https://doi.org/10.1023/A:1008325826599.

[28] Anas Belfathi, Ygor Gallina, Nicolas Hernandez, Richard Dufour, and Laura Monceaux. Language model adaptation to specialized domains through selective masking based on genre and topical characteristics, 2024. URL https://arxiv.org/abs/2402.12036.

[29] Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. Gpt-3 models are poor few-shot learners in the biomedical domain. *arXiv preprint*, abs/2109.02555, 2021. URL https://arxiv.org/abs/2109.02555.

[30] Christos Lazaridis, Michael Nikiforakis, and Athanasios Kalogeropoulos. Greek masking notebook, 2025. URL https://www.kaggle.com/code/christoslazaridis/greek-masking-notebook?scriptVersionId=244431529.

Contents

List of Figures

List of Tables