LawToken: a single token worth more than its constituents

Anonymous ACL submission

Abstract

Legal citations require correctly recalling the law references of complex law article names and article numbering, which large language models typically treat as multi-token sequences. Motivated by the form-meaning pair of constructionist approaches, we explore treating these multi-token law references as a single holistic law token and examining the implications for legal citation accuracy and differences in model interpretability. We train and compare two types of models: LawToken models, which encode the legal citations as a single law token, and LawBase models, which treat them as multi-token compounds. The results show that LawToken models outperform LawBase models on legal citation tasks, primarily due to fewer errors in the article numbering components. Further model representation analysis reveals that, while both models achieve comparable semantic representation quality, the multi-token-based LawBase suffers from degraded representations in multistep decoding, leading to more errors. Taken together, these findings suggest that form-meaning pairing can operate in a larger context, and this larger unit may offer advantages in future modeling of legal reasoning. In practice, this approach can significantly reduce the likelihood of hallucinations by anchoring legal citations as discrete, holistic tokens, thereby minimizing the risk of generating nonexistent or incorrect legal references.

1 Introduction

Recalling the correct legal citations, e.g., the law articles, regulations, or precedents, poses a great challenge to the large language models and raises an interesting question to computational linguistics (Guha et al., 2024; Dahl et al., 2024). While the autoregressive models are so adept at working with legal texts in certain, but not all, scenarios and tasks (Katz et al., 2024; Rodgers et al., 2023), generating the correct without producing non-existent articles or hallucinating remains a challenge to the modern models (Weiser, 2023; Henderson et al., 2023). While finding efficient ways to train LLMs adept at legal citations may potentially be addressed in future models, the linguistic intrigues nevertheless persist regarding how models encode the explicit textual forms and their impacts on the model's representations.

In current large language models (Dubey et al., 2024; Yang et al., 2024; Achiam et al., 2023a), these legal citations are treated as normal texts: processed by the tokenizer, they are chunked into a sequence of tokens. For example, the legal citation form in Taiwan generally is the article name followed by the article and paragraph numbers, such as "Road traffic safety regulations, Article 94, Paragraph 3." The model needs to learn how the multi-token sequence is related to the intended meanings in context.

The intended meaning of a cited law reference may entail the following three layers, in the order of their context-dependence: (1) the compositional meaning from the tokenized components, which, for instance, are the composite meanings of road traffic, safety, and others (Bell and Schäfer, 2016; Cordeiro et al., 2019; Alipoor and Schulte im Walde, 2020); (2) the semantic extensions of the legal text content, specifying the legal obligation of the driver (Tseng et al., 2023; Noraset et al., 2017; Mickus et al., 2019); and (3) the pragmatical usage of the law in the court verdict when determining the liability (Ruis et al., 2023; Louis et al., 2020; Parrish et al., 2021). In practice, the large language models might be good at deriving pragmatics and resolving the intended sense of the ambiguous words (tokens) from the constituting lexical semantics; but, in contrast, the hallucination (Guha et al., 2024; Bommasani et al., 2023; Dahl et al., 2024) suggests the model may struggle with decoding back from the context-specific pragmatic to the underlying constituent tokens.

An alternative approach is to map between the

layers *as direct as possible*; that is, treating law references as a single holistic form-meaning pair, where the entire law citation – including the law names and article or paragraph numbering – is recognized as one *law token*. These additional law tokens are motivated by the constructionist approach (Goldberg, 2024; Lakoff, 1987; Bybee, 2010). As linguistic units, from single words to multi-word idioms, function as form-meaning pairs, there is no theoretical limit on their scope except for cognitive constraints. However, computationally, large language models may already have enough capacity to capture the complex form-meaning mapping, provided they have clear cue-meaning mappings from tokenization.

This paper aims to empirically study the effect of tokenization on legal citations, focusing on both task performances and how tokenization affects the model's prediction probabilities and representation. Using the court verdicts of Taiwan, we compile a LawToken dataset containing 675M tokens. The dataset is used to fine-tune two types of models: LawBase models, which use the unmodified tokenizer, and LawToken models, which use an augmented tokenizer that includes frequently-used *law references* as new *law tokens*. When referring generically to using law tokens or references in the texts, we use the term *legal citation.* We first establish that LawToken models outperform LawBase models in legal citation tasks, and we next further analyze model representations, revealing that the performance difference may stem from the degraded contextualized representation during the multistep decoding in LawBase models.

This paper is organized as follows. After briefly summarizing the related works in Section 2, Section 3 describes the preprocessing steps, dataset, training, and evaluation of LawToken and LawBase models. Section 4 examines the model representations and explores how they differ in the two models. Section 5 concludes the paper.

2 Related Works

A legal reference, consisting of law or act names and article numbers, is composed of multiple tokens, which the language model has to learn to determine the intended meaning of the multi-token compound. However, past literature suggests that the compound meaning is not always transparent in terms of its constituent. Some are semantically transparent, such as "swimming pool," where the compound meaning is directly composite of the constituents; some are opaque, such as "hot dog." However, even a seemingly transparent compound may be challenging to pinpoint the relationships between its constituents; for instance, "airport" and "airplane" (compounds written without spaces), the role of "air" may be unexpectedly complicated (Bell and Schäfer, 2016; Reddy et al., 2011; Zwitserlood, 2014). Modeling the semantic transparency of compounds remains difficult, even when using static or contextualized semantic vectors (Shwartz and Dagan, 2019; Miletić and im Walde, 2023).

Some multi-token(word) expressions are not usually considered compounds but nevertheless convey meanings more than their parts. For example, "hazard a guess," or more idiom-like expression, "I hope this mail finds you well." These expressions, gaining their meaning through repeated uses by the language community and, therefore, form a static form-meaning pair, are *constructions* (Goldberg, 2013).

Along this line of reasoning, the law references can act as a construction. However, if the law reference is an opaque multi-token expression, the LLMs should already handle them to some extent (Goldberg, 2024). Yet, a previous study argued that the LLM's task performances are form-dependent (Ohmer et al., 2024), indicating that the models rely more on the surface form rather than the underlying meaning to complete the task. Consequently, even though the law reference is a construction, the way they are tokenized can significantly influence the model's task behavior.

Tokenizing law reference as a single law token has implications beyond linguistic theory. Using law tokens implies the model operates with a fixed set of "law vocabulary," which prevents the model from producing nonexistent law articles (Guha et al., 2024; Dahl et al., 2024). Although specialized legal-domain LLMs have become more prevalent, they are fine-tuned or continuously pretrained on legal texts or using retrieval-augmented generation without changing tokenization specifically for legal references (Colombo et al., 2024; Wiratunga et al., 2024; Lee, 2023; Cui et al., 2023). Furthermore, from an information-theoretic perspective, tokenization is the pre-compression in the LLM (Deletang et al., 2024). It is therefore interesting to observe how using a law token will change the compression behavior.

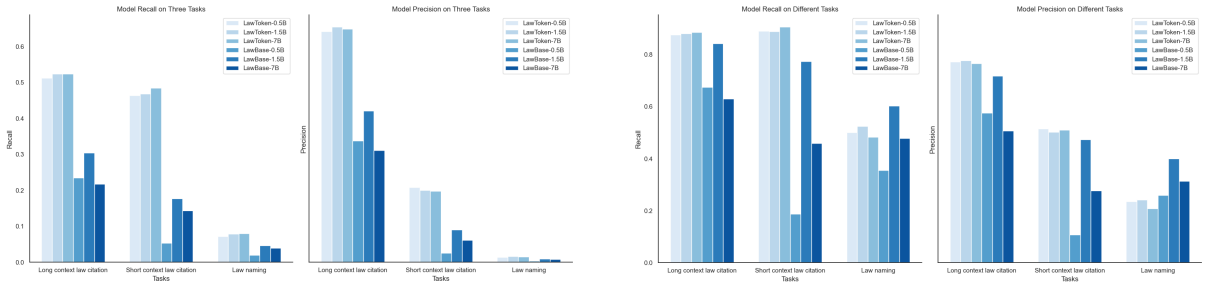


Figure 1: Evaluation results using full metric (left) and partial metric (right).

3 LawToken & LawBase models

3.1 Dataset

The LawToken Datasets[1] consist of legal documents publicly available in Taiwan, encompassing both law articles and court verdicts. The dataset has three parts. The first and second parts, composed of court verdicts and law articles, respectively, standardize law references in natural language by representing them in the following format: <*LAW_NAME*|*ARTICLE_NUMBER*>. For instance, a reference to 道路交通安全規則第94條第3項"road traffic safety regulations, Article 94, Paragraph 3." is transformed into the format <道路交通安全規則|94|3>. Conversely, the third group, derived solely from the court verdicts, employs a different transformation: legal references are removed from their original positions in the main text and then appended at the end of each court verdict, enclosed between a start-of-citation marker "<cite>" and an end-of-citation marker "</cite>." Examples of each group are provided in the Appendix. The three groups are combined and randomly shuffled. Subsequently, a train and test split is generated at a ratio of 9:1, resulting in a training set with 545.4k instances and a testing set with 60.6k instances.

3.2 Model Training

The three base models employed in this paper are Qwen2 of sizes 0.5B, 1.5B, and 7B [2]. We select the frequently occurred law references, namely, the total frequencies of the law references in the court verdicts need to be higher than 100 times, resulting in 13,083 law tokens. Subsequently, we train LawToken models with the high-frequency law tokens added into the tokenizer. The integration of the law tokens into the tokenizer enables the models to recognize the law references as single tokens and learn the contexts in which they are referenced. On the other hand, the LawBase uses the unmodified tokenizer. In other words, the mentions of law references in natural languages are represented as single tokens in LawToken models, whereas in the baseline LawBase models, they are interpreted as multi-token sequences.

Overall, six models are trained [3]. The fine-tuning uses 4 nVIDIA H100s and takes around 30 hours for all models. The evaluation cross-entropy losses of the LawToken models are .86, .79, and .69 for 0.5B, 1.5B, and 7.0B model sizes, respectively, and they are .82, .76, and .65 for the LawBase models. The evaluation loss decreases as the model size increases, whereas LawBase model losses are consistently lower than those of LawToken models.

3.3 Evaluation

The evaluation tasks include a long-context law citation task, a short-context law citation task, and a law naming task. These tasks, derived from the testing set, involve the same objective: predicting relevant LawTokens based on the provided context, with "<cite>" serving as the special token for prediction.

In the long-context law citation task, the model is provided with the full context of court verdicts, with law references removed, and is asked to predict the relevant legal citations. Conversely, the short-context law citation focuses on a more localized context, where sentences containing legal citations are identified, and the model is provided with only the preceding sentence as context to predict the relevant citations. The law naming task, on the other hand, is derived from law articles. Here, the model is presented solely with the content from a certain law article and is required to predict the

[1] https://huggingface.co/datasets/*******/LawToken. (masked during anonymous review)

[2] Models obtained from https://huggingface.co/Qwen

[3] All six models are available on HuggingFace, for instance the 7B finetuned model could be found at https://huggingface.co/********/LawToken-7B-a2.

correct law name and article number in the standardized format. Examples of each evaluation task are included in Appendix.

Overall, LawToken models demonstrate a significant advantage over LawBase models. The accuracies are estimated through recall and precision, in which recall calculates the numbers of correctly predicted law tokens divided by the numbers of correct law tokens, and precision is the numbers of correctly predicted law tokens divided by the numbers of predicted law tokens. In addition, out of all the unique law reference predictions produced by the LawBase models, 6.6% of them do not exist in those generated by the 0.5B model, 8.2% by the 1.5B model, and 7.6% by the 7B models. That is, the LawBase models still experience hallucinations after being specifically fine-tuned in the current dataset.

Figure 1 also visualizes the recall and precision of the six models on three different tasks. Notice that the sub-figures demonstrate the results using full metric (left) and partial metric (right). Given that the short-context law citation task and the law naming task provide relatively fewer contextual clues for the models, we observe that while models often accurately predict the law names, they tend to struggle more with the corresponding article or paragraph numbers. To address this, in addition to evaluating the models with the full metric, we also assess their performance using a partial metric, where predictions are considered accurate if the correct law name alone is identified. The recall and precision under this partial metric are presented in the right panel of Figure 1.

Finally, we randomly sampled 1,000 instances from each evaluation task to assess the performance of the state-of-the-art model, OpenAI's GPT-4o models (Achiam et al., 2023b). The generation method employs the batch API, with greedy decoding (temperature set to 0) and model specified to "GPT-4o-mini-2024-07-18". We use one-shot prompt design for GPT-4o-mini to understand the task better and produce the answer in the same format of LawTokens. The prompt example is provided in the Appendix.

The results are presented in Table 1. Overall, GPT-4o-mini does not perform at a level comparable to LawToken models. While we find that GPT-4o-mini is quite competitive when provided with ample contextual information, for example, in long-context law citation task, nearly matching the performance of the fine-tuned LawBase models, its effectiveness diminishes significantly in tasks with limited context, such as the short-context law citation task and the law naming task.

Model	Long		Short		Naming	
	R	P	R	P	R	P
LawTok-0.5B	0.54	0.65	0.46	**0.25**	0.08	0.02
LawTok-1.5B	**0.55**	**0.67**	0.44	0.22	0.08	0.02
LawTok-7.0B	0.53	0.65	**0.46**	0.22	**0.09**	**0.02**
LawBas-0.5B	0.23	0.33	0.06	0.03	0.02	0.01
LawBas-1.5B	0.31	0.42	0.20	0.11	0.05	0.01
LawBas-7.0B	0.21	0.30	0.18	0.09	0.05	0.01
GPT-4o-mini	0.28	0.41	0.03	0.02	0.01	0.01

Table 1: Comparison of recalls and precisions in different models in the 1000-dataset.

4 Examing model representations

While both model types show competitive results across the three legal tasks, LawToken consistently outperforms the LawBase models, with the only difference between the two being tokenization. This raises the question of what underlies this difference. On the one hand, the better performance of LawToken seems counterintuitive, as it uses fewer tokens to represent the legal mentions, thus fewer "buffering tokens" when decoding (Goyal et al., 2024; Herel and Mikolov, 2024). On the other hand, retrieving a legal mention is arguably distinct from reasoning; thus, LawToken may benefit from using an explicit, holistic token, allowing it to escape the complex structure within the legal mention comprising long compounds of act names and highly ambiguous article numbers.

In what follows, we investigate why the LawToken and LawBase models behave differently in the task. First, we demonstrate that the input embeddings learned by LawToken models reflect a general structure. Next, we examine the type-level representation similarities by comparing the model (hidden) states at different layer depths to the embeddings of the law's textual content. Finally, we analyze the token-level prediction probability as an index to how difficult the model finds certain tokens. These analyses provide further insight into the underpinning of the models' performance differences.

4.1 Input embeddings

Figure 2 shows the visualization of the law token's input embeddings of the top 3 common laws extracted from the LawToken model. Each point in the panel represents a law token; for example, arti-

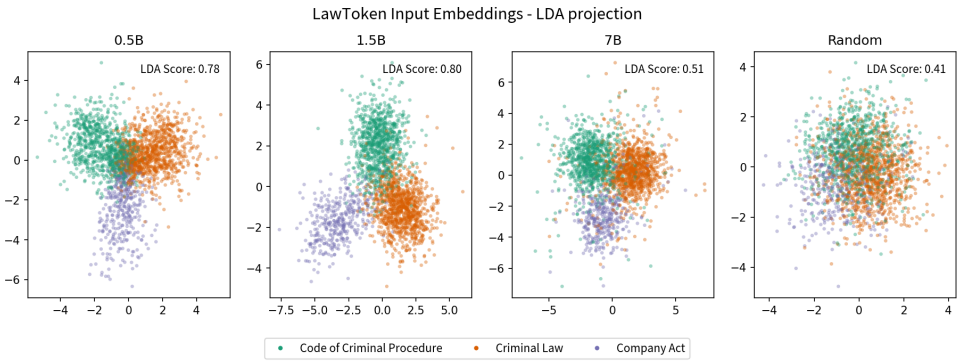


Figure 2: The input embeddings of the LawToken models, color-coded with the law article names: Code of Criminal Procedure (刑事訴訟法), Criminal Law (中華民國刑法), and Company Act (公司法). Only three laws are included for better visualization. The random Gaussian embeddings (Random) are shown as a baseline.

cles number 330 and 107 in the Code of Criminal Procedure are coded as two green dots. We use linear discriminant analysis to show how law tokens of different laws can be separable by a linear hyperplane. The underlying rationale is that law tokens coming from different laws should already reflect different usage patterns. Indeed, all classification accuracies are above the random chance level, while the 7B model is the worst of the three.

However, while classifying for law names is a simple and intuitive method to explore the embedding structure, it is not ideal. Law tokens of the same law may not necessarily be more similar than those of the different ones. To better gauge the semantic representation of the law tokens and the law references, we next examine the text embeddings of the legal text content.

4.2 Type-level representation similarity

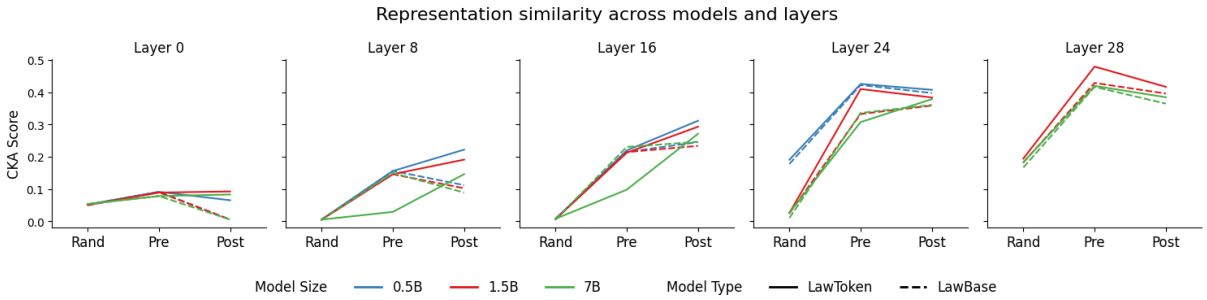
To better independently assess the quality of semantic representation encoded by the LawToken and LawBase models, we obtain the text embeddings of legal text content [4] with the commercial embedding models [5]. These embeddings are compared to the model's hidden states in various layer depths when encoding the selected sentences in the test split. A total of 13,215 sentences were selected, which included 2,211 unique legal citations. These sentences were selected to better evaluate the effect on the surrounding contexts, where there is only one law token or reference occurring before or after the 100-character window. We compute the centered kernel alignment scores (CKA; Kornblith et al., 2019) to measure the similarity between the model-encoded representation and the embedding of legal text content, where a higher score indicates a better correspondence between two representations.

However, caveats remain when using such text embeddings. The legal text content is the semantic extension of a legal citation – what it normatively refers to – whereas the model encodes how a legal token or reference is functionally used in the legal texts. They are inevitably different. In addition, LawToken and LawBase both encode the usage in the context, meaning that each law token occurrence induces a different model state, while the legal text embedding stays the same. Therefore, although we use legal text embeddings as a reference for semantic representation, they are only an operationalization of the law token's meaning.

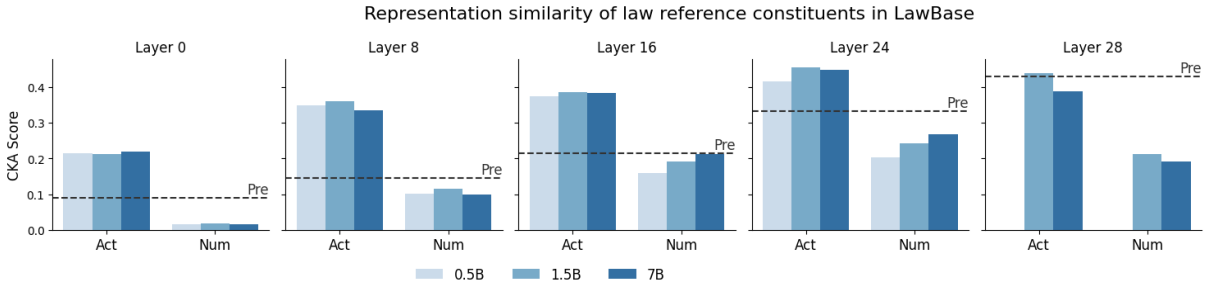
Figure 3a shows in each panel the results of representational similarities from the input layer (Layer 0) to the last layer of 0.5B model (Layer 24) or of 1.5B and 7B model (Layer 28). Each panel also shows three sites of interest. The `Rand` site denotes a random location before the target law token or reference, the `Pre` site is one token just before the target law token or the law reference, and the `Post` site is the token at the end of the target, which is the law token itself in the LawToken model and the last token of the law reference. Put in a more functional perspective, the `Rand` site provides a baseline estimate of the similarity possible to achieve only with the preceding context; the `Pre` site sheds light on the model states at which the model is about to predict the target law token or the first token in the law reference; and the `Post` site is

[4] For example, the text embedding for law token `<Labor Standards Act|43>` is the vector representation of the legal text content: "Workers may request leave for reasons such as marriage, [...]" (texts were in Taiwanese Mandarin.)

[5] Open AI's text-embedding-3-large



(a) Representation similarity scores across different sites. **Rand**: random location before the target law token or law reference; **Pre**: the token before the target; **Post**: the last token of the target, which is the law token itself and the last token of the law reference. Higher CKA scores indicate better alignment of the vectors with the law's semantics extensions.



(b) The representation similarities of the two constituents. **Act** refers to the name of the law article, and **Num** refers to the article number. As a visual reference, the dashed lines indicate the values of the Pre site of 1.5B LawBase model.

Figure 3: Representation similarities in different layers and different sites across model type and sizes.

when the models take into account of the law token or the law reference itself.

As shown in Figure 3a, the representation similarities increase throughout the layers and deeper into the sentence context. At the early layers of 0, 8, and 16, the `Rand` site scores are close to zero, reflecting there is only very local information at this stage, and they do not correlate well with the law semantics. In contrast, the `Pre` sites are more indicative of the law content, potentially because the immediate pre-context of the target law token and reference are already informative enough to the legal mentions. Interestingly, the `Post` sites start to show diverging patterns between the representation of LawToken and LawBase, where the scores from LawToken are consistently higher than those from LawBase. The pattern effectively demonstrates the effects of tokenizing legal mentions as a whole in the LawToken model, showing that the embeddings of the law tokens carry rich lexical information.

However, this advantage is not irreplaceable. As we move into the deeper layers of 24 and 28, the contextual effect is more pronounced. The diverging trends observed in the earlier layers are closing in on Layer 24, especially for the 0.5B model, which is the last layer, and on Layer 28, where all models' scores are similar. Nevertheless, in the last layers, the `Pre` sites have higher scores than the `Post` sites, which hints at three potential explanations: (1) the model's hidden states at `Pre` site should be the most indicative for the legal references, as they are ones used to generate final token logits. (2) The scores may inevitably decrease after the `Pre` site, as the models shift from focusing on the legal reference to predicting the subsequent context. (3) Alternatively, the drop may potentially be a consequence of the internal structure of the legal references.

To instantiate the impact of the internal structure of the legal reference, we compute their representation similarity scores on `Act` and `Num` sites. The `Act` and `Num` sites, applying only to the LawBase models, are two constituents in the law references: the former being the last token of the act name and the latter the last token of article numbering. Each panel clearly shows that while act name representations contribute more as we move from `Pre` site to `Act` site, especially in the early layers, the `Num` sites consistently reduce the scores. This suggests the numbering constituents of the law references are less informative than the article numbering or even the preceding context. In fact, incorporating the article numbering seems to negatively impact the representation of the law references.

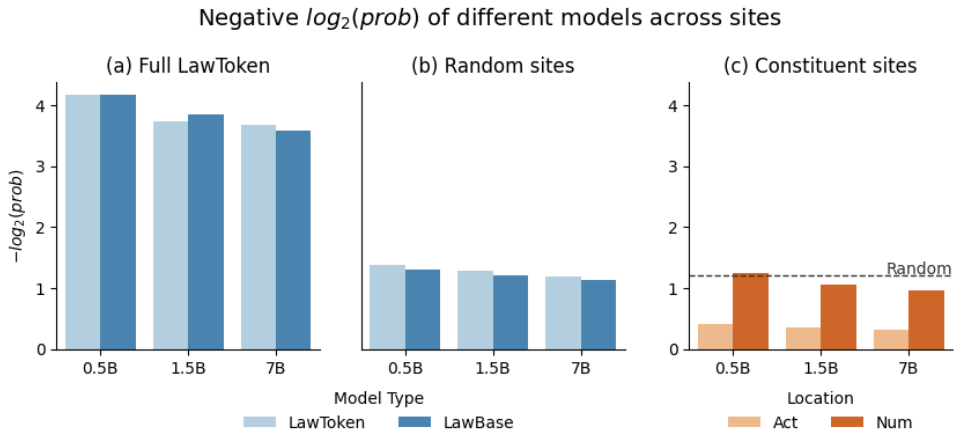


Figure 4: Negative $log_2(prob)$ of next-token predictions of different sites across models. (a) **Full LawToken** refers to the true law token and the multi-token sequences of the law reference. (b) The **Random sites** are the random locations before the target. (c) **Constituent sites** are the Act and Num sites. The dashed line is added as a visual reference, which are the values of 1.5B LawBase in the Random sites.

Representation similarities show the (mis-)alignments with law content semantics across different model layers and different sites, but they nevertheless only offer a coarse-grained view of the individual context each law token or reference is embedded. Being a context-independent measure of semantic extension, law content semantics is only based on the law content and has no access to the context information encoded by the LawBase or LawToken model. It is very well possible the misalignment we observed, for example, the reduced similarity scores of the Article numbering site, is because that the model has captured the context information that is not encoded in the static law content semantics. Therefore, we move to token-level probabilities to investigate the model behaviors further.

4.3 Token-level probability

The token-level probability provides complementary information for evaluating model behaviors. Distinct from the representation similarities where the token-based model states are compared to a type-based law content semantics, the prediction probabilities (of the true targets) are computed and evaluated in their context. There are two advantages of such a measure. (1) The prediction probabilities come directly from the model states of the hidden layer after accounting for all the other possible candidates. It effectively measures how good or close the last hidden states are to the true embeddings in that context. (2) The prediction probabilities also have explicit interpretations, which are surprisals as used in psycholinguistics studies (Goodkind and Bicknell, 2018; Wilcox et al., 2020), and information content or the compressed message length in bits if the law token or reference were to be compressed with an optimal compressor (Deletang et al., 2024; Tseng et al., 2024). That is to say, the prediction probabilities, particularly when transformed with a 2-based logarithm, signify the degree of difficulty the model has in predicting the law tokens or the law references based on the context it has encountered so far.

Figure 4 presents the results of prediction probabilities. Interestingly, despite the drastically different tokenization – where the law reference in LawBase has 11.90 tokens and only one in LawToken – their information contents (the $log_2(prob)$, summed over all tokens in law references) are largely the same across model sizes. However, this does not suggest intrinsic differences in decoding capacities between model types. As shown in Figure 4(b), LawBase models are not generally more efficient than the LawToken ones as the information contents remain comparable in the random sites where the predicted tokens occur before the law token. The findings are consistent with the previous representation similarities results, where the model states of the last hidden layers are almost the same in the *Pre* sites (except for the 1.5B model size, Figure 3a). Furthermore, this makes sense when considering the law token or reference conceptually: they are only two realizations of the same concept in input tokens, so both model types are expected to encode the law token or reference

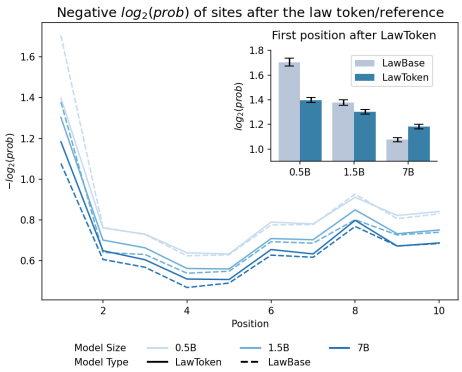


Figure 5: The negative $log_2(prob)$ of the sites after the target law token (LawToken) or reference (LawBase). The horizontal axis shows how many tokens are after the target. The inset highlights the first token after the target, where the LawToken models show higher predictability than the LawBase ones, except for the 7B model.

with similar information contents.

However, the similar information contents of the law tokens and references do not fully account for the observed differences in law citation tasks. As suggested by the previous model states findings, both LawBase and LawToken models achieve similar qualities of model states, as indicated by the CKA scores. It is only when LawBase models begin decoding token by token that the representation similarities decrease, especially at the article numbering sites. This pattern is consistently reflected in Figure 4(c). When comparing the Act name (`Act`) and article numbering sites (`Num`), the `Act` sites show very low information contents, significantly lower than the Random sites. In contrast, the `Num` site has higher values comparable to the Random ones. These token-level prediction probability results align with the type-level representation similarity findings: although LawToken models exhibit better lexical representation in the early layers, both models ultimately encode a similar amount of information through context. The key difference is that the LawBase models decode the law reference in multiple steps, and the best decoding representations are already achieved before the first token of law reference. Afterward, the LawBase models struggle with the highly ambiguous tokens from article numbers (`Num` sites), as evidenced by the reduced type-level representation similarities and the lower token-level information content.

Finally, Figure 5 presents the prediction probabilities following the law tokens and references. Neither the LawToken nor LawBase models show significant effects after the legal mentions, except the 0.5B and 1.5B models do show small but significant differences in the immediate token following position. This result is not surprising; as shown earlier, both model types encode comparable information content of legal mentions and can eventually compensate for the lexical information carried by the law token using context. Therefore, the holistic tokenization of law tokens only has a very limited effect on the following tokens.

5 Conclusion

Motivating by the form-meaning pairs of cognitive linguistics, we propose that the legal citations involving multi-word constituents can be processed not only as multi-token compounds but as holistic tokens. This paper empirically tests and investigates how different tokenizations affect model behaviors and representations. We train two model types: LawToken models, which consider the whole legal citation as one law token, and LawBase models, where the same citation is treated as multiple tokens. Our results show that LawToken models outperform LawBase models in legal citation tasks, particularly due to the article numbering component. We further analyze the model representations and find that both LawToken and LawBase models achieve comparable semantic representation quality. However, the LawBase model suffers from degraded representation in the multi-step decoding process, potentially increasing errors and hallucinations.

The implications of the present findings extend beyond linguistic theory. Indeed, the ability of LawToken models to encode what requires multiple tokens in LawBase ones already highlights that the form-meaning mappings can operate in a larger scope. Furthermore, in addition to better task performance, treating legal citations as law tokens has significant implications for future legal reasoning studies, particularly when examining potential circuits (Tigges et al., 2024; Prakash et al., 2024). Ultimately, while the model is likely to continue improving, understanding how it works – and ideally linking this back to our existing knowledge of language – is always an ongoing theme for computational linguistics.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,

Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023a. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023b. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Pegah Alipoor and Sabine Schulte im Walde. 2020. Variants of vector space reductions for predicting the compositionality of English noun compounds. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4379–4387, Marseille, France. European Language Resources Association.

Melanie J Bell and Martin Schäfer. 2016. Modelling semantic transparency. *Morphology*, 26:157–199.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Joan Bybee. 2010. *Language, usage and cognition*. Cambridge University Press.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and 1 others. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Gregoire Deletang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. 2024. Language modeling is compression. In *The Twelfth International Conference on Learning Representations*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Adele E. Goldberg. 2013. Constructionist Approaches. In *The Oxford Handbook of Construction Grammar*. Oxford University Press.

Adele E Goldberg. 2024. Usage-based constructionist approaches and large language models. *Constructions and Frames*.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. Think before you speak: Training language models with pause tokens. In *The Twelfth International Conference on Learning Representations*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Peter Henderson, Tatsunori Hashimoto, and Mark Lemley. 2023. Where's the liability in harmful ai speech? *J. Free Speech L.*, 3:589.

David Herel and Tomas Mikolov. 2024. Thinking tokens for language modeling. *arXiv preprint arXiv:2405.08644*.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR.

G. Lakoff. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press.

Jieh-Sheng Lee. 2023. Lexgpt 0.1: pre-trained gpt-j models with pile of law. In *Proceedings of the Seventeenth International Workshop on Juris-Informatics 2023 (JURISIN 2023)*, pages 15–24.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online. Association for Computational Linguistics.

Timothee Mickus, Denis Paperno, and Matthieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the First NLPL Workshop on Deep Learning for*

Natural Language Processing, pages 1–11, Turku, Finland. Linköping University Electronic Press.

Filip Miletić and Sabine Schulte im Walde. 2023. A systematic search for compound semantics in pretrained bert architectures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1499–1512.

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.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. 2024. From form (s) to meaning: Probing the semantic depths of language models using multisense consistency. *Computational Linguistics*, pages 1–51.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *The Twelfth International Conference on Learning Representations*.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th international joint conference on natural language processing*, pages 210–218.

Ian Rodgers, John Armour, and Mari Sako. 2023. How technology is (or is not) transforming law firms. *Annual Review of Law and Social Science*, 19(1):299–317.

Laura Eline Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. 2024. LLM circuit analyses are consistent across training and scale. In *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024)*, pages 290–303, Bangkok, Thailand. Association for Computational Linguistics.

Yu-Hsiang Tseng, Pin-Er Chen, Da-Chen Lian, and Shu-Kai Hsieh. 2024. The semantic relations in LLMs: An information-theoretic compression approach. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024*, pages 8–21, Torino, Italia. ELRA and ICCL.

Yu-Hsiang Tseng, Mao-Chang Ku, Wei-Ling Chen, Yu-Lin Chang, and Shu-Kai Hsieh. 2023. Vec2Gloss: definition modeling leveraging contextualized vectors with Wordnet gloss. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 679–690, Hong Kong, China. Association for Computational Linguistics.

Benjamin Weiser. 2023. Here's what happens when your lawyer uses chatgpt. *The New York Times*, 27.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, page 1707–1713.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*, pages 445–460. Springer.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Pienie Zwitserlood. 2014. The role of semantic transparency in the processing and representation of dutch compounds. In *Morphological Structure, Lexical Representation and Lexical Access (RLE Linguistics C: Applied Linguistics)*, pages 341–368. Routledge.