LawToken: a single token worth more than its constituents

Yu-Hsiang Tseng [1], Hsin-Yu Chou [2], Shu-Kai Hsieh [2]
[1]Department of Linguistics, University of Tübingen
[2] Graduate Institute of Linguistics, National Taiwan University

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., 2023), 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 performance 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 constituents. 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.

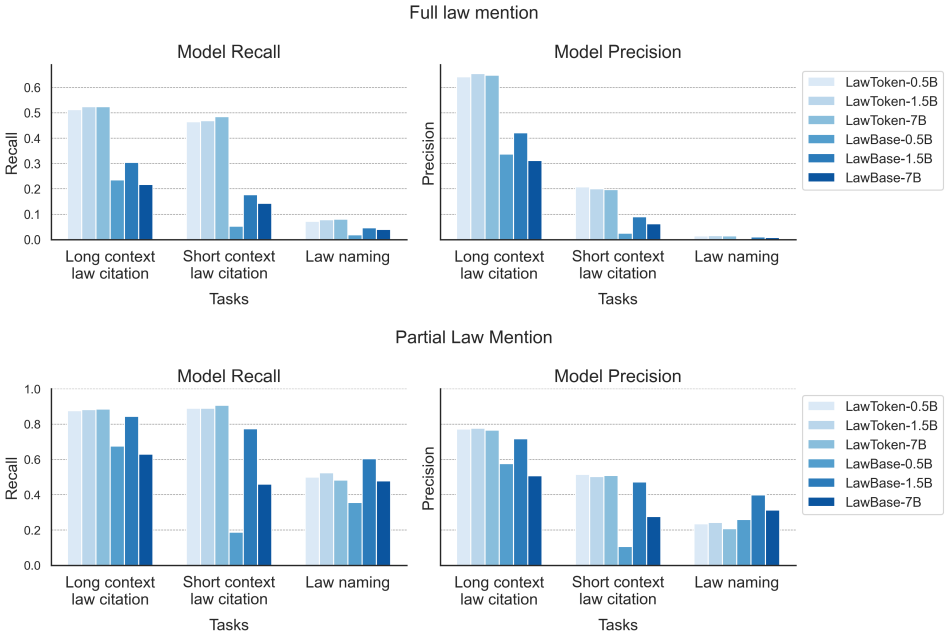


Figure 1: Evaluation results using full law mentions (upper panel) and partial law mentions (lower panel). The performances are evaluated using recall and precision, where recall is the proportion of correctly predicted law tokens among all true tokens, and precision is the proportion of correctly predicted tokens among all predicted ones.

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.

[1]https://huggingface.co/datasets/amy011872/LawToken.

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

[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/amy011872/LawToken-7B-a2.

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

Figure 1 presents the recall and precision of the six models across three different tasks. The upper panel indicates the measures evaluated using full law mentions, where a prediction is counted as correct only if both the law name and article number match the ground truth. The results show that the LawToken models consistently outperform the LawBase ones, regardless of the tasks and model sizes. These patterns may suggest that LawToken encodes better representations of law mentions, or simply sidesteps the challenge of predicting article number, which the LawBase model often struggles with. To investigate, we re-evaluate using partial law mentions, where the predictions are considered correct when the law name alone matches with the true ones. The results are shown in the lower panel of Figure 1. Again, LawToken still outperforms LawBase in most cases, although the performance gap narrows, especially with LawToken 1.5B, and in the law naming task. 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 fine-tuned explicitly in the current dataset.

The patterns in Figure 1 further reveal three notable observations: (1) Task difficulties vary with the richness of pragmatic context: the more context a task provides, the better the model performs. This effect is particularly evident in precision scores, where both models achieve the highest precisions in the long context task and the worst in the law naming task, where only the legal text content is available, with no additional pragmatic context. (2) Pragmatic context helps the LawBase model predict law names but not article numbers. This is shown in the partial law mention evaluations, where the LawBase's performance closes in on that of LawToken. This pattern is consistent with the fact that both LawBase and LawToken are trained on the same data, and the law names are lexical tokens that LawBase can learn their contextual usages during fine-tuning. By contrast, article numbers are highly ambiguous tokens reused across different law mentions and LawBase, having no specialized tokenization, struggles to disambiguate them. This is where LawToken has an advantage. (3) We also observed that 1.5B model size in the LawBase family is the best-performing one in both full and partial law mentions across the board. This suggests that, given the moderate size of our fine-tuning data (675M tokens), 1.5B may represent the optimal model size under data constraints, assuming no changes to the tokenization.

Finally, to further compare the task performance of LawToken to other models, we randomly sampled 1,000 instances from each evaluation task to assess the performance of one of the commercial models (Achiam et al., 2023) (GPT-4o-mini). 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 the 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. The compar-

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.

ison crucially demonstrates that the tasks cannot be solved solely by superficial textual cues included in the context, which the GPT-4o models will take advantage of.

Taken together, these results show LawToken models consistently outperform LawBase models. Moreover, the comparison between full and partial law mention evaluations suggests the crucial differences stem from how the model handles law names versus article numbers. To better understand the model representations of the law tokens and their law names and article number constituents, we next examine the representational differences between LawToken and LawBase models.

4 Examining 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 of how difficult the model finds certain tokens. These analyses provide further insight into the underpinnings of the models' performance differences.

4.1 Input embeddings

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

[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

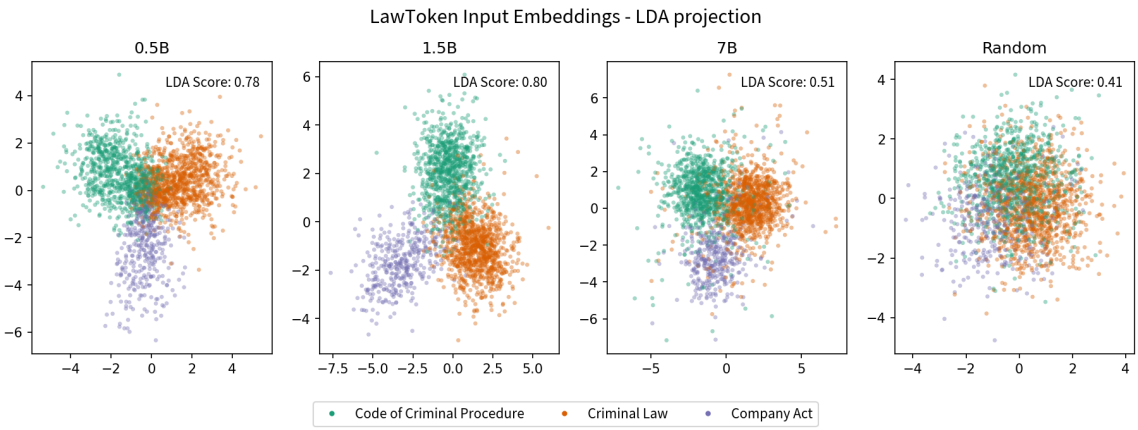


Figure 2: The input embeddings of the LawToken models, color-coded with the law article names: Code of Criminal Procedure (刑事訴訟法), Criminal Law (中華民國刑法), and Company Act (公司法). Only three laws are included for better visualization. The random Gaussian embeddings (Random) are shown as a baseline.

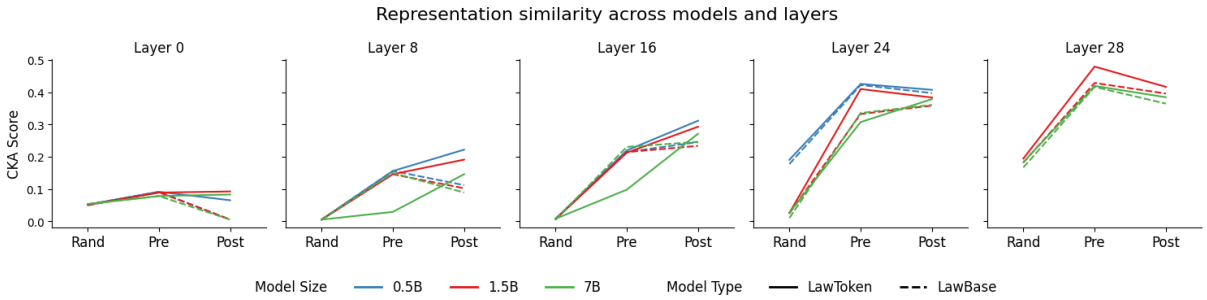
a better correspondence between two representations.

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.

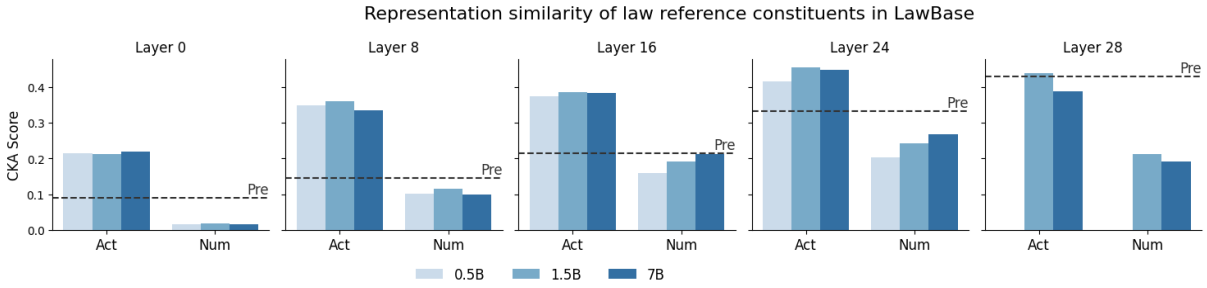
Figure 3a shows in each panel the results of representational similarities from the input layer (Layer 0) to the last layer of 0.5B model (Layer 24) or of 1.5B and 7B model (Layer 28). Each panel also shows three sites of interest. The `Rand` site denotes a random location before the target law token 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 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 that 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 is already informative enough about 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



(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.

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.

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 that the misalignment we observed, for example, the reduced similarity scores of the Article numbering site, is because 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's behavior 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

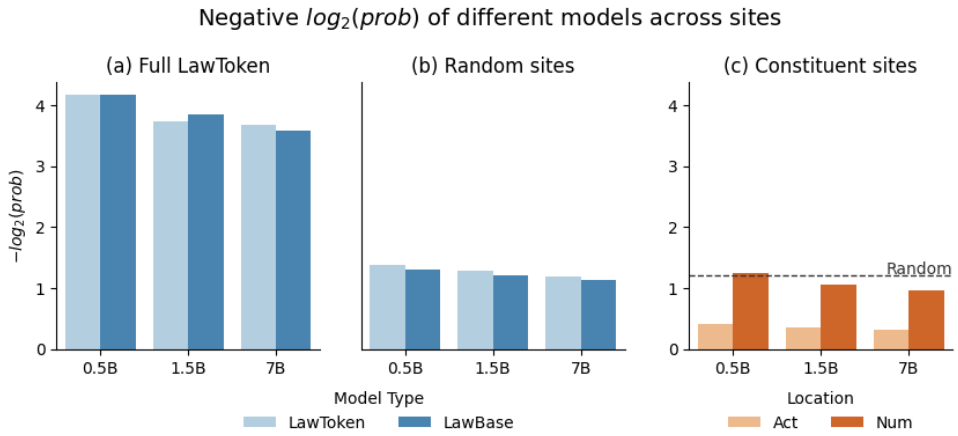


Figure 4: Negative $log_2(prob)$ of next-token predictions of different sites across models. (a) **Full LawToken** refers to the true law token and the multi-token sequences of the law reference. (b) The **Random sites** are the random locations before the target. (c) **Constituent sites** are the Act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

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 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 that the 0.5B and 1.5B models do show small but significant differences in the immediate token fol-

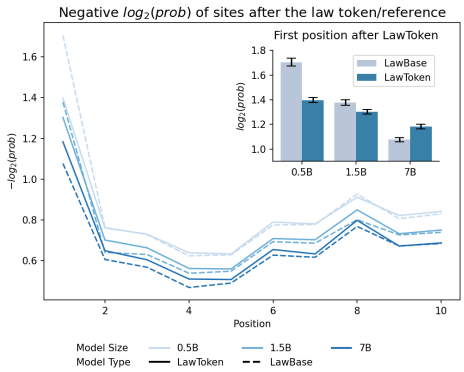


Figure 5: The negative $log_2(prob)$ of the sites after the target law token (LawToken) or reference (LawBase). The horizontal axis shows how many tokens are after the target. The inset highlights the first token after the target, where the LawToken models show higher predictability than the LawBase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

lowing 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

Motivated 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.

It may seem counterintuitive that treating an entire legal mention as a holistic law token improves task performance instead of leading to overfitting. However, this becomes understandable when we consider the compositionality problem inherent in the legal mention. In the mention, the article number component is the least informative constituent in a compound: it is constantly reused, lacks intrinsic connection to the intended meaning, and can only be resolved by context. The fact that LawBase models can achieve higher performance through fine-tuning, yet still fall short of LawToken models, suggests an upper bound to what contextualization alone can achieve. Beyond that, the model may need a more efficient or more discriminative cue, i.e., a law token in this case, to link with the intended semantics. In this sense, the model either considers the legal mention as a single "word" or compress a compound as a token, depending on one's definition of "word." Regardless, this line of reasoning align with linguistic models that do not assume the compositionality processing of compounds or a fixed and static notion of words (Baayen et al., 2019; Libben, 2022). Moreover, while this study shows that a manually defined law token is beneficial, whether such tokens can be learned dynamically (Pagnoni et al., 2025) remains an open question.

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, 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). Linguistic theories may not directly inform the development of LLMs. Instead, the growing use of LLMs now makes it possible for linguists to empirically test theoretical claims that were previously out of reach. When combined with such implementations, linguistic theories can begin to move toward "an integrated model that generates precise quantitative predictions for vast arrays of empirical findings" (Baayen, 2024), opening new pathways that connect LLMs with our existing knowledge of language.

Acknowledgments

This study is supported by the ERC Horizon Europe Programme, ERC-2021-ADG, No. 101054902 (SUBLIMINAL). We also thank NVIDIA-NTU Artificial Intelligence Joint Research Center for providing computational resources.

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. 2023. 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.

R Harald Baayen. 2024. The wompom. *Corpus Linguistics and Linguistic Theory*, 20(3):615–648.

R Harald Baayen, Yu-Ying Chuang, Elnaz Shafaei-Bajestan, and James P Blevins. 2019. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity*, 2019(1):4895891.

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.

Gary Libben. 2022. From lexicon to flexicon: The principles of morphological transcendence and lexical superstates in the characterization of words in the mind. *Frontiers in Artificial Intelligence*, 4:788430.

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.

Artidoro Pagnoni, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, and 1 others. 2025. Byte latent transformer: Patches scale better than tokens. In *ICML 2025 Workshop on Tokenization and Beyond (TokShop)*, Vancouver Convention Center, Vancouver, BC, Canada. Extended version available at arXiv:2412.09871.

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.

Appendix

A Limitation

This paper examines how tokenization impacts model performance in predicting legal citations and shaping semantic representations, using Taiwan's legal citation system as our dataset. However, we acknowledge that the citation formats vary across different countries, especially considering that Taiwan follows the civil law system, in contrast to the common law system, as in the British or the United States. Although we believe the findings are relevant to other surface forms, this work remains constrained by the dataset upon which it is trained and tested. On the theoretic side, our results indicate that the multi-token models (i.e., LawBase) suffer from degraded representations during multi-step decoding. Yet, it remains unclear whether the degradation stems from the nature of legal citation or can be generalized to a more general form-meaning mapping problem, such as those found in compounds or multi-word expressions. Addressing these questions requires more experiments and analyses in future studies.

B Examples of training data

The dataset comprises laws and verdicts in Taiwan. The examples of training data shown below, other than the "Question", "Answer" and the law citation tokens, are all in traditional Chinese. Personal names are anonymized, although they appear in the original verdicts. English translations are provided for clarity but are never seen by the model.

B.1 Example 1

理 由 一、本件原裁定以抗告人陳○○因不服臺灣新北地方檢察署 101年度執更丑字第4313號執行指揮書而聲明異議，經原審以10 7年度聲字第544號裁定駁回，並囑託法務部矯正署宜蘭監獄長官於民國107 年4月2日向抗告人合法送達，此有送達證書附卷可查。其抗告期間之末日為同年月7 日星期六，翌日為星期日，均為休息日。其提起抗告，僅可於休息日次日即同年月9 日星期一為之。乃竟遲至同年月11日始向法務部矯正署宜蘭監獄長官提起抗告，有抗告人所提刑事抗告狀在卷可證。已逾5日抗告期間，因依<刑事訴訟法|411>前段規定駁回其抗告。經核尚無不合。 二、抗告意旨徒以107年4月4日至同年月8日為休假期日，依社會通念休假期日不計算期日，同年月12日才是抗告終止日等語，係憑己見指摘原裁定不當。其抗告為無理由，應予駁回。 據上論結，應依<刑事訴訟法|412>，裁定如主文。

English Translation

Reasoning

1. The original ruling was based on the fact that the appellant, Chen xx-xxx, objected to the execution order No. 4313 (Year 101, Re-Execution Chou Character) issued by the Taiwan New Taipei District Prosecutors Office, and filed an objection accordingly. The original court dismissed the objection in Ruling No. 544 (Year 107, Objection Character), and entrusted the Yilan Prison Warden of the Agency of Corrections, Ministry of Justice to serve the ruling lawfully to the appellant on April 2, 2018 (Year 107 of the Republic of China calendar). This is confirmed by the certificate of service included in the case file.

The last day of the appeal period was Saturday, April 7 of the same year, and the next day, Sunday, was also a rest day. Therefore, an appeal could only be filed on the next business day, which was Monday, April 9 of the same year. However, the appellant did not file the appeal until April 11 of the same year, submitting it to the Yilan Prison Warden. This is proven by the criminal appeal document submitted by the appellant on record. Since this was beyond the 5-day appeal period, the appeal is dismissed according to the first part of Article 411 of the Code of Criminal Procedure. Upon review, this decision is deemed proper.

2. The grounds for appeal merely argue that the period from April 4 to April 8, 2018, was a holiday, and that under common social understanding, holidays are not counted toward deadlines, hence April 12 should be considered the last day to appeal. This is a subjective interpretation and an unfounded criticism of the original ruling. The appeal lacks merit and should be dismissed.

In conclusion, pursuant to Article 412 of the Code of Criminal Procedure, the ruling is made as stated in the main text.

B.2 Example 2

<土地法|46-2>重新實施地籍測量時，土地所有權人應於地政機關通知之限期內，自行設立界標，並到場指界。逾期不設立界標或到場指界者，得依左列順序逕行施測： 一、鄰地界址。 二、現使用人之指界。 三、參照舊地籍圖。 四、地方習慣。 土地所有權人因設立界標或到場指界發生界址爭議時，準用第五十九

條第二項規定處理之。

English Translation

<Land Act|46-2> When a cadastral resurvey is being conducted, the landowner shall, within the deadline specified in the notice issued by the land administration authority, install boundary markers and appear on-site to indicate the boundaries. If the landowner fails to install boundary markers or appear on-site within the prescribed period, the survey may proceed directly according to the following order of priority: (1) The boundaries of adjacent parcels. (2) The boundary indications provided by the current user of the land. (3) Reference to the old cadastral maps. (4) Local customs. If a boundary dispute arises due to the installation of boundary markers or the on-site boundary indication by the landowner, the provisions of Paragraph 2, Article 59 shall apply mutatis mutandis.

B.3 Example 3

原審以：被上訴人主張上訴人為系爭支票發票人，伊為執票人等情，為上訴人所不爭執，且有系爭支票影本可稽，堪信為真實。 上訴人抗辯：鄭○○詐稱呂○○對其負有債務，且將來會負責支付系爭支票票款，要求開立系爭支票與被上訴人，但實際上呂○○未積欠鄭○○錢，呂○○被鄭○○及被上訴人詐欺，陷於錯誤交付系爭支票等語，可知上訴人因認呂○○對鄭○○負有債務始簽發系爭支票，嗣因呂○○與鄭○○間發生債務糾葛，呂○○始否認對鄭○○負有債務，此由鄭○○於上訴人簽發系爭支票後，嗣後另案起訴請求呂○○返還投資款即明，復有民事起訴狀影本可參，況上訴人自始未提出任何證據佐證其被詐欺或脅迫而簽發系爭支票，上訴人此部分抗辯，不足為採。上訴人另抗辯系爭支票之原因關係不存在，惟票據係文義證券及無因證券，屬不要因行為，故執票人祇須就該票據作成之真實負證明之責，關於票據給付之原因，並不負證明之責任，票據債務人仍應就其抗辯之原因事由，先負舉證責任。然上訴人未就其抗辯事由負舉證責任，則被上訴人請求上訴人給付系爭支票款1100萬元，及自105年8月1日起至清償日，按週年利率6%計算之利息，為有理由，應予准許等詞，因而維持第一審所為上訴人敗訴之判決，駁回其上訴，經核於法並無不合。 按票據乃文義證券及無因證券，票據上之權利義務悉依票上所載文義定之，與其基礎之原因關係各自獨立，票據上權利之行使不以其原因關係存在為前提。是執票人行使票據上權利時，就其基礎之原因關係確係有效存在不負舉證責任。僅於票據債務人以自己與執票人間所存抗辯事由對抗執票人，而該票據基礎之原因關係經確立者，法院就此項原因關係進行實體審理時，當事人於該原因關係是否有效成立或已否消滅等事項有所爭執，始應適用各該法律關係之舉證責任分配原則。查上訴人為系爭支票發票人，被上訴人為執票人，為原審所確定。被上訴人主張上訴人係為返還伊投資款而簽發系爭支票，上訴人則抗辯係鄭○○詐稱呂○○對其負有債務，且將來會負責支付系爭支票票款，而簽發交付系爭支票予被上訴人，就被上訴人取得系爭支票之原因關係，各執一詞，並未確立，依上說明，仍應由上訴人就其抗辯之原因關係，負舉證之責。原審因上訴人未舉證證明系爭支票之原因關係，而為其不利之認定，自不違背舉證責任分配原則。至上訴人援引之本院判決，或係就該票據基礎之原因關係經確立情形所為之闡述，或與本件事實有所差異，均無從比附援引。上訴論旨，指摘原判決不當，聲明廢棄，非有理由。 據上論結，本件上訴為無理由。依、、、，判決如主文。<cite><民事訴訟法|436-2|2>,<民事訴訟法|78>,<民事訴訟法|449|1>,<民事訴訟法|481></cite>

English Translation

The court of first instance found that: the appellee asserted that the appellant was the issuer of the check in dispute, and that the appellee was the holder of said check—facts not contested by the appellant and supported by a copy of the disputed check, which is deemed credible and authentic.

The appellant contended that Cheng xxx-xxx falsely claimed that Lu xxx-xxx was indebted to him and would be responsible for the payment of the disputed check, and thus requested the issuance of the check jointly with the appellee. However, in fact, Lu xxx-xxx owed no debt to Cheng xxx-xxx, and the check was delivered under a mistake caused by the fraud committed by Cheng xxx-xxx and the appellee. From this, it is clear that the appellant issued the check under the belief that Lu xxx-xxx owed Cheng xxx-xxx a debt. After a dispute arose between Lu xxx-xxx and Cheng xxx-xxx regarding said debt, Lu xxx-xxx denied owing any such debt. This is evident from the fact that, after the appellant issued the check, Cheng xxx-xxx filed a separate lawsuit seeking return of his investment from Lu xxx-xxx; a copy of that civil complaint is also on record. Moreover, the appellant never submitted any evidence to support the claim of having

been defrauded or coerced into issuing the check. Therefore, this part of the appellant's defense lacks merit.

The appellant further argued that there was no underlying transaction or cause for the issuance of the check in dispute. However, as a negotiable instrument, a check is a documentary and abstract security—its legal force derives from the wording on the instrument itself and is independent of the underlying cause. Accordingly, the holder of the check only bears the burden of proof with respect to the authenticity of the check itself, and not regarding the underlying cause of payment. On the contrary, it is the debtor on the check who must bear the burden of proof for any defenses raised against payment. Since the appellant failed to provide proof supporting the grounds for their defense, the appellee's claim for payment of NT$11 million, along with interest calculated at an annual rate of 6% from August 1, 2016 until the date of repayment, is well-founded and should be granted. Therefore, the judgment of the court of first instance, which ruled against the appellant, is upheld, and the appeal is dismissed. Upon review, this judgment is in accordance with the law.

According to law, negotiable instruments are documentary and abstract in nature. The rights and obligations indicated on the face of the instrument govern, independently of any underlying transaction. The exercise of rights under a negotiable instrument does not require proof of the existence of the underlying relationship. Thus, when a holder of an instrument seeks to exercise such rights, they bear no burden of proof regarding the validity of the underlying relationship. Only when the debtor on the instrument raises a defense based on their own relationship with the holder—and the underlying cause of the instrument is thereby established—does the court proceed to substantively examine that cause. In such cases, the burden of proof is allocated according to the relevant substantive legal relationships.

The appellant is confirmed to be the issuer of the check in dispute, and the appellee its holder, as determined by the lower court. The appellee claims the check was issued by the appellant to repay an investment, while the appellant claims the check was issued under the false impression—due to misrepresentation by Cheng xxx-xxx—that Lu xxx-xxx owed Cheng a debt and would pay the amount. Each party presents a different version of the reason behind the check's issuance, and no cause has been established. According to the principles stated above, it remains the appellant's responsibility to prove their asserted cause. Since the appellant failed to meet that burden, the lower court's unfavorable ruling does not violate the principle of burden of proof allocation.

As for the judgments cited by the appellant, those either concern cases where the underlying cause of the negotiable instrument was established, or differ in facts from the present case, and are therefore inapplicable. The grounds of appeal, which challenge the lower judgment as improper and request its reversal, are without merit.

In conclusion, the appeal in this case is groundless. Pursuant to,,, judgment is rendered as stated in the main text. <cite><Civil Procedure Code|436-2|2>,<Civil Procedure Code|78>,<Civil Procedure Code|449|1>,<Civil Procedure Code|481></cite>

C Examples of evaluation tasks

C.1 Example of long-context citation task

Question: ``四、原審已依吳○○就醫之相關病歷資料、診斷證明書、臺中榮民總醫院函文、勞動部勞工保險局函文等資料，載敘吳○○傷勢及結果甚詳，上訴人及其辯護人於審理中並未爭執有何記載錯誤、不實之處，則原審綜合全案證據資料，依其所採取之證據及得心證理由之說明，已足以認定吳○○受有右眼創傷性黃斑部裂孔造成僅能辨識眼前指數10公分，且右眼視野缺損、最佳矯正視力為0.01，達一目視能嚴重減損之重傷害，而未再為其他無益之調查，自無上訴意旨所指適用法則不當、調查未盡之違法情形可言。又本院為法律審，不為事實之調查，上訴人上訴於本院，始提出其蒐得吳○○工作之照片作為新證據資料，執以指摘原判決違誤，亦非上訴第三審之合法理由。 五、綜合前旨及其他上訴意旨，無非係置原判決所為明白論斷於不顧，仍持已為原判決指駁之陳詞再事爭辯，或對於事實審法院取捨證據與自由判斷證據證明力之職權行使，徒以自己之說詞，為相異評價，任意指為違法，或單純為事實上枝節性之爭辯，要與法律規定得為第三審上訴理由之違法情形，不相適合。本件上訴違背法律上之程式，應予駁回。 據上論結，應依刑事訴訟法前段，判決如主文。<cite>''

Answer: ``<刑事訴訟法|395>,<刑事訴訟法|377>''

English translation

Question: 4. The original trial court had already

reviewed relevant medical records, diagnostic certificates, correspondence from Taichung Veterans General Hospital, and documents from the Bureau of Labor Insurance of the Ministry of Labor. These materials provided a detailed description of Wu xxx-xxx's injuries and medical outcomes. During the proceedings, neither the appellant nor their defense counsel disputed any inaccuracies or falsehoods in those records. Therefore, the trial court, based on the totality of the evidence and its reasoning for the credibility of the accepted proof, was fully justified in concluding that Wu xxx-xxx sustained a traumatic macular hole in his right eye, rendering him able to perceive only hand motion at 10 cm in front of the eye. He also suffers from a loss of visual field and a best-corrected visual acuity of 0.01 in that eye—constituting a serious injury causing severe impairment to monocular vision. As such, the court did not engage in further unnecessary investigation, and there is no indication of improper application of the law or failure to investigate, as alleged in the appeal.

Furthermore, this Court serves as a court of law, not of fact. The appellant's submission of photographs allegedly showing Wu xxx-xxx at work—presented for the first time on appeal to this Court as new evidence and cited as grounds to challenge the lower court's decision—does not constitute a legitimate reason for a third-instance appeal.

In sum, the foregoing and the rest of the appeal merely disregard the clear reasoning of the original judgment, reasserting arguments already addressed and rejected by the lower court, or challenge the trial court's discretion in evaluating and weighing evidence by offering alternative interpretations based on the appellant's own narrative. Such arguments are factual disputes over minor points and do not qualify as legal grounds for a third-instance appeal under the law. This appeal thus violates procedural requirements and shall be dismissed.

Based on the above reasoning, and pursuant to the first part of the Code of Criminal Procedure, judgment is rendered as stated in the main text. <cite>

Answer: <Code of Criminal Procedure|395>, <Code of Criminal Procedure|377>

C.2 Example of short-context citation task

Question: ``按當事人因無資力支出訴訟費用而聲請訴訟救助者，關於無資力支出訴訟費用之事由，應提出可使法院信其主張為真實並能即時調查之證據，以釋明之。此觀之規定自明。<cite>''

Answer: ``<民事訴訟法|109|2>''

English translation According to the law, when a party applies for litigation aid on the grounds of inability to afford litigation costs, they must provide evidence that is sufficient to convince the court of the truthfulness of their claim and that can be promptly verified by the court, in order to clarify the grounds for their financial inability. This is clearly stipulated by law. <cite>

Answer: <Civil Procedure Code|109|2>

C.3 Example of law-naming task

Question: ``物之發明之實施，指製造、為販賣之要約、販賣、使用或為上述目的而進口該物之行為。<cite>''

Answer: ``<專利法|58|2>''

English translation
Question: The implementation of an invention of a product refers to the acts of manufacturing, offering for sale, selling, using, or importing the product for the above purposes. <cite>

Answer: "<Patent Act|58|2>"

D Prompt design for GPT-4o-mini

D.1 System prompt

"你是一名熟悉中華民國法條的法律專業人士，在任何情境下，你都能援引最適切的法條予以回應。"

English translation
You are a legal professional well-versed in the laws of the Republic of China (Taiwan), and in any situation, you are able to cite the most appropriate legal provisions in your response.

D.2 Prompt template

```
---

<判決書>
[....]
<cite>
<法條>
<刑事訴訟法|449|1>,<刑事訴訟法|449|3>,<毒品危害防制條例|20>,<毒品危害防制條例|23|2>,<刑事訴訟法|454|1>,<毒品危害防制條例|23>,<毒品危害防制條例|10|2>
```

<判決書>
{Question}
<cite>
<法條>
"

English translation

Your task is to identify the most relevant legal provisions. First, refer to the judgments and their associated legal articles in the examples below. Then, a second judgment will be presented—this is your task. Based on the content and subject of that judgment, please provide the applicable legal provisions in the specified format and return them as a JSON file.

<verdict>
[....]
<cite>
<laws>
<Code of Criminal Procedure|449|1>,<Code of Criminal Procedure|449|3>,<Narcotics Hazard Prevention Act|20>,<Narcotics Hazard Prevention Act|23|2>,<Code of Criminal Procedure|454|1>,<Narcotics Hazard Prevention Act|23>,<Narcotics Hazard Prevention Act|10|2>

<verdict>

{Question}
<cite>
<laws>