Jurnal Komputer, Informasi dan Teknologi Vol: 5, No 1, 2025, Page: 1-9

Artificial Intelligence in Tax Law: Case Classification and Topic Modelling Using Large Language Models

Faisal Labib Zulfiqar[1]*, Ayu Rosalia[2]

1 Ministry of Finance, Indonesia
2 Universitas Mercubuana, Indonesia

DOI: https://doi.org/10.53697/jkomitek.v5i1.2815
*Correspondence: Faisal Labib Zulfiqar
Email: faisal.labib@gmail.com

Received: 16-06-2025
Accepted: 23-06-2025
Published: 30-06-2025

Abstract: This study explored the application of the GPT API framework to automate case complexity classification and topic modelling in Indonesian Tax Court decisions. Using a dataset of 5,000 anonymized tax dispute summaries, we designed a prompt-based classification pipeline supported by expert-labelled benchmarks. The case complexity classification task, categorizing cases into low and high complexity. The GPT-based model achieving 87% precision. This indicates the model's practical ability to simulate legal judgment in triaging case difficulty. Simultaneously, topic modelling was performed to identify key dispute themes across grouped cases. The three most frequently recurring themes were: (1) input VAT correction errors in VAT disputes (26.1%), (2) net income adjustments in income tax cases (9.1%), and (3) customs valuation issues in import transactions (17.4%). These model-derived clusters aligned closely with expert taxonomies and provided useful summaries of dispute patterns over time. The methodology built using Python, Google Colab, and the OpenAI GPT API. By structuring Indonesia's growing corpus of tax litigation into actionable categories, this approach strengthens the country's digital justice transformation. It enables better resource allocation and faster dispute resolution.

Keywords: Tax Dispute Analytics, Large Language Models (LLM), Case Complexity Classification, Topic Modelling, Judicial Decision Support

Introduction

Indonesia's tax court system plays a critical role in maintaining the integrity of national revenue and upholding taxpayer rights. With more than 13,000 tax disputes processed annually, the burden on administrative and judicial resources continues to grow. These disputes range widely in complexity, monetary value, and legal nuance. It poses significant challenges for timely and consistent resolution (Ding, 2024). Amid ongoing digital transformation efforts, such as the implementation of the e-Tax Court system, there is a growing urgency to enhance efficiency, transparency, and accountability in tax dispute management. As Indonesia strengthens its commitment to digital governance, the integration of advanced technologies such as artificial intelligence (AI) and natural language processing (NLP) has emerged as a promising pathway to support judicial decision-making.

This policy will improve institutional responsiveness and reinforce legal certainty across the fiscal justice system (Andriati et al., 2024).

The sheer volume and complexity of tax dispute cases often lead to prolonged litigation and inconsistent judgments. It can erode public trust in the legal process and hinder state revenue collection. Many of these cases involve recurring themes, such as transfer pricing, input-output VAT, or formal versus material defects, that, while differing in details, follow similar legal trajectories (Zulfiqar et al., 2023). However, the absence of systematic case classification and the lack of data-driven support tools for judges and administrators impede efforts to streamline case handling (Afiyati et al., 2022). Existing digital infrastructure has yet to fully harness the potential of AI to assist in identifying patterns and forecasting case duration. As a result, judicial processes remain heavily reliant on manual review, subject to human variability, and limited in scalability (Elliot & Thomas, 2020). This context underscores the need for a more intelligent, responsive system that augments human judgment with machine learning capabilities.

Despite significant progress in digitizing court processes, such as electronic filing and digital document management, the Indonesian Tax Court still struggles to manage incoming caseloads effectively. For example, in 2024 alone, the court received 11,835 new dispute cases and issued 17,200 verdicts. It indicates a substantial inflow that challenges current workflow capacities. This dataset of 5,000 (2008 to 2020) metadata-rich dispute records reflects this complexity and volume, offering a valuable testing ground for AI-driven processing. By integrating machine learning tools into existing systems, there is the potential to automatically detect high-complexity cases, assist in digital docket prioritization, and reroute less demanding cases for streamlined adjudication (Sari et al., 2024). Such an approach has the potential to enhance the court's responsiveness, reduce administrative bottlenecks, and support systematic improvements in judicial operations without overhauling foundational legal procedures (Han et al., 2024).

Recent advancements in artificial intelligence (AI) and natural language processing (NLP) have significantly improved legal text analysis, particularly in the classification and prediction of legal cases. Transformer-based models like BERT and its variants are widely recognized for their superior performance in handling complex legal documents. Imran et al. (2023) demonstrated that RoBERTa outperformed other transformer models in classifying European Court of Human Rights (ECHR) cases, achieving an F1-score of 86.7%, thereby emphasizing the effectiveness of transformer models for legal classification tasks. Similarly, Costa et al. (2025) proposed a two-stage deep learning approach that combined BERT and BiLSTM encoders with an SVM classifier to automate petition classification in Brazil's judicial system, achieving impressive accuracy and real-world application. These findings are supported by Siino et al. (2025), who reviewed 61 studies and found that large

language models (LLMs) like GPT and BERT were consistently applied in tasks such as legal case classification and contract analysis due to their contextual understanding and scalability.

Complementing these developments, researchers have also explored hybrid and interpretable machine learning approaches. Sukanya & Priyadarshini (2024) introduced a CNN-transformer hybrid model that surpassed traditional classifiers in predicting Indian legal judgments, while Budhiraja (2024) presented an ensemble learning framework combining semantic networks and models like SVM and CNN, achieving over 90% accuracy in legal outcome predictions. Dokumacı (2024) added an econometric dimension by integrating AI with pattern recognition models to support legal decision-making using statistical and NLP tools.

Despite the growing application of transformer-based and hybrid models in legal analytics, most existing studies are concentrated in high-resource jurisdictions such as Europe, India, and Brazil. Exploration in the low to middle income countries like Indonesia is limited. However, legal texts are diverse, inconsistently structured, and underrepresented in global datasets. Furthermore, prior research has largely focused on judgment prediction or legal classification using structured data, whereas the challenge of interpreting unstructured tax dispute summaries remains underexplored. This study addresses these gaps by leveraging the GPT API to classify case complexity and extract thematic insights from Indonesian Tax Court decisions. By combining LLM-generated embeddings and integrating them into traditional machine learning workflows, this research introduces a scalable and context-aware method tailored to the nuances of Indonesia's tax litigation domain.

Methodology

This study used a natural language processing (NLP) approach to analyse unstructured legal texts from Indonesian Tax Court decisions. The research focused on two main tasks: classifying case complexity and identifying common dispute themes.

The dataset consisted of anonymized tax dispute summaries. Basic text pre-processing was applied to clean and standardize the data before analysis. For the complexity classification task, a few-shot prompting strategy was used directly with the OpenAI API. The prompts were designed to simulate how a legal assistant would interpret and classify the complexity level of each dispute narrative. To evaluate the model's classification performance, 150 cases were manually labelled based on expert judgment from practitioners with experience in tax litigation. These expert-labelled cases served as a benchmark for assessing the accuracy and consistency of the model's predictions (Zhang et al., 2022).

For the topic modelling task, the GPT model was prompted to extract and summarize recurring themes from grouped dispute descriptions. This approach replaced traditional topic modelling methods and allowed the model to identify legal themes based on semantic context, without relying on predefined taxonomies. All steps were performed using Python in a Google Colab environment. The model was accessed via API calls without any additional training or fine-tuning.

Results and Discussion

From the 5,000 tax dispute from 2008 to 2018 analysed, the model classified 4,477 cases (89.5%) as high complexity and 523 (10.5%) as low complexity. This strong skew toward high-complexity cases highlights the intricate nature of disputes that reach litigation in Indonesia. By successfully interpreting unstructured descriptions of dispute content, the GPT-based model demonstrated its ability to simulate human-like legal reasoning at scale. It's an innovation particularly relevant for jurisdictions with large case volumes and limited expert resources.

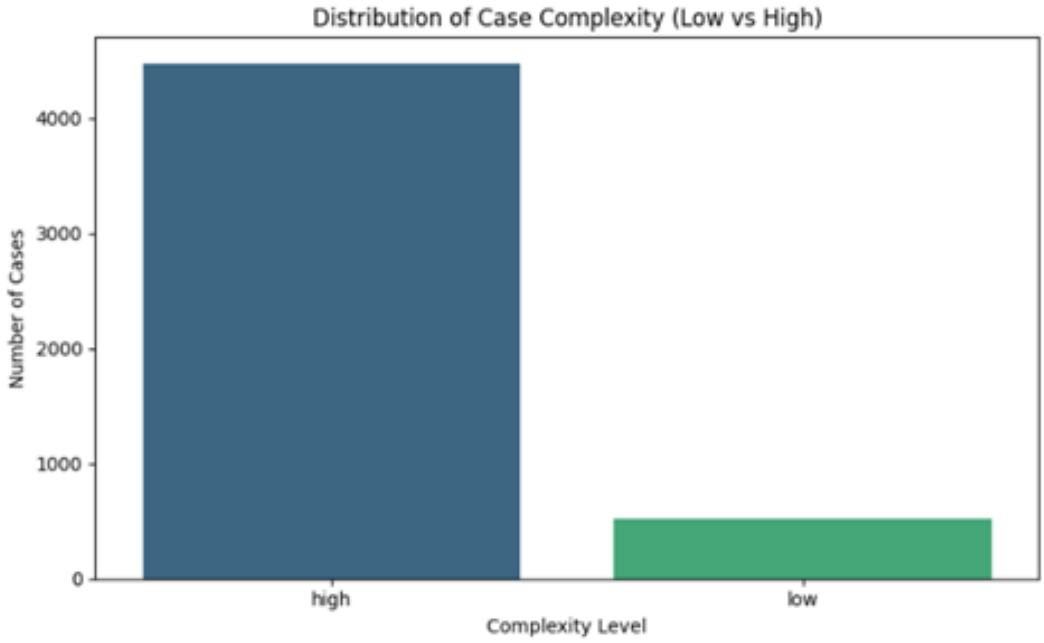

Figure 1. Tax Dispute Complexity

The GPT-powered topic modelling process extracted dominant legal themes from unstructured dispute descriptions. The most frequent topic was Customs Valuation Determination, present in 478 cases, followed by Input Tax Corrections with 286 cases, and Import Duty Tariff Disputes with 146 cases. Other recurring themes included disputes over the tax base, input VAT calculations, and net income assessments. These results validate the model's capacity to group semantically similar cases without relying on pre-coded taxonomies.

Table 1. Five Most Frequent Dispute Theme

No.	Dispute Theme	Frequency
1	Customs Value Determination	476
2	Input Tax Correction	175
3	Import Duty Tariff	146
4	Customs Value Appeal	138
5	Tax Base Correction	120

An additional layer of analysis examined the distribution of case complexity across decision years from 2008 to 2020. The results consistently show a dominance of high-complexity cases in every year of observation. For example, in 2011, 472 out of 481 cases (98.1%) were classified as high complexity; in 2014, 1,235 out of 1,432 cases (86.2%) fell into the same category. Even in years with relatively fewer cases, such as 2012, 2015, and 2019, the share of high-complexity cases remained above 90%.

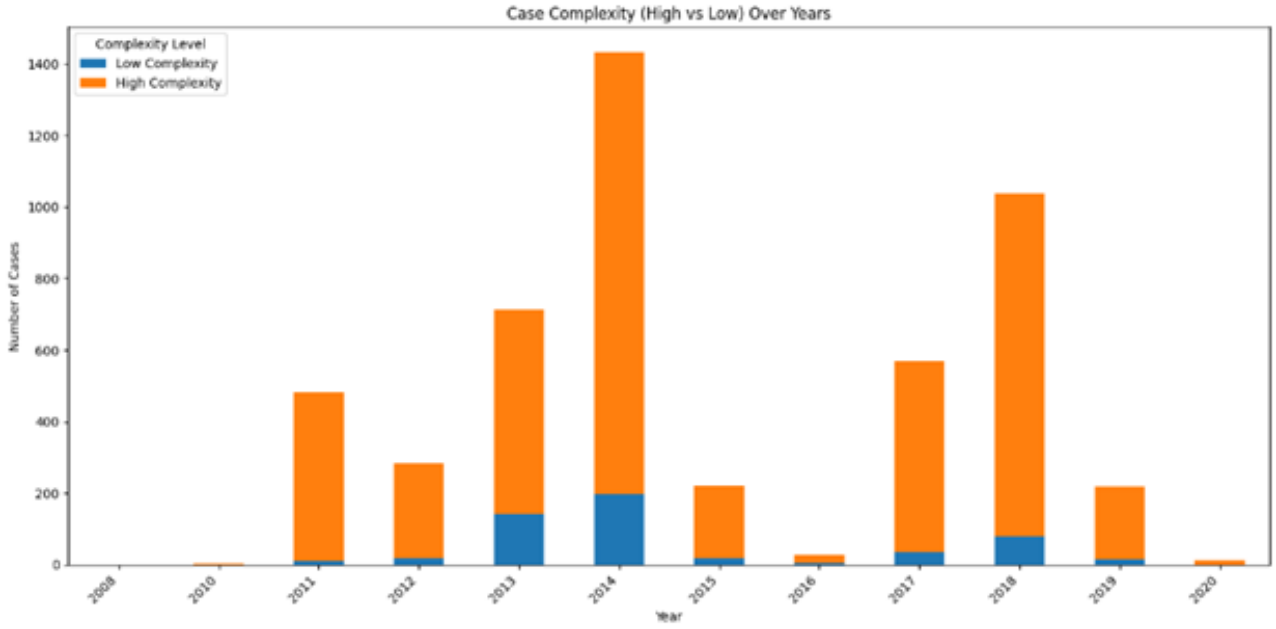

Figure 2. Case Complexity over Time

In contrast, low-complexity cases never exceeded 20% of the total in any given year. This persistent pattern suggests a structural characteristic of the Indonesian Tax Court caseload: cases that escalate to the litigation stage are rarely straightforward. This may reflect either the nature of the disputes themselves, such as technical VAT or customs valuation issues, or inefficiencies in pre-litigation resolution channels (Anyebe, 2020).

Table 2. Most Tax Dispute Theme in VAT, Income Tax, and Customs Dispute

Tax Category	Dispute Theme
Income Tax	Net Income Correction
	Withholding Tax – Article 23
	Withholding Tax – Article 26
Value-Added Tax	Input VAT Correction
	Input VAT Recalculation
	VAT Base Adjustment
Customs	Customs Valuation Determination
	Others Customs Valuation
	Import Duty Tariff Dispute

Across the main tax categories, the model effectively surfaced the most prevalent dispute themes. In Income Tax cases, the most common themes were net income adjustments (n=67), withholding tax disputes under Article 23 (n=56), and non-resident withholding issues under Article 26 (n=41). It highlights a recurring pattern of corrections related to income reporting and withholding procedures. Meanwhile, Value-Added Tax (PPN) disputes were dominated by input VAT issues, with corrections to input tax (n=283) and its recalculation (n=119), followed by adjustments to the VAT tax base (n=99). In the Customs (Bea Cukai) domain, the primary concerns included customs valuation determinations (n=470), appeals against such valuations (n=138), and disputes over import duty tariffs (n=125). These concentrated themes indicate persistent friction points within each tax domain.

Case Study VAT and Income Tax

To further illustrate how the model distinguishes between complexity levels, two VAT dispute cases were examined in detail. In a high-complexity case (Put-103679.16/2014/PP/M.IIIA Tahun 2018), the dispute centered on substantial corrections to input VAT involving cross-period transactions, the eligibility of tax invoices, and the legal interpretation of taxable events under Indonesian VAT Law. The case required the interpretation of multi-layered documentation and legal reasoning around substantive versus administrative compliance. The court ultimately granted the appeal in full. It indicates that the taxpayer's arguments, though complex, were legally valid.

In contrast, a low-complexity case (Put-080133.16/2011/PP/M.VA Tahun 2018) also involved a dispute over input VAT, but the issue was procedural. It focused primarily on incomplete invoice details and the timing of VAT credit claims. There were no contested interpretations of law or complex factual disputes, and the resolution largely hinged on verifying standard administrative documents. The case was also fully granted in favor of the taxpayer, but with far less judicial reasoning required.

A parallel comparison was also conducted for disputes involving Income Tax, starting with a high-complexity case related to net income correction. In PUT-106083.14/2011/PP/M.IVB Tahun 2018, the taxpayer challenged the fiscal authority's upward adjustment of net income based on audit findings. The core of the dispute centered on the deductibility of cost of goods sold and operational expenses, which the tax authority had disallowed, citing insufficient documentation. The taxpayer contended that the corrections were not only unsupported by evidence but also failed to reflect the actual business circumstances. This case required careful judicial evaluation of accounting methods, inventory calculations, and audit procedures. It demonstrates the depth of legal and technical scrutiny typical in net income disputes. Its classification as a high-complexity case reflects the level of financial and evidentiary analysis involved.

In contrast, a clear example of a low-complexity income tax dispute can be found in Put-39057/PP/M.XVI/15/2012, which concerned discrepancies in reported business revenue and cost of goods sold (COGS). The tax authority issued a correction after identifying mismatches between sales figures and purchase invoices. The taxpayer did not contest the legal framework but rather provided additional documentation to reconcile the figures. The issue was limited to administrative verification of invoices and accounting entries, without the need for deeper legal or financial interpretation. The Tax Court fully granted the appeal following basic factual clarification. It illustrates the type of dispute that could be resolved efficiently and possibly avoided through improved pre-litigation procedures.

Model Evaluation

To evaluate the reliability of the GPT-based classification model, a benchmark test was conducted using 150 manually labelled tax dispute summaries from the Indonesian Tax Court dataset. Each case was independently assessed for its complexity level, "High" or "Low", based on legal, procedural, and evidentiary criteria. These ground-truth labels were then compared to the model's predictions, which were generated using a few-shot prompting strategy through the OpenAI GPT API. The results showed that the model achieved a precision score of 0.87, indicating that when it classified a case as high complexity, it was correct 87% of the time. This level of accuracy reflects the model's strong contextual reasoning ability, even when interpreting unstructured legal summaries written in varied formats and styles.

This performance aligns well with findings from other legal NLP studies. For example, transformer-based models such as RoBERTa and BERT-based hybrids have achieved precision scores between 0.85 and 0.91 in legal classification tasks in high-resource jurisdictions like Europe, India, and Brazil (Imran et al., 2023; Costa et al., 2023; Siino et al., 2025). While this study did not fine-tune a transformer model, the GPT-based pipeline

effectively replicated expert judgment through few-shot learning and language understanding, demonstrating a scalable alternative for low- to middle-income legal systems.

Conclusion

This study demonstrates the potential of large language models, to classify legal case complexity and extract thematic insights from unstructured Indonesian Tax Court decisions. By analysing 5,000 tax dispute summaries from 2008 to 2018, the model effectively categorized the vast majority of cases as high complexity and identified recurring themes such as customs valuation and VAT corrections. A benchmark evaluation using 150 manually labelled cases showed a precision score of 0.87, affirming the model's ability to simulate expert-level legal reasoning. These results not only validate the model's accuracy but also highlight its scalability and adaptability for legal systems dealing with high volumes of unstructured data.

The application of machine learning and natural language processing shows strong potential to improve consistency in tax classification and automate the identification of dispute-prone topics. It offers valuable support for more efficient dispute resolution. Nonetheless, the model exhibited some limitations. It occasionally misclassified borderline cases, especially when dispute narratives were brief, poorly structured, or lacked clear legal arguments. Furthermore, predictions were sometimes sensitive to prompt phrasing, indicating the need for improved input standardization and further model refinement.

References

Aastha Budhiraja. (2024). Machine Learning Infused Approach for Advancing Legal Predictive Analytics. *Communications on Applied Nonlinear Analysis*, *31*(8s), 352–364. https://doi.org/10.52783/cana.v31.1506

Afiyati, R., Sudarsono, Negara, T. A. S., & Koeswahyono, I. (2022). Tax dispute settlement mediation arrangements in the future tax court. *International Journal of Research in Business and Social Science (2147- 4478)*, *11*(5), 503–511. https://doi.org/10.20525/ijrbs.v11i5.1867

Andriati, S. L., Rizki, I. K., & Malian, A. N. B. M. (2024). Justice on Trial: How Artificial Intelligence is Reshaping Judicial Decision-Making. *Journal of Indonesian Legal Studies*, *9*(2). https://doi.org/10.15294/jils.v9i2.13683

Anyebe, P. A. (2020). Tax Disputes Resolution In Nigeria: Going Beyound The Traditional Court And Administrative Resolution System. *Advances in Social Sciences Research Journal*, *6*(12), 236–252. https://doi.org/10.14738/assrj.612.7574

Costa, Y. D. R., Oliveira, H., Nogueira, V., Massa, L., Yang, X., Barbosa, A., Oliveira, K., & Vieira, T. (2025). Automating petition classification in Brazil's legal system: a two-step deep learning approach. *Artificial Intelligence and Law*, *33*(1), 227–251. https://doi.org/10.1007/s10506-023-09385-4

Ding, Z. (2024). A Study on the Multiple Dispute Resolution Mechanisms of Systemic Jurisprudence in the Context of Big Data. *Applied Mathematics and Nonlinear Sciences*, *9*(1). https://doi.org/10.2478/amns-2024-0859

Dokumacı, M. (2024). AI-Driven Econometric Models for Legal Issues. *Human Computer Interaction*, *8*(1), 137. https://doi.org/10.62802/btfvze98

Elliot, M., & Thomas, R. (2020). The Effectiveness and Impact of Judicial Review. In *Public Law*. Oxford University Press.

Han, W., Shen, J., Liu, Y., Shi, Z., Xu, J., Hu, F., Chen, H., Gong, Y., Yu, X., Wang, H., Liu, Z., Yang, Y., Shi, T., & Ge, M. (2024). LegalAsst: Human-centered and AI-empowered machine to enhance court productivity and legal assistance. *Information Sciences*, *679*, 121052. https://doi.org/10.1016/j.ins.2024.121052

Imran, A. S., Hodnefjeld, H., Kastrati, Z., Fatima, N., Daudpota, S. M., & Wani, M. A. (2023). Classifying European Court of Human Rights Cases Using Transformer-Based Techniques. *IEEE Access*, *11*, 55664–55676. https://doi.org/10.1109/ACCESS.2023.3279034

Sari, I., Kosasih, R., & Indarti, D. (2024). Predicting levels of legal case difficulties using machine learning. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *13*(4), 4364. https://doi.org/10.11591/ijai.v13.i4.pp4364-4371

Siino, M., Falco, M., Croce, D., & Rosso, P. (2025). Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. *IEEE Access*, *13*, 18253–18276. https://doi.org/10.1109/ACCESS.2025.3533217

Sukanya, G., & Priyadarshini, J. (2024). Hybrid CNN: An Empirical Analysis of Machine Learning Models for Predicting Legal Judgments. *International Journal of Advanced Computer Science and Applications*, *15*(7). https://doi.org/10.14569/IJACSA.2024.01507124

Zhang, L., Ma, Y., Herman, D., & Chen, J. (2022). Testing calibration of phenotyping models using positive-only electronic health record data. *Biostatistics*, *23*(3), 844–859. https://doi.org/10.1093/biostatistics/kxab003

Zulfiqar, F. L., Ulupui, I. G. K. A., & Respati, D. K. (2023). A qualitative analysis on transfer pricing tax audit performance in Indonesia. *AKURASI: Jurnal Riset Akuntansi Dan Keuangan*, *5*(1), 73–84. https://doi.org/10.36407/akurasi.v5i1.805