Legal Case LLM: An Open-Source Fine-Tuned Model for Indonesian Human Trafficking Jurisprudence

Muhammad Hariz Faizul Anwar[1], Nizam Avif Anhari[2], Galih Wasis Wicaksono[3*], Nur Putri Hidayah[4]

[1,2,3]*Informatics, University of Muhammadiyah Malang, Indonesia*
[4]*Law, University of Muhammadiyah Malang, Indonesia*

*corr-author: galih.w.w@umm.ac.id

Abstract - This paper presents Legal-Case LLM, an open-source, fine-tuned language model tailored for Indonesian human-trafficking jurisprudence. General-purpose large language models exhibit high fluency but risk factual hallucination and limited jurisprudential fidelity when applied to legal texts. The objective is to develop a reproducible model that improves factual recall, legal terminology use, and jurisprudential alignment for Indonesian trafficking cases. Methods: We assembled a curated corpus of 400 court decisions from the Direktori Putusan Mahkamah Agung, extracted structured metadata and summaries, and generated question–answer pairs via large models followed by multi-stage cleaning and expert validation. We fine-tuned open models from the LLaMA family variants using parameter-efficient techniques (LoRA), evaluated with automatic metrics (ROUGE, BLEU, BERTScore, BARTScore), and a focused qualitative audit. Results: The fine-tuned model demonstrates marked improvements in content recall and semantic alignment versus zero-shot baselines, produces more jurisprudentially aligned phrasing (accurate use of terms such as amar putusan, Majelis Hakim, and percobaan), and reduces hallucination propensity in statute-related outputs. Conclusion and impact: Legal-Case LLM offers a reproducible, transparent tool to assist legal practitioners and researchers in Indonesia, while emphasising human-in-the-loop verification and citation-matching to ensure legal reliability and ethical deployment.

Keywords: Indonesian human trafficking; legal LLM; jurisprudence; legal AI, transformers.

I. INTRODUCTION

Large Language Models (LLMs) have fundamentally transformed numerous domains through their remarkable ability to understand, process, and generate human-like text with unprecedented accuracy and fluency [1-4]. These sophisticated AI systems have demonstrated exceptional performance in diverse applications ranging from machine translation and multilingual communication to automated code generation and software development, revolutionizing how professionals perform complex linguistic and analytical tasks [5-8]. The success of LLMs in these domains stems from their capacity to capture intricate patterns in language, understand contextual nuances, and generate coherent responses that often rival human-level performance [9-11].

However, despite these remarkable achievements in general-purpose applications, the adoption and implementation of LLMs in the legal profession remains notably limited and underdeveloped. This limitation is particularly pronounced, given the unique characteristics of legal language, fundamental issues of factual reliability, and the critical jurisprudential concerns and data confidentiality that are central to the legal domain. These demand precision, adherence to established precedents, and a deep understanding of jurisprudential principles that differ significantly from those required in everyday language processing tasks. As legal professionals increasingly face voluminous, complex documents that are challenging to navigate, analyze, and interpret within reasonable timeframes, a clear and urgent need emerges for specialized LLMs designed for legal applications [12,13].

The modern legal landscape is characterized by an exponential growth in document complexity and volume, encompassing statutory texts, case law, regulatory frameworks, contracts, and legal briefs that require sophisticated analytical capabilities to process effectively [14,15]. This unprecedented complexity presents a significant challenge yet a substantial opportunity for AI systems specifically engineered to assist with the comprehensive analysis, interpretation, and synthesis of legal documents. The development of such specialized systems could address critical inefficiencies in legal practice, reduce the time required for document review and analysis, and enhance the

accuracy of legal research and decision-making, thereby transforming the delivery of legal services and improving access to justice.

Recent studies highlight the diverse applications of LLMs in the legal field. For public legal aid, "LegalBot-EC" used a BERT-like model to democratize access to information in Ecuador, achieving 94% accuracy [16]. In tax law and AI governance, research on GPT-4 indicates that it can improve efficiency and reduce costs, though it currently lacks reliable autonomous reasoning [17]. Similarly, a study on food safety and GDPR compliance compared multiple models (including GPT-4 and Mixtral) for automating requirement classification; BERT achieved the highest performance with an 87% F-score [18]. Collectively, these findings suggest that while LLMs significantly enhance efficiency, they still require further development to achieve fully reliable autonomy.

By addressing the specific needs of legal practitioners through a specialized pre-training approach that utilises domain-specific legal datasets, our work represents a significant advancement in AI applications for legal purposes. This initiative introduces the first open-weight LLM, specifically fine-tuned for Indonesian law and designed to empower legal professionals and accelerate innovation at the intersection of artificial intelligence and jurisprudence. Our research aims to enhance both the theoretical understanding and practical application of legal language processing within the Indonesian legal system.

The selection of human trafficking as the focus is driven by its rising prevalence in Indonesia, migrant exploitation, and digital technology misuse [19,20]. Publicly available data from the Supreme Court Decision Directory enables high-quality dataset curation. However, general models often fail due to language bias and hallucinations from Indonesian legal language issues, such as hybrid terminology (e.g., 'delik', 'actus reus', 'cyber-trafficking'), rigid formatting per Decree No. 359/KMA/SK/XII/2022 (e.g., 'Majelis Hakim', 'Pertimbangan Hukum'), and code-mixing with regional dialects. These impair LLMs' jurisprudential fidelity, making this domain ideal for testing LLM fine-tuning in local law with high social impact, such as supporting trafficking eradication and justice access. Thus, in this study, we summarize the main contributions as follows.

This work contributes a factual, jurisprudence-specific LLM for Indonesian Human Trafficking Law. While large models like GPT-4 possess strong zero-shot capabilities, their generative nature often leads to hallucinations of verdicts or case numbers, risking legal malpractice [21]. Furthermore, English-heavy pre-training introduces cultural biases that are ill-suited to interpreting Indonesian law [22]. Unlike prior encoder-based models (e.g., LegalBERT, Lawformer) focused on English or Chinese classification tasks [14,23], Legal-Case LLM employs a decoder-only architecture (LLaMA 3.2) instruction-tuned for generative reasoning. This approach addresses the specific challenges of Indonesian jurisprudence, a hybrid of Dutch civil, Adat, and Islamic law, which requires accurate terminology (e.g., Majelis Hakim) absent in general benchmarks. Methodologically, we use a parameter-efficient synthetic data pipeline to address factual hallucination, improving the ROUGE score to 24, compared to 11 and 12 for zero-shot baselines. As the first open-source model in this domain, Legal-Case LLM enables transparent reproducibility and research impossible with closed-source alternatives.

II. METHOD

This research was conducted in four stages, as presented in the methodological framework in Fig. 1. The first stage was collecting decision documents from Supreme Court Decision Directory of Indonesia (Direktori Putusan Mahkamah Agung Indonesia), followed by validating the collected documents to ensure their validity, extracting summaries from the decision documents, and the fourth stage, creating questions and answer data to train the fine-tuned model.

A. Data Acquisition

The first stage is data collection. We compiled a legal dataset by gathering decision files from the Supreme Court Decision Directory of Indonesia, focusing on human trafficking cases. We obtained 408 documents, of which 404 were deemed eligible after verification. These public decisions are anonymized by courts under SK KMA No. 1-144/KMA/SK/I/2011 and PERMA No. 3/2017, particularly for minor cases (e.g., victims as 'Korban'). For adults, redactions are partial; we applied post-processing scripts for further obfuscation (e.g., name placeholders) and excluded potential leaks to ensure ethical training without harming judicial transparency.

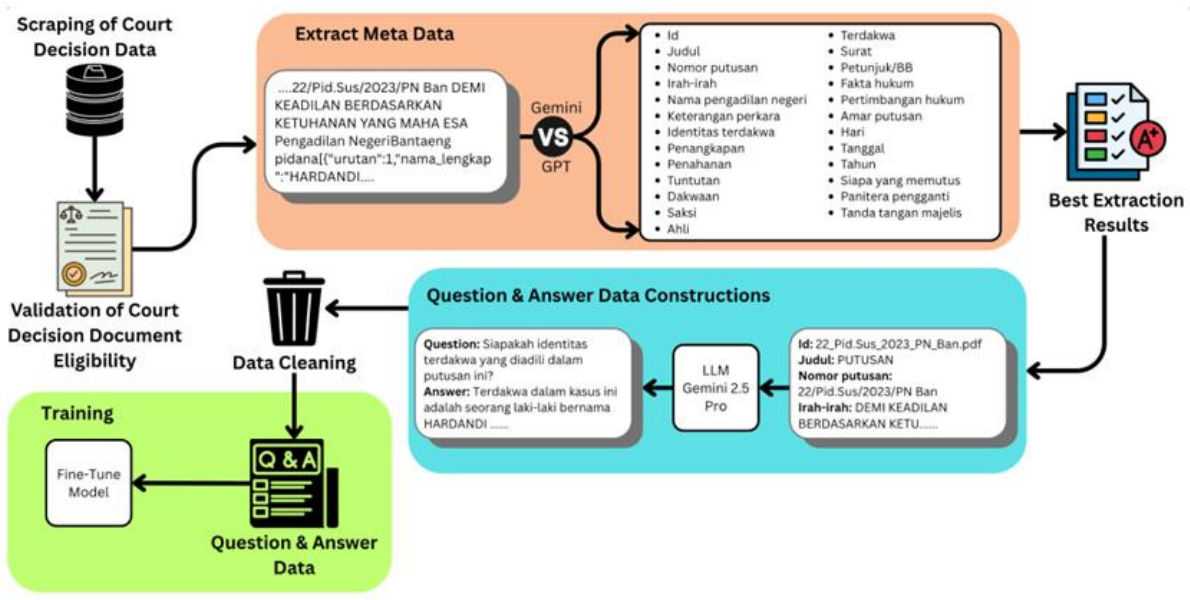

Fig. 1 Overview of our methodology flow

The dataset is available on Mendeley Data under a CC BY 4.0 license [24], allowing reuse with attribution. Original decisions are in the public domain per Law No. 14 of 2008 concerning Public Information Disclosure. Anonymization complies with Law No. 27 of 2022 concerning Personal Data Protection, with courts redacting identifiers (e.g., names to initials). The structure uses JSON files with fields for case ID, metadata (e.g., verdict number, judges), full text, summaries, and Q&A pairs. Annotations follow Supreme Court Decree No. 359/KMA/SK/XII/2022. Q&A pairs were synthetically generated via Gemini 2.5 Pro and cleaned for relevance (e.g., removing hallucinations).

Due to resource limits, formal inter-annotator agreement (e.g., Cohen's κ) was not measured. Instead, we used a hybrid validation: legal experts reviewed for compliance with Indonesian standards, while GPTScore [23] assessed factuality and relevance, leveraging generative models for multifaceted quality checks without large annotated samples.

Following the collection, documents underwent rigorous verification to address the variable quality of Indonesian court classifications. Legal experts validated each decision for consistency and adherence to legal principles. This study utilizes a dataset originally curated for prior unpublished work, now openly available on Mendeley Data, to ensure reproducibility [24].

B. Metadata Extraction

This research implements automatic extraction of court decision documents using Large Language Model (LLM) technology, including Google Gemini and OpenAI GPT. The extraction process was conducted in accordance with the standardization structure stipulated in the Decree of the Chief Justice of the Supreme Court, Number 359/KMA/SK/XII/2022, concerning Court Decision Templates/Guidelines. Important parts of the verdict that need to be extracted include ID, Title, Verdict Number, Irah-irah, District Court Name, Case Details, Defendant Identity, Arrest, Detention, Indictment, Indictment Letter, Witnesses, Experts, Defendant Testimony, Letters, Evidence (BB), Legal Facts, Legal Considerations, Verdict, Day, Date, Year, Presiding Judges, Deputy Court Clerk, and Signatures of the Presiding Judges [25].

After successfully extracting the documents, legal professionals examined the summaries to determine which was best between the results from OpenAI GPT and Google Gemini.

C. Question & Answer Data Construction

We opted to employ Google's Gemini 2.5 Pro for generating synthetic data for the Q&A dataset over alternatives like GPT-4o.

Fig. 2 shows Gemini achieved lower factuality variance (σ=0.501 vs. σ=0.812) and higher relevance (3.501 vs. 3.408) than GPT, offering the stability needed for legal definitions. Experts favored Gemini's precision over GPT's verbosity. To ensure scalability and reduce bias, we generated synthetic Q&A pairs using Gemini 2.5 Pro based on texts from Section B. This validated dataset, detailed in Table I, was used for fine-tuning [25,26].

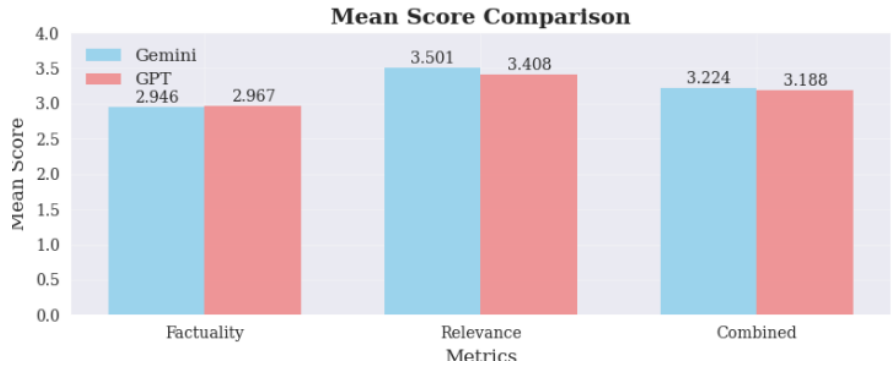

Fig. 2 Gemini & GPT performance comparison

TABLE I
SAMPLE OF QUESTION AND ANSWER DATASET

Question	Answer	Source Distill
Bagaimana status akhir dari barang bukti berupa telepon genggam dan sejumlah uang tunai yang disita dalam perkara ini berdasarkan amar putusan hakim?	*Dalam amar putusannya, Majelis Hakim menetapkan bahwa [...] untuk negara. Sementara itu, barang bukti berupa satu buah kondom bekas pakai dirampas untuk dimusnahkan.*	Gemini 2.5 Pro
Dari empat dakwaan alternatif yang diajukan, mengapa Majelis Hakim menyatakan Terdakwa bersalah berdasarkan Pasal 88 Undang-Undang Perlindungan Anak?	*Majelis Hakim memilih dakwaan keempat, yaitu Pasal 88 Jo. Pasal 76I [...] terbukti dibandingkan dakwaan lain seperti perdagangan orang.*	Gemini 2.5 Pro

D. Data Cleaning

In the Data Cleaning phase, we implemented a multi-stage filtering approach to identify and remove unsuitable data points from the initially generated Q&A pairs. The cleaning process involved both automated and manual review procedures, with which we systematically evaluated each question-answer pair for relevance, accuracy, and linguistic quality. Questions that exhibited poor formulation, lacked clear connection to the source material, or generated responses that were inconsistent or off-topic were systematically excluded. Any punctuation, such as slashes, was removed from the text. This data-cleaning approach ensured that only relevant Q&A pairs were retained for model training, thereby improving the effectiveness of the overall fine-tuning process and enhancing the model's ability to generate accurate responses.

E. Experimental setup

In this section, we fine-tuned the Llama 3.2 3B base model (meta-llama/Llama-3.2-3B) model using the dataset described in section C. The selection of this compact architecture is substantiated by recent evaluations in legal NLP, which demonstrate that Small Language Models (SLMs), when optimized with parameter-efficient methods, achieve comparable fidelity to larger foundational models in specialized legislative tasks while significantly reducing deployment latency [27]. We loaded the model with BF16 precision to balance performance and resource efficiency. Fine-tuning was performed using the Unsloth framework. We configured LoRA with a rank (r) of 64 and a scaling factor (α) of 128, resulting in a scaling factor of α/r=2. The adaptation matrices were applied to the q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj modules, with a max sequence length of 2048 with RoPE scaling [28]. The weight update can be expressed as

$$w = w_0 + \frac{a}{r} BA \quad (1)$$

Where w is the updated weight matrix, w_0 is the original weight matrix, and B x A are the low-rank decomposition.

To ensure convergence on a single L4 GPU (22 GB VRAM), we utilized an effective batch size of 8 (per-device size 2, gradient accumulation 4), following [25], The model was trained for 3 epochs (309 steps) on 822 samples to balance learning with overfitting risks. Data

was structured in JSON format (keys: role, content) to support instruction tuning, as detailed in Table II.

To prevent data leakage, we performed the train-test split at the document level. All Q&A pairs derived from a distinct court decision (Case ID) were assigned exclusively to either the training or test set. We utilized stratified sampling based on case years (2023-2024) to ensure temporal balance, resulting in an 80/20 split of the original documents (approx. 822 training pairs and 204 test pairs). The training set was used for fine-tuning with LoRA, and the test set for quantitative evaluation (ROUGE, BLEU, BERTScore, BARTScore, and Citation-F1) and qualitative expert audits. This split ratio promotes model robustness and prevents data leakage.

The model was trained for 3 epochs with a maximum of 309 steps, using the AdamW optimizer with BF16 to reduce memory footprint and increase training throughput while maintaining numerical stability. Unlike INT8 quantization, BF16 maintains the same dynamic range as FP32 (8-bit exponent) by truncating the mantissa, thereby eliminating the need for complex scaling or quantization formulas.

Logging was performed every step, and training runs were tracked using Weights & Biases. We compared the performance of our fine-tuned models with Qwen 2.5 3B and Microsoft phi 3 3B, assessing performance using a combination of quantitative and qualitative metrics. For the quantitative analysis, we calculated metrics using ROUGE, BLEU, BERTScore, and BARTScore.

All environment runs were tracked with Weights & Biases and executed on Google Colab using a single L4 GPU with 24 GB of VRAM and 20 GB of dedicated memory. Our code was based on Python 3.12 and Huggingface Transformers.

III. RESULT AND DISCUSSION

In this section, we present the key findings from our dataset curation, model fine-tuning, and evaluation, followed by a discussion of their implications for Indonesian legal AI. Results are illustrated through figures, tables, and metrics to demonstrate the model's performance improvements over baselines.

As shown in Fig. 3, a total of 408 district court decisions related to human trafficking cases were collected from the Indonesian Supreme Court Decision Directory website. These decisions comprise 260 from 2023 and 148 from 2024. After verification by legal experts in accordance with the Decree of the Chief Justice of the Supreme Court Number 359/KMA/SK/XII/2022 concerning the Template/Guidelines for Court Decisions, 404 documents were declared eligible and appropriate. From these documents, approximately 1,019 Q&A pairs were collected for fine-tuning.

To further assess the quality of our model's output, we employed several evaluation metrics, including ROUGE [29] and BLEU [30]. However, we recognize that n-gram overlap metrics are insufficient for capturing jurisprudential validity, as they cannot distinguish between legally distinct but textually similar terms (e.g., 'murder' vs. 'manslaughter'). To address this, we implemented two domain-specific enhancements: BERTScore [31], Citation F1, and BARTScore [32]. We developed a custom Regex-based metric to evaluate citation accuracy. This function extracts legal references (e.g., Undang Undang, Pasal, Peraturan Pemerintah) from both the generated and reference texts to calculate an F1 score specifically for statutory citations [33]. This serves as a proxy for measuring hallucination, ensuring the model does not fabricate non-existent laws. ROUGE measures the overlap of n-grams and the longest common subsequence between the generated output and the reference texts, focusing on recall-oriented metrics such as content coverage. BLEU evaluates precision by comparing n-gram matches, commonly used for machine translation but adaptable to other generation tasks. BERTScore leverages contextual embeddings. We calculated BERTScore using the indobenchmark/indobert-base-p1 model. Unlike generic BERT models, this localized embedding better captures the nuances of Indonesian legal syntax and semantics. BARTScore, derived from BART's sequence-to-sequence probabilities, quantifies faithfulness and fluency by estimating the likelihood of the output, given the input.

TABLE II
HYPERPARAMTER SETUP

Hyperparameter	Value
Batch size	2
Gradient accumulation step	4
Epochs	3
Optimizer	AdamW
Learning rate	1e-4
Weight decay	0.01
Scheduler	Cosine
Logging steps	1
Seed	3407
Output directory	outputs
W & B tracking	Enabled

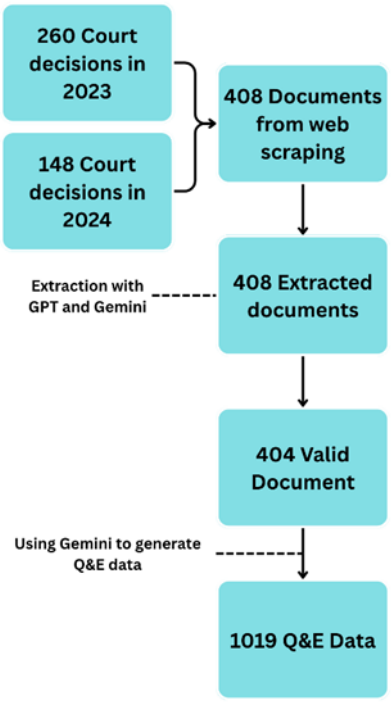

Fig. 3 Distribution of court decision documents data

We compared the performance of our fine-tuned models with Qwen 2.5 3B Instruct and Microsoft Phi-3 Mini 4k Instruct. Then, we selected these two models to establish a rigorous performance benchmark against the most current State-of-the-Art (SOTA) open-weight models in the "small language model" (SLM) parameter class (3B-4B). Qwen 2.5 3B was chosen to represent the peak of current dense model performance at this size. Unlike previous baselines, this model was pre-trained on a massive dataset of 18 trillion tokens, providing a foundation of expert knowledge and reasoning capabilities that rivals those of much larger models [34]. Conversely, Microsoft Phi-3 Mini was selected to represent the "data-centric" optimization approach. As detailed in the Phi-3 Technical Report, this model challenges traditional scaling laws by utilizing a dataset of highly filtered "textbook-quality" web data and synthetic content. This allows us to test our specialized model against a generalist baseline that achieves reasoning capabilities comparable to larger models (e.g., Llama-2-7B) through superior data curation rather than raw parameter count [35]. The quantitative results are summarized in Table III below.

Table III highlights the superiority of our fine-tuned model over zero-shot baselines. Our model nearly doubled baseline ROUGE-L scores (24.33 vs. ~12.0), indicating superior legal phrase recall and structural integrity. While our BLEU score (11.49) vastly outperforms baselines (~2.0), the low absolute value reflects valid synonym usage (e.g., "menjatuhkan pidana" vs. "memutuskan hukuman") rather than the semantic errors hypothesis confirmed by our high Citation F1 (69.06), which proves the statutory core remains accurate despite phrasing variations. Finally, while Phi-3 slightly edged out BARTScore (-3.41) due to general fluency, our model's leading BERTScore (59.89) confirms that generalist models lack the specific domain grounding required for accurate Indonesian jurisprudence.

We conducted a domain-expert validation of model outputs using a three-dimensional assessment framework derived from [25], encompassing informativeness (coverage of legally relevant facts), factualness (adherence to source material), and fluency (professional linguistic quality). The evaluation reveals a critical trade-off between generative coherence and juridical accuracy when deploying language models on unseen trafficking-in-persons (TPPO) cases from the Indonesian judicial database (Decision Directory).

In the first case assessment, the model processed a scenario involving MiChat-based prostitution facilitation characterized by a pseudonymous account ("Amanda") and a negotiable fee range (Rp 200,000 - Rp 500,000). The output achieved high informativeness (4/5) by correctly identifying core juridical elements, specifically the digital platform, the account's pseudonymous nature, and the commercial exploitation framework. However, factualness scored poorly (2/5) due to material hallucinations. The model fabricated a secondary alias ("SALSA") and erroneously assigned it to the victim while constructing a non-existent "rental fee" transactional schema involving dual-defendant coordination. This represents a conjunctive hallucination—the synthetic merging of plausible but unverified operational details. Additionally, fluency (3/5) was compromised by register inconsistency, notably an unwarranted code-switch into English ("respectively"), which violates Indonesian legal drafting conventions. These errors are juridically significant; under KUHAP Article 184, the precise identification of digital artifacts constitutes material evidence, and such creative extrapolation risks misleading investigative focus or generating inadmissible synthetic facts.

TABLE III
SYNTACTIC AND SEMANTIC METRIC EVALUATION RESULT

Model	Bart	Bert	Bleu	Rouge	Citation F1
Llama 3.2 3B (Ours)	46 ± 1.8	59 ± 0.5	11.48 ± 0.62	24 ± 0.7	69.1 ± 0.6
Qwen 2.5 3B	70 ± 1.5	46 ± 0.1	2.07 ± 0.04	11 ± 0.1	34 ± 2.0
Microsoft Phi 3 mini 4K	41 ± 8.3	44 ± 0.3	1.84 ± 0.06	12.0 ± 0.1	57 ± 0.2

A second evaluation focused on a witness tasked with victim transportation for sexual services, receiving compensation of Rp50,000-Rp100,000 per assignment. Here, the model attained optimal informativeness (5/5) and fluency (5/5), producing a professionally articulated summary that accurately captured the witness's logistical function and compensation structure. Nevertheless, factualness remained suboptimal (3/5) due to associative hallucination: the model inferred the witness's age as "16 years old" and appended unverified duties, such as "housing the victim," which were absent from the source dossier. While these additions are probabilistically reasonable within TPPO networks, they constitute speculative jurisprudence. This is particularly critical in the context of Indonesian child protection law, where precise age determination triggers mandatory procedural safeguards.

The discrepancy between these samples exposes a fundamental limitation, namely the model's inability to distinguish between probable contextual inference and evidentiary fact. While the fine-tuned Llama 3.2 3B model demonstrates an exceptional capacity to mimic legal analytical style and extract salient exploitation patterns consistent with prosecutorial narratives, its tendency toward plausible confabulation, particularly when source texts are elliptical, poses acute risks. The persistent factualness deficit (averaging 2.5/5) indicates that without robust grounding mechanisms, the model prioritizes narrative completeness over the evidentiary purity required for TPPO case processing.

To rigorously validate the model's reliability, we conducted a granular failure analysis on a stratified sample of 50 generated responses. Adhering to the reference-free evaluation protocols (G-Eval) [36] and the multifaceted scoring framework [23], we employed an automated auditor to categorize errors into distinct taxonomies. The corresponding results are presented in Fig. 4 and summarized in Table IV.

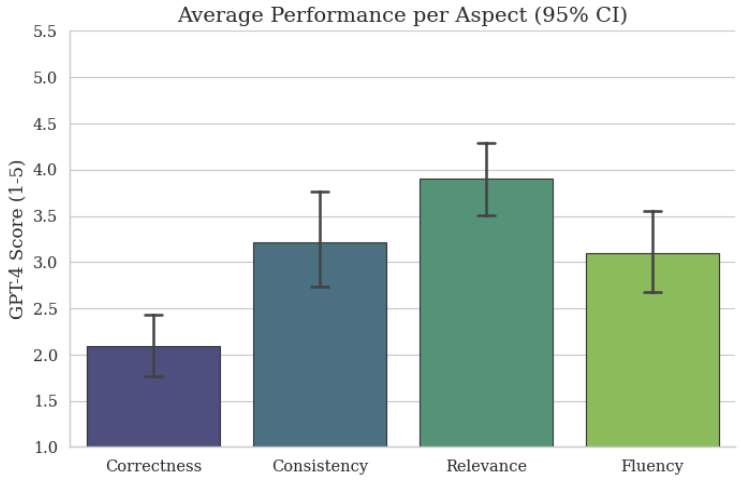

Fig. 4 Model Performance comparison by metric

TABLE IV
GPTScore & G-EVAL ERROR ANALYSIS

Error Type	Contex Input	Generated Output	Score	Analysis
Correctness	*Q: Bagaimana modus operandi yang digunakan oleh ketiga terdakwa dalam mengeksploitasi anak korban, Muzdalifah Muchsin? Ref: Berdasarkan fakta persidangan, ketiga terdakwa... memfasilitasi anak korban... [...] serta dibayari makan dan rokok.*	*Ketiga terdakwa, Didi Komaling, Muh. Taufik Sangkala, dan Dhimas Galih Prandika, bekerja sama […] sebagai "admin" utama yang mengelola seluruh komunikasi, […] mengambil keuntungan, yang menunjukkan adanya tujuan eksploitasi ekonomi.Ã±ldÄ±ÄŸÄ±n[...]ç¥ž é©¬æ''¶å½•Ä±ldÄ±ÄŸÄ±"> Ñ‹ÑŸN*	1.0	The generated answer starts correctly by describing the modus operandi of the defendants in exploiting the victim. However, it then devolves into a series of nonsensical characters and phrases that are not related to the question or the legal context. This makes the answer factually incorrect and incomplete. The hallucinations in the text also make it difficult to understand the intended message.
Consistency	*Q: Bagaimana Majelis Hakim menjelaskan terpenuhinya unsur "turut serta melakukan perbuatan" dalam kasus Terdakwa Sultan, meskipun ia sendiri juga berstatus sebagai calon pekerja migran? Ref: Majelis Hakim menjelaskan bahwa unsur "turut serta melakukan perbuatan" terpenuhi... Terdakwa secara aktif melaksanakan permintaan Sdr. Hasan... [...]*	*Majelis Hakim menjelaskan terpenuhinya unsur [...] secara non-prosedural dianggap sebagai bentuk bantuan aktif (medeplichtigheid) yang memenuhi unsur pidana bersama-sama.assistantespoÅˆÐ¸Ð»Ð° ÑÑÃ¼slÃ[...]¼ã,«ãƒ¼ujÃ-cÃ-m*	1.0	The generated answer starts off consistently, discussing the actions of the defendant and their involvement in a crime. However, the text suddenly devolves into a nonsensical string of words and symbols from various languages, which is highly inconsistent with the initial part of the text. This makes the overall text highly inconsistent.
Relevance	*Q: Apa yang menjadi inti dari argumen pembelaan Terdakwa, dan mengapa argumen tersebut ditolak oleh Majelis Hakim? Ref: Inti dari argumen pembelaan Terdakwa adalah bahwa ia hanya menjalankan "perintah jabatan" (Pasal 51 KUHP)... [...] Namun, hakim menolak karena perbuatan tersebut secara inheren melawan hukum.*	*Inti dari argumen pembelaan Terdakwa adalah bahwa ia hanya membantu saksi korban [...] secara teknis sudah memenuhi unsur "melakukan percobaan".useruseruserÑƒÐ² Ð°Ñ,Ð¸ÑÑuseruser[...]Ñ‹ÑŸN Ð¸Ð»Ð°ÑÑ*	1.0	The generated answer starts off relevant to the question, discussing the defendant's argument and why it was rejected by the judges. However, it quickly devolves into a string of nonsensical text and unrelated languages, making it largely irrelevant to the question asked.
Fluency	*Q: Bagaimana modus operandi yang digunakan oleh Terdakwa untuk mengeksploitasi saksi korban, yang merupakan pacarnya sendiri? Ref: Berdasarkan fakta persidangan, Terdakwa menggunakan hubungan asmaranya sebagai modus utama untuk mengeksploitasi korban.[...] total keuntungan pribadi mencapai sekitar Rp 2.000.000,-.*	*Berdasarkan fakta persidangan, Terdakwa yang berprofesi sebagai karyawan swasta, menjalankan modus [...] keuntungan pribadi sebesar Rp 100.000,- sebagai fee atau imbalan jasa.Ã¼slÃ¼manassistantÐ¸Ð »Ð°Ð°Ñ,Ð¸ Ñ…Ð°Ñ€Ð°Ð°Ñ,ÐµÑ€[...]anc Ä±*	1.0	The generated text starts off with coherent and grammatically correct Indonesian language, but it suddenly devolves into a string of unrelated words and characters from various languages, making it unreadable and nonsensical. The fluency is severely compromised due to this.

The observed generation anomalies, characterized by the transition from coherent legal synthesis into repetitive artifacts or stochastic noise, represent a failure mode identified as Post-Termination Degeneration caused by a Truncation-Termination Mismatch. This error is not attributable to semantic reasoning deficits or Rotary Positional Embedding (RoPE) failures, as the model maintains high fidelity throughout the logical response; instead, it results from the systematic excision of End-of-Turn (`<|eot_id|>`) tokens during the Supervised Fine-Tuning (SFT) phase due to fixed-length data truncation. Consequently, the model failed to learn a semantic stop condition, leading to a distributional collapse during inference, where, upon completing the valid output, it autoregressively samples from the low-probability tail of the pre-training distribution or leaks internal chat template structures (e.g., `useruseruser`) in an attempt to continue the sequence indefinitely.

Table V presents the ablation study quantifying the performance loss due to quantization. When comparing the BF16 baseline against the 8-bit quantized variant, we observed a minimal performance degradation: a decrease of only 0.0863 in ROUGE-1 and 0.0236 in BERTScore. These results indicate that the semantic capabilities of Legal-Case LLM are robust to quantization. While BF16 offers the theoretical upper bound in performance, the 8-bit variant retains 98.4% of the BERTScore performance while significantly lowering the memory barrier for deployment. This confirms that the model can be effectively deployed on cost-efficient hardware (e.g., T4 or consumer GPUs) with negligible loss in legal reasoning fidelity.

In this study, we benchmark Legal-Case LLM against Microsoft-Phi3 and Qwen 2.5. Our fine-tuned model demonstrates superior semantic performance, exhibiting improved content recall and stylistic alignment with Indonesian legal terminology. However, it displays lower syntactic quality than the comparison models; we attribute this distinction to fine-tuning a non-reasoning architecture, whereas the baselines possess inherent reasoning capabilities. Despite this, the model exhibits significant input fault tolerance, effectively disambiguating intent from noisy prompts, a capability derived from training on diverse, real-world text data.

Ethically, this research prioritizes strict adherence to Law No. 27 of 2022 regarding sensitive data. We ensured no de-anonymization or extraneous data collection occurred, limiting the dataset to academic purposes under a CC BY 4.0 license. We emphasize that the model is a support tool, not a substitute for professional legal advice, defining "responsible deployment" strictly as the minimization of hallucinations (false citations). Users must verify outputs against primary sources; the current scope excludes broader deployment mechanics in favor of rigorous accuracy benchmarking.

We acknowledge limitations in our methodology. Synthetic data generation risks propagating English-dominant biases, potentially oversimplifying statutes like Law No. 21 of 2007. Factual hallucinations could cascade into the fine-tuned model, reducing reliability and raising ethical concerns regarding misinformation. To mitigate this, we employed multi-stage cleaning and human-in-the-loop oversight. We acknowledge that standard metrics such as ROUGE and BLEU serve only as imperfect proxies for legal validity. Furthermore, our qualitative audit used a limited sample; future work requires more annotators and formal inter-annotator agreement [21]. Finally, resource constraints, specifically a single L4 GPU, necessitate bfloat16 (bf16) precision. While efficient, this restricts full-precision training or experimentation with larger architectures. Future work will integrate Retrieval-Augmented Generation (RAG) and expand the corpus to mitigate hallucinations. We recommend immediate human-in-the-loop verification and automated statute validation to ensure safety. Long-term goals include aligning with UNESCO ethics guidelines and conducting reproducible bias audits using frameworks such as Fairlearn.

IV. CONCLUSION

This research presents Legal-Case LLM, the first open-source model fine-tuned for Indonesian human trafficking jurisprudence. By training on 400 verdicts using a decoder-only architecture, the model significantly outperforms zero-shot baselines in factual accuracy and hallucination reduction. These findings demonstrate that domain-specific fine-tuning can effectively bridge the gap between AI capabilities and the unique complexities of Indonesia's hybrid legal system. While currently limited by dataset size, this work provides a transparent foundation for reproducible legal AI. Future efforts will expand the corpus to corruption and agrarian law, further supporting evidence-based policymaking and judicial consistency.

TABLE V
BF 16 VS 8-BIT PERFORMANCE COMPARISON

Metric	8-bit (T4 Optimized)	bfloat 16 (Unconstrained)	Difference
ROUGE-1	0.4510	0.5373	+0.0863
ROUGE-L	0.3424	0.4471	+0.1047
BERTScore	0.8625	0.8861	+0.0236

REFERENCES

[1] H. Alamleh, A. A. S. AlQahtani, and A. ElSaid, "Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning," in *2023 Systems and Information Engineering Design Symposium (SIEDS)*, Charlottesville, VA, USA: IEEE, Apr. 2023, pp. 154–158. doi: 10.1109/SIEDS58326.2023.10137767.

[2] D. Stap, E. Hasler, B. Byrne, C. Monz, and K. Tran, "The Fine-Tuning Paradox: Boosting Translation Quality Without Sacrificing LLM Abilities," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 6189–6206. doi: 10.18653/v1/2024.acl-long.336.

[3] Y. Wang, H. Le, A. Gotmare, N. Bui, J. Li, and S. Hoi, "CodeT5+: Open Code Large Language Models for Code Understanding and Generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, 2023, pp. 1069–1088. doi: 10.18653/v1/2023.emnlp-main.68.

[4] C. S. K. Aditya and F. D. S. Sumadi, "Combination of Term Weighting with Class Distribution and Centroid-based Approach for Document Classification," *Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control*, vol. 8, no. 4`, Nov. 2023, doi: 10.22219/kinetik.v8i4%60.1793.

[5] K. S. Kalyan, "A survey of GPT-3 family large language models including ChatGPT and GPT-4," *Nat. Lang. Process. J.*, vol. 6, p. 100048, Mar. 2024, doi: 10.1016/j.nlp.2023.100048.

[6] X. Yuan, S. Yuan, Y. Cui, T. Lin, X. Wang, R. Xu, J. Chen, and D. Yang, "Evaluating Character Understanding of Large Language Models via Character Profiling from Fictional Works," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 8015–8036. doi: 10.18653/v1/2024.emnlp-main.456.

[7] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, 2021, pp. 483–498. doi: 10.18653/v1/2021.naacl-main.41.

[8] J. R. K. Suseno, A. E. Minarno, and Y. Azhar, "Implementation of Pretrained VGG16 Model for Rice Leaf Disease Classification using Image Segmentation," *Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control*, Mar. 2023, doi: 10.22219/kinetik.v8i1.1592.

[9] D. M. Anisuzzaman, J. G. Malins, P. A. Friedman, and Z. I. Attia, "Fine-Tuning Large Language Models for Specialized Use Cases," *Mayo Clin. Proc. Digit. Health*, vol. 3, no. 1, p. 100184, Mar. 2025, doi: 10.1016/j.mcpdig.2024.11.005.

[10] S. Zhou, Z. Xu, M. Zhang, C. Xu, Y. Guo, Z. Zhan, Y. Fang, S. Ding, J. Wang, K. Xu, L. Xia, J. Yeung, D. Zha, D. Cai, G. B. Melton, M. Lin, and R. Zhang, "Large language models for disease diagnosis: a scoping review," *Npj Artif. Intell.*, vol. 1, no. 1, p. 9, June 2025, doi: 10.1038/s44387-025-00011-z.

[11] V. Liévin, C. E. Hother, A. G. Motzfeldt, and O. Winther, "Can large language models reason about medical questions?," *Patterns*, vol. 5, no. 3, p. 100943, Mar. 2024, doi: 10.1016/j.patter.2024.100943.

[12] X. Yang, L. Pan, X. Zhao, H. Chen, L. R. Petzold, W. Y. Wang, and W. Cheng, "A Survey on Detection of LLMs-Generated Content," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 9786–9805. doi: 10.18653/v1/2024.findings-emnlp.572.

[13] J. Lai, W. Gan, J. Wu, Z. Qi, and P. S. Yu, "Large language models in law: A survey," *AI Open*, vol. 5, pp. 181–196, 2024, doi: 10.1016/j.aiopen.2024.09.002.

[14] C. Xiao, X. Hu, Z. Liu, C. Tu, and M. Sun, "Lawformer: A pre-trained language model for Chinese legal long documents," *AI Open*, vol. 2, pp. 79–84, 2021, doi: 10.1016/j.aiopen.2021.06.003.

[15] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "GPT-4 Passes the Bar Exam," *SSRN Electron. J.*, 2023, doi: 10.2139/ssrn.4389233.

[16] F. Rivas-Echeverría, L. T. Ramos, J. L. Ibarra, S. Zerpa-Bonillo, S. Arciniegas, and M. Asprino-Salas, "LegalBot-EC: An LLM-Based Chatbot for Legal Assistance in Ecuadorian Law," *IEEE Access*, vol. 13, pp. 106817–106833, 2025, doi: 10.1109/access.2025.3580488.

[17] J. J. Nay, D. Karamardian, S. B. Lawsky, W. Tao, M. Bhat, R. Jain, A. T. Lee, J. H. Choi, and J. Kasai, "Large language models as tax attorneys: a case study in legal capabilities emergence," *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 382, no. 2270, Apr. 2024, doi: 10.1098/rsta.2023.0159.

[18] Y. Wu, C. Wang, E. Gumusel, and X. Liu, "Knowledge-Infused Legal Wisdom: Navigating LLM Consultation through the Lens of Diagnostics and Positive-Unlabeled Reinforcement Learning," in *Findings of the Association for Computational Linguistics ACL 2024*, Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, 2024, pp. 15542–15555. doi: 10.18653/v1/2024.findings-acl.918.

[19] A. Kusumowijoyo, A. Marta, and K. Natali Boasrifa, "The Artificial Intelligence as a One-Stop Point for Dealing with Online Human Trafficking Scams in Indonesia," *J. Sustain. Dev. Regul. Issues JSDERI*, vol.

1, no. 3, pp. 189–211, Sept. 2023, doi: 10.53955/jsderi.v1i3.18.

[20] N. Earlyana and K. L. L. Aung, "LEGAL PROTECTION OF INDONESIAN MIGRANT WORKERS INVOLVED IN ILLEGAL ACTIVITIES OF THE ONLINE SCAMMER SECTOR IN CAMBODIA," vol. 02, no. 01, 2025.

[21] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, H. S. Chan, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.

[22] Y. Tao, O. Viberg, R. S. Baker, and R. F. Kizilcec, "Cultural bias and cultural alignment of large language models," *PNAS Nexus*, vol. 3, no. 9, p. pgae346, Sept. 2024, doi: 10.1093/pnasnexus/pgae346.

[23] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, and N. Aletras, "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 4310–4330. doi: 10.18653/v1/2022.acl-long.297.

[24] G. W. Wicaksono, N. P. Hidayah, C. S. K. Aditya, A. F. P. Dewa, H. Fatikasari, M. A. P. Insani, M. H. F. Anwar, and N. A. Anhari, "Human Trafficking Court Decisions (Indonesia) — Structured Dataset." Mendeley Data, Sept. 15, 2025. doi: 10.17632/8GTBKY7R9X.1.

[25] P. Italiani, G. Moro, and L. Ragazzi, "Enhancing legal question answering with data generation and knowledge distillation from large language models," *Artif. Intell. Law*, July 2025, doi: 10.1007/s10506-025-09463-9.

[26] M. Goyal and Q. H. Mahmoud, "A Systematic Review of Synthetic Data Generation Techniques Using Generative AI," *Electronics*, vol. 13, no. 17, p. 3509, Sept. 2024, doi: 10.3390/electronics13173509.

[27] M. Etcheverry, T. Real-del-Sarte, and P. Chavallard, "Algorithm for Automatic Legislative Text Consolidation," in *Proceedings of the Natural Legal Language Processing Workshop 2024*, Miami, FL, USA: Association for Computational Linguistics, 2024, pp. 166–175. doi: 10.18653/v1/2024.nllp-1.13.

[28] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "RoFormer: Enhanced transformer with Rotary Position Embedding," *Neurocomputing*, vol. 568, p. 127063, Feb. 2024, doi: 10.1016/j.neucom.2023.127063.

[29] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries".

[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001, p. 311. doi: 10.3115/1073083.1073135.

[31] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[32] W. Yuan, G. Neubig, and P. Liu, "BARTSCORE: evaluating generated text as text generation," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, in NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.

[33] C. Ryu, S. Lee, S. Pang, C. Choi, H. Choi, M. Min, and J.-Y. Sohn, "Retrieval-based Evaluation for LLMs: A Case Study in Korean Legal QA," in *Proceedings of the Natural Legal Language Processing Workshop 2023*, D. Preoțiuc-Pietro, C. Goanta, I. Chalkidis, L. Barrett, G. Spanakis, and N. Aletras, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 132–137. doi: 10.18653/v1/2023.nllp-1.13.

[34] Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 Technical Report," Jan. 03, 2025, *arXiv*: arXiv:2412.15115. doi: 10.48550/arXiv.2412.15115.

[35] M. Abdin, "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone," Aug. 30, 2024, *arXiv*: arXiv:2404.14219. doi: 10.48550/arXiv.2404.14219.

[36] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, 2023, pp. 2511–2522. doi: 10.18653/v1/2023.emnlp-main.153.