

Penyusunan Amar Putusan Berbasis *Reinforcement Learning* untuk Meningkatkan *Reasoning* Hukum Pengadilan Indonesia

Mohamad Maulana Firdaus Ramadhan¹, Rizqika Mulia Pratama², Ikhwan Al Hakim³, Rinaldi Munir⁴

¹Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung, Kota Bandung, 40132, email: 18222140@mahasiswa.itb.ac.id

²Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung, Kota Bandung, 40132, email: 13522126@mahasiswa.itb.ac.id

³Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung, Kota Bandung, 40132, email: 13522147@mahasiswa.itb.ac.id

⁴Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung, Kota Bandung, 40132, email: rinaldi@itb.ac.id

Corresponding Author: Mohamad Maulana Firdaus Ramadhan

INTISARI — Penelitian ini membahas pengembangan model generasi amar putusan berbasis *reinforcement learning* untuk meningkatkan *reasoning* hukum dan interpretabilitas hasil generasi pada dokumen putusan pengadilan Indonesia. Model generatif berbasis *supervised learning* seperti *sequence-to-sequence* dan transformer selama ini banyak digunakan, namun seringkali hanya meniru pola teks tanpa melakukan penalaran hukum yang mendalam dan argumentatif. Untuk mengatasi keterbatasan tersebut, penelitian ini mengusulkan pendekatan GRPO atau *Group Relative Preference Optimization* dengan *fine-tuning parameter-efficient* menggunakan LoRA pada Qwen2.5-14B-Instruct. Dataset yang digunakan adalah IndoLaw, yang telah diproses dan diekstraksi bagian-bagian pentingnya, serta dilakukan pencarian pasal relevan untuk setiap kasus. *Reward function* pada *training* mengombinasikan skor BLEU, BERTScore, METEOR, ROUGE-L, serta penalti untuk format dan kelengkapan *reasoning*. Evaluasi dilakukan secara kuantitatif menggunakan metrik BLEU, ROUGE, METEOR, dan BERTScore, serta kualitatif melalui analisis struktur, kesesuaian konteks, dan perbandingan dengan amar putusan asli. Hasil eksperimen menunjukkan bahwa model mampu menghasilkan *reasoning* hukum yang lebih terstruktur, lengkap, dan sesuai format hukum, meskipun masih terdapat beberapa kesalahan seperti halusinasi dan ketidaklengkapan *reasoning*. Penelitian ini membuktikan bahwa pendekatan GRPO dapat meningkatkan kualitas *reasoning* hukum dan interpretabilitas generasi amar putusan, serta memberikan kontribusi baru dalam pengembangan sistem legal AI di Indonesia.

KATA KUNCI — Generasi Amar Putusan, *Reinforcement Learning*, GRPO, *Reasoning* hukum, LoRA.

I. PENDAHULUAN

A. Latar Belakang

Amar putusan merupakan bagian inti dari dokumen putusan pengadilan yang berisi pertimbangan hukum dan alasan logis yang mendasari keputusan hakim atas suatu perkara [1], [2]. Kualitas *reasoning* hukum dalam amar putusan sangat penting untuk menjamin keadilan, transparansi, dan akuntabilitas sistem peradilan [1]. Di Indonesia, jumlah putusan pengadilan yang dipublikasikan terus meningkat secara signifikan. Berdasarkan data Mahkamah Agung Republik Indonesia, hingga saat ini telah tersedia lebih dari 9,9 juta putusan yang dapat diakses publik, dengan lebih dari 86 ribu putusan baru setiap bulannya [3]. Volume data yang sangat besar ini menimbulkan tantangan tersendiri bagi hakim dan praktisi hukum dalam menjaga konsistensi, kualitas, dan efisiensi penyusunan amar putusan.

Seiring berkembangnya teknologi kecerdasan buatan, berbagai metode telah diusulkan untuk mengotomatisasi proses generasi amar putusan, khususnya dengan pendekatan *Natural Language Processing* (NLP) dan *Legal Judgment Prediction* (LJP) [1], [4], [5]. Metode yang paling banyak digunakan adalah model generatif berbasis *supervised learning*, seperti *sequence-to-sequence* (seq2seq), transformer, dan *supervised fine-tuning* (SFT) pada *pretrained language model* [1], [2]. Model-model ini mampu menghasilkan teks amar putusan yang secara sintaksis baik dan mirip dengan data pelatihan.

Namun, berbagai studi menunjukkan bahwa model generatif konvensional tersebut memiliki kelemahan mendasar, yaitu *reasoning* hukum yang dihasilkan cenderung dangkal, tidak argumentatif, dan seringkali hanya meniru pola teks tanpa melakukan penalaran hukum yang mendalam [1], [2]. Selain itu, model ini juga sering gagal memberikan justifikasi hukum yang jelas dan dapat ditelusuri (*traceable*), sehingga hasil generasi sulit dipercaya oleh praktisi hukum [1], [4]. Keterbatasan ini

menjadi hambatan utama dalam adopsi sistem generasi amar putusan otomatis di lingkungan peradilan.

Tren terbaru dalam pengembangan *large language model* (LLM) *reasoning*, seperti Qwen3 [6] dan DeepSeek-R1 [7], menunjukkan bahwa *reinforcement learning* (RL) mampu secara signifikan meningkatkan kemampuan *reasoning* model, baik melalui *training* murni RL maupun *distillation* dari model *reasoning* besar. Qwen3 bahkan mengintegrasikan *thinking* mode dan *thinking* budget untuk *reasoning* adaptif, serta membuktikan bahwa RL dan *distillation* dapat meningkatkan *reasoning* dan generalisasi model, bahkan pada model yang lebih kecil [6]. DeepSeek-R1 membuktikan bahwa RL murni (tanpa SFT) dapat menghasilkan *reasoning* yang *emergent* dan *self-evolving* pada LLM, serta *reasoning pattern* yang dihasilkan dapat didistilasi ke model yang lebih kecil secara efisien [7]. Selain itu, kedua model ini menekankan pentingnya desain *reward*, format *reasoning*, dan mode *switching* untuk aplikasi *reasoning* di dunia nyata, termasuk domain hukum.

Namun, riset tentang penerapan RL, khususnya GRPO (*Group Relative Preference Optimization*), untuk generasi amar putusan masih sangat terbatas. Studi sistematis yang membandingkan efektivitas RL dengan *supervised learning* dalam konteks *reasoning* hukum dan interpretabilitas amar putusan juga masih jarang ditemukan [1], [4], [5]. Oleh karena itu, diperlukan penelitian lebih lanjut untuk mengembangkan dan mengevaluasi model generatif berbasis RL yang mampu menghasilkan amar putusan dengan *reasoning* hukum yang lebih dalam dan justifikasi yang lebih transparan.

Meskipun model bahasa besar (LLM) komersial seperti Gemini 2.5 Pro, Claude 4 Sonnet, dan ChatGPT memiliki kemampuan *reasoning* yang sangat baik, terdapat beberapa

kekurangan utama yang perlu diperhatikan. Model-model ini umumnya tidak dioptimalkan untuk konteks hukum Indonesia, baik dari sisi data pelatihan, struktur dokumen, maupun terminologi yudisial yang khas. Selain itu, sifat closed-source membatasi fleksibilitas adaptasi, seperti penyesuaian reward function, integrasi knowledge base hukum, atau evaluasi sesuai standar nasional. Penggunaan LLM komersial juga menghadapi kendala akses, biaya yang tinggi, serta isu privasi dan keamanan data, terutama jika data sensitif harus diproses di luar infrastruktur lokal. Selain itu, transparansi proses reasoning dan parameter model sering kali terbatas, sehingga sulit untuk diaudit atau disesuaikan lebih lanjut sesuai kebutuhan spesifik di bidang hukum.

Penelitian ini bertujuan untuk mengembangkan model generasi amar putusan berbasis *reinforcement learning*, khususnya dengan pendekatan GRPO, guna meningkatkan kualitas *reasoning* hukum dan interpretabilitas pada hasil generasi amar putusan. Model yang diusulkan akan dievaluasi secara komprehensif terhadap *benchmark reasoning* hukum dan dibandingkan dengan model generatif konvensional berbasis *supervised learning*.

Penelitian ini menawarkan kebaruan dengan mengimplementasikan Group Relative Policy Optimization (GRPO) sebagai pendekatan reinforcement learning (RL) untuk tugas generasi amar putusan, yang sebelumnya belum banyak dieksplorasi pada domain hukum di Indonesia. Berbeda dengan penelitian terdahulu yang umumnya menggunakan supervised learning, model yang diusulkan difokuskan untuk menghasilkan reasoning hukum yang lebih dalam, argumentatif, dan justifikasi hukum yang lebih jelas serta dapat ditelusuri. Selain itu, penelitian ini memberikan kontribusi sebagai berikut:

- Melakukan evaluasi komprehensif pada data putusan pengadilan Indonesia, baik secara kuantitatif maupun kualitatif, untuk mengukur kualitas reasoning hukum dan struktur amar putusan yang dihasilkan model.
- Mengembangkan reward function yang ketat dan sensitif terhadap format serta kelengkapan reasoning hukum, sehingga mendorong model menghasilkan output yang lebih sistematis dan sesuai kaidah hukum.
- Menyediakan analisis perbandingan dengan metode terkini (state-of-the-art), serta mengidentifikasi keunggulan dan kelemahan model dalam menangani kasus nyata.
- Memberikan insight baru untuk pengembangan sistem Legal AI yang lebih transparan, dapat dipercaya, dan siap diadopsi dalam ekosistem peradilan elektronik nasional.

Adapun struktur tulisan ini disusun sebagai berikut. Bagian pertama merupakan pendahuluan yang memuat latar belakang, identifikasi dan rumusan masalah, tujuan, dan batasan penelitian. Bagian kedua membahas penelitian terdahulu. Bagian ketiga menjelaskan metodologi yang diajukan dan teori terkait. Bagian keempat menjelaskan metodologi penelitian, mulai dari lingkungan eksperimen, persiapan data, hingga detail training dan evaluasi. Bagian kelima memaparkan hasil eksperimen dan analisis baik secara kuantitatif maupun kualitatif. Terakhir, bagian keenam berisi Kesimpulan serta saran untuk penelitian mendatang.

B. Identifikasi Masalah

Model generatif berbasis *supervised learning*, seperti *sequence-to-sequence* (seq2seq), transformer, dan *supervised fine-tuning* (SFT), telah banyak digunakan untuk tugas generasi amar putusan otomatis. Namun, berbagai studi menunjukkan

bahwa model-model ini cenderung hanya meniru pola teks dari data pelatihan tanpa melakukan *reasoning* hukum yang mendalam dan argumentatif [1], [2]. Akibatnya, amar putusan yang dihasilkan seringkali dangkal, tidak konsisten dengan logika hukum, dan sulit memberikan justifikasi hukum yang jelas serta dapat ditelusuri [1], [4].

Selain itu, penelitian terbaru menunjukkan bahwa pendekatan *reinforcement learning* (RL), khususnya dengan algoritma seperti GRPO, mampu meningkatkan kemampuan *reasoning* model pada berbagai domain [6], [7]. Namun, riset tentang penerapan RL untuk generasi amar putusan, khususnya dalam konteks *reasoning* hukum dan interpretabilitas, masih sangat terbatas dan belum banyak dieksplorasi pada data putusan pengadilan Indonesia [4], [5].

C. Rumusan Masalah

Berdasarkan identifikasi masalah di atas, rumusan masalah penelitian ini adalah:

1. Bagaimana mengatasi keterbatasan *reasoning* hukum yang dangkal dan kurang argumentatif pada model generatif amar putusan berbasis *supervised learning*?
2. Bagaimana meningkatkan interpretabilitas dan justifikasi hukum pada hasil generasi amar putusan otomatis?
3. Sejauh mana penerapan *reinforcement learning* (GRPO) dapat meningkatkan kualitas *reasoning* hukum dan interpretabilitas pada generasi amar putusan dibandingkan model konvensional?

D. Tujuan

Tujuan dari penelitian ini adalah:

1. Mengembangkan model generasi amar putusan berbasis *reinforcement learning* (GRPO) untuk meningkatkan reasoning hukum dan interpretabilitas hasil generasi.
2. Mengevaluasi kualitas *reasoning* dan justifikasi hukum hasil generasi amar putusan model RL pada data putusan pengadilan Indonesia.
3. Memberikan kontribusi pada pengembangan sistem generasi amar putusan otomatis yang lebih argumentatif, transparan, dan dapat dipercaya untuk mendukung efisiensi dan kualitas sistem peradilan di Indonesia.

E. Batasan

Penelitian ini masih memiliki beberapa keterbatasan yang perlu diperhatikan. Pertama, proses *training* model hanya dilakukan dalam 50 langkah (*steps*), sehingga belum merepresentasikan proses pelatihan yang optimal untuk model bahasa besar. Kedua, jumlah data yang digunakan relatif kecil, yaitu 256 dokumen untuk *training* dan 32 dokumen untuk test, sehingga hasil yang diperoleh belum dapat digeneralisasi untuk seluruh variasi kasus hukum di Indonesia. Ketiga, hingga saat ini belum ada keterlibatan langsung dari praktisi atau ahli hukum dalam proses anotasi label *reasoning* maupun evaluasi kualitatif, sehingga validitas hasil *reasoning* model masih perlu divalidasi lebih lanjut oleh pihak yang berkompeten. Keterbatasan-keterbatasan ini menjadi catatan penting untuk penelitian lanjutan agar hasil yang diperoleh semakin robust dan representatif.

II. PENELITIAN TERDAHULU

Penelitian terkait dalam bidang generasi amar putusan dan *legal judgment prediction* (LJP) telah berkembang pesat dalam beberapa tahun terakhir. Salah satu penelitian awal yang relevan adalah oleh Ye et al. [1], yang memperkenalkan tugas *court view generation* (CVG) untuk meningkatkan interpretabilitas sistem

prediksi dakwaan berbasis model label-conditioned Seq2Seq dengan *attention*. Penelitian ini menyoroti pentingnya *reasoning* hukum yang argumentatif dan *charge-discriminative*, namun model yang diusulkan masih sering menghasilkan *reasoning* yang kurang detail dan tidak selalu terstruktur secara lengkap.

Selanjutnya, Huang et al. [8] mengusulkan pendekatan *dependency learning* berbasis unified text-to-text Transformer untuk LJP. Mereka menunjukkan bahwa dengan membangun *dependency* antar sub-tugas seperti prediksi pasal, dakwaan, dan amar putusan, model dapat menghasilkan *reasoning* yang lebih interpretable dan akurat. Penelitian ini juga menekankan pentingnya *court view* dan konten pasal hukum sebagai penjelasan prediksi model, yang krusial untuk membangun sistem LJP yang dapat dipercaya.

Wu et al. [3] memperkenalkan metode *Attentional and Counterfactual-based Natural Language Generation* (AC-NLG) untuk mengatasi bias data dan meningkatkan kualitas *reasoning* pada *court view generation*. Dengan menggabungkan *claim-aware encoder* dan *counterfactual decoders*, model ini mampu menghasilkan *reasoning* yang lebih robust terhadap *imbalance* data dan *confounding bias*.

III. METODOLOGI

Penelitian ini mengusulkan model generasi amar putusan berbasis *reinforcement learning* (RL) dengan algoritma *Group Relative Policy Optimization* (GRPO) untuk meningkatkan *reasoning* hukum dan interpretabilitas hasil generasi amar putusan. Model generatif berbasis *supervised learning*, seperti *sequence-to-sequence* (seq2seq), transformer, dan *supervised fine-tuning* (SFT) pada *pretrained language* model, telah menjadi standar dalam tugas *court view generation* [1], [4]. Namun, berbagai studi menunjukkan bahwa model ini hanya meniru pola data pelatihan tanpa melakukan *reasoning* hukum yang mendalam [1], [2].

Secara umum, *reinforcement learning* adalah paradigma pembelajaran mesin di mana agen belajar mengambil keputusan optimal melalui interaksi dengan lingkungan, dengan tujuan memaksimalkan *reward* kumulatif [9]. Dalam konteks *natural language processing* (NLP), RL telah banyak digunakan untuk mengoptimalkan model bahasa dalam menghasilkan output yang tidak hanya akurat secara sintaksis, tetapi juga relevan secara semantik dan sesuai dengan tujuan spesifik, seperti *reasoning* hukum atau interpretabilitas [10].

Language model (LM) merupakan model probabilistik yang mempelajari distribusi kata atau token dalam sebuah bahasa, dan digunakan untuk memprediksi urutan kata berikutnya dalam sebuah teks [11]. Perkembangan LM kemudian melahirkan *Large Language Model* (LLM), yaitu model bahasa berskala besar dengan miliaran parameter yang dilatih pada data teks dalam jumlah masif. LLM mampu memahami konteks panjang, menangkap pola semantik yang kompleks, dan menghasilkan teks yang lebih natural serta variatif [12].

Group Relative Policy Optimization (GRPO) adalah algoritma *reinforcement learning* efisien yang mengoptimalkan *policy* model tanpa memerlukan *value* model besar seperti pada PPO. Pada GRPO, *baseline advantage* diestimasi dari *reward* rata-rata dalam satu grup *output*, sehingga *training* menjadi lebih hemat memori dan stabil [8]. Secara matematis, objective GRPO dirumuskan sebagai berikut:

$$L^{\text{GRPO}}(\theta) = E_{(q, \{o_i\}_{i=1}^G) \sim \mathbb{D}} \left[\frac{1}{G} \sum_{i=1}^G \left(\frac{\pi_{\theta}(o_i | q)}{\pi_{\theta_{\text{old}}}(o_i | q)} \cdot \hat{A}_i \right) \right]$$

Dimana

- $L^{\text{GRPO}}(\theta)$: Merupakan fungsi objektif GRPO untuk memperbarui parameter kebijakan (policy) θ .
- $E_{(q, \{o_i\}_{i=1}^G) \sim \mathbb{D}}$: Menunjukkan ekspektasi (nilai harapan) dari *batch* kueri q dan grup-grup output sampel $\{o_i\}_{i=1}^G$ yang bersesuaian, yang diambil dari dataset $\mathbb{D}$.
- $\pi_{\theta}(o_i | q)$: Adalah probabilitas untuk menghasilkan output o_i berdasarkan kueri q di bawah kebijakan saat ini dengan parameter θ .
- $\pi_{\theta_{\text{old}}}(o_i | q)$: Adalah probabilitas untuk menghasilkan output o_i berdasarkan kueri q di bawah kebijakan lama (referensi) dengan parameter θ_{old} .
- $\hat{A}_i$: Adalah *advantage* (keunggulan) relatif grup untuk output ke- i yaitu o_i , yang biasanya dihitung sebagai:

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_1, \dots, r_G\})}{\text{std}(\{r_1, \dots, r_G\}) + \epsilon}$$

Dengan:

- r_i adalah *reward* (imbalance) untuk output o_i .
- $\text{mean}(\{r_1, \dots, r_G\})$ dan $\text{std}(\{r_1, \dots, r_G\})$ adalah rata-rata (*mean*) dan deviasi standar (*standard deviation*) dari *reward* di dalam grup.
- ϵ adalah sebuah konstanta kecil untuk menjaga stabilitas numerik.

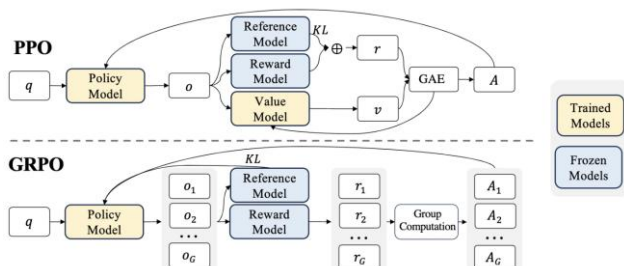
Pendekatan RL, khususnya GRPO, telah terbukti pada model-model mutakhir seperti Qwen3 [6] dan DeepSeek-R1 [7] mampu meningkatkan kemampuan *reasoning*, generalisasi, dan interpretabilitas model, bahkan pada model yang lebih kecil. Perbedaan utama metode yang diusulkan dengan metode terkini adalah: RL (GRPO) mengoptimalkan *reasoning* model berdasarkan *reward* yang dirancang khusus untuk *reasoning* hukum, bukan sekadar meniru pola data; model mampu menghasilkan *reasoning* yang lebih dalam, argumentatif, dan justifikasi hukum yang lebih jelas serta dapat ditelusuri; serta pendekatan ini juga memungkinkan *distillation* ke model yang lebih kecil tanpa kehilangan kemampuan *reasoning* [6], [7].

IV. EKSPERIMEN

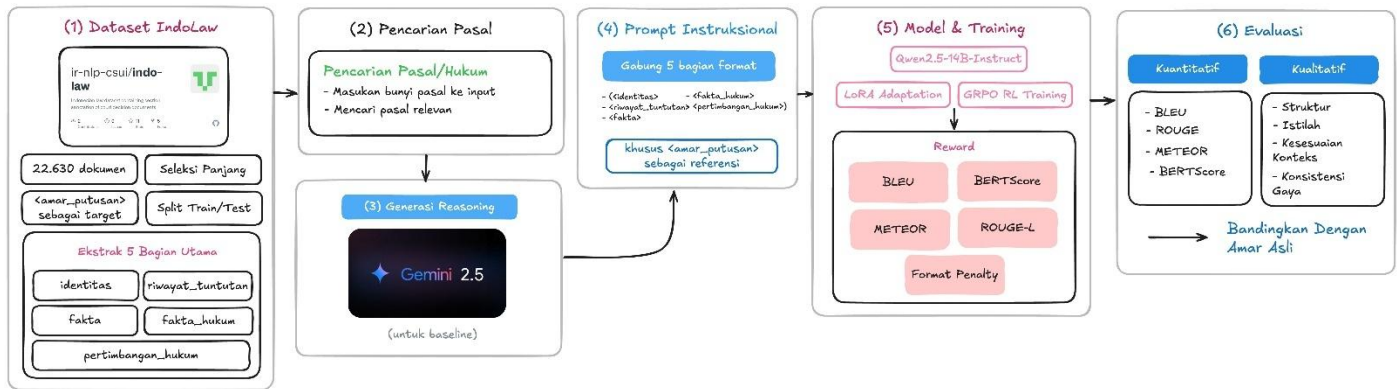
Gambar 2 memperlihatkan alur eksperimen yang dimulai dari persiapan dan filtering dataset, pencarian pasal relevan, pembuatan prompt, baseline Gemini 2.5 Pro, training model Qwen2.5-14B-Instruct dengan LoRA dan GRPO, hingga evaluasi hasil secara kuantitatif dan kualitatif.

A. Lingkungan Eksperimen

Eksperimen dilakukan pada server dengan GPU NVIDIA H100 SXM 80GB, CPU Intel Xeon Platinum 8481C, RAM 235,8 GB dan disk NVMe 64 GB.



Gambar 1. Diagram perbandingan arsitektur PPO dan GRPO.



Gambar 2. Diagram alir eksperimen penelitian. Diagram ini menggambarkan tahapan utama mulai dari persiapan dataset IndoLaw, *filtering* dan ekstraksi bagian penting, pencarian pasal/hukum relevan, pembuatan *prompt* instruksional, proses *baseline* dengan Gemini 2.5 Pro, *training* model Qwen2.5-72B-Instruct dengan LoRA dan GRPO, hingga evaluasi kuantitatif dan kualitatif hasil *reasoning* hukum

B. Dataset dan Persiapan Data

Penelitian ini menggunakan *dataset* IndoLaw [14] yang dapat diakses secara publik melalui repositori di <https://github.com/ir-nlp-csui/indo-law>. *Dataset* IndoLaw terdiri dari 22.630 dokumen putusan pengadilan Indonesia yang telah dianotasi ke dalam 11 bagian struktural dalam format XML, di antaranya <kepala_putusan>, <identitas>, <riwayat_penahanan>, <riwayat_perkara>, <riwayat_tuntutan>, <riwayat_dakwaan>, <fakta>, <fakta_hukum>, <pertimbangan_hukum>, <amar_putusan>, dan <penutup>. Dokumen-dokumen ini merupakan hasil ekstraksi teks bersih dari file PDF Direktori Putusan Mahkamah Agung RI (putusan3.mahkamahagung.go.id), yang telah dibersihkan dan disusun ulang dalam struktur XML untuk memudahkan pemrosesan dan segmentasi bagian [9].

Langkah-langkah pra-pemrosesan data dilakukan sebagai berikut. Pertama, dilakukan *filtering* dokumen untuk memastikan kelengkapan bagian-bagian penting dan validitas struktur XML. Dari total 22.630 dokumen, hanya 6.691 dokumen yang valid dan lengkap secara struktur XML. Selanjutnya, lima bagian utama diekstrak sebagai input model, yaitu <identitas>, <riwayat_tuntutan>, <fakta>, <fakta_hukum>, dan <pertimbangan_hukum>, karena bagian-bagian ini dinilai memberikan informasi substantif dalam pembentukan amar putusan [14].

Seleksi tambahan dilakukan terhadap panjang dokumen, di mana panjang gabungan input dibatasi hingga kurang dari 3.000 kata agar dapat diproses oleh model dengan konteks penuh, dan panjang amar putusan sebagai target dibatasi hingga kurang dari 256 kata. Dokumen yang melebihi batas panjang tersebut diabaikan untuk mencegah pemotongan *input/output* yang dapat menurunkan kualitas prediksi.

Setelah proses seleksi, diperoleh 256 dokumen untuk *training* dan 32 dokumen untuk *test*. Pembagian *dataset* dilakukan secara *random sampling*.

Sebelum proses generasi *reasoning* hukum, dilakukan pencarian pasal atau hukum yang relevan dengan setiap kasus, dan bunyi pasal tersebut dimasukkan ke dalam input agar model dapat melakukan analisis berdasarkan dasar hukum yang tepat. Pendekatan ini sejalan dengan temuan [1] yang menunjukkan bahwa memasukkan label pasal atau dakwaan secara eksplisit ke dalam input dapat meningkatkan interpretabilitas dan ketepatan *reasoning* model dalam menghasilkan *court view* yang *charge-discriminative*. Selain itu, pada penelitian [2] juga menegaskan bahwa penyediaan konten pasal hukum sebagai bagian dari input tidak hanya membantu model dalam menjelaskan prediksi

yang dihasilkan, tetapi juga penting untuk membangun sistem legal *judgment prediction* yang dapat dipercaya dan transparan.

Proses generasi *reasoning* hukum dilakukan dengan memanfaatkan *prompt* instruksional terstruktur, di mana setiap input dokumen hasil konkatenasi lima bagian utama dibungkus dalam format khusus dan bagian <amar_putusan> digunakan sebagai referensi jawaban. Model Gemini 2.5 Pro dari Google Generative AI digunakan untuk menghasilkan *reasoning* hukum, karena kemampuannya memahami konteks panjang dan mendukung Bahasa Indonesia dengan baik.

Untuk proses *training* GRPO, digunakan format *prompt* instruksional yang terdiri dari SYSTEM PROMPT (berisi instruksi peran hakim dan format output yang diharapkan) dan USER PROMPT (berisi deskripsi kasus, pasal terkait, bunyi pasal, dan instruksi analisis).

TABEL 1

SYSTEM DAN USER PROMPT YANG DIGUNAKAN UNTUK PELATIHAN GRPO

SYSTEM_PROMPT = ""

Anda adalah seorang hakim yang profesional dan objektif. Tugas Anda adalah memeriksa dan menganalisis perkara pidana berikut secara teliti, berdasarkan fakta, alat bukti, dan ketentuan hukum yang berlaku.

Lakukan penalaran hukum secara sistematis, uraikan setiap unsur pasal, hubungkan dengan fakta, dan berikan pertimbangan memberatkan/meringankan serta alasan pemaaf/pembenar jika ada.

Akhiri dengan amar putusan yang tegas dan jelas.

Jawablah dalam format berikut:

<reasoning>

Bunyi Pasal:

...

Unsur-unsur:

...

Analisis Unsur:

...

Barang Bukti:

...

Status Barang Bukti:

...

Pertimbangan Pidana:

...

Alasan Pemaaf/Pembenar:

...

Kesimpulan:

```
...
</reasoning>
<answer>
...
</answer>
""""
```

USER_PROMPT = """"

Berikut adalah sebuah perkara pidana yang harus Anda analisis dan putuskan.

Kasus:
{input_text}

Pasal Terkait:
{pasal_terkait}

Bunyi Pasal:
{bunyi_pasal}

Tugas Anda:

- Analisis perkara ini secara teliti dan objektif.
- Hubungkan fakta dengan unsur-unsur pasal yang relevan.
- Berikan penalaran hukum yang logis dan sistematis.
- Sertakan pertimbangan memberatkan/meringankan serta alasan pemaaf/pembenar jika ada.
- Akhiri dengan amar putusan yang tegas dan jelas, sesuai format yang telah ditentukan.

""""

C. Model dan Training

Model yang digunakan dalam penelitian ini adalah Qwen2.5-14B-Instruct, sebuah model bahasa besar *open-source* yang telah dioptimalkan untuk berbagai tugas *reasoning* [10]. Model ini di-load menggunakan library Unsloth dan diadaptasi menggunakan teknik LoRA (*Low-Rank Adaptation*). LoRA merupakan metode *parameter-efficient fine-tuning* yang hanya menambahkan sejumlah kecil parameter baru dalam bentuk matriks low-rank pada layer tertentu, sementara bobot utama model tetap dibekukan. Dengan demikian, LoRA memungkinkan adaptasi model ke task baru secara efisien tanpa perlu melakukan update seluruh parameter model [16].

Training dilakukan menggunakan algoritma *Group Relative Policy Optimization* (GRPO). *Reward function* yang digunakan mengombinasikan beberapa aspek penting, dengan penekanan utama pada skor BLEU untuk memastikan kesesuaian dan ketepatan *reasoning* model terhadap referensi. Selain BLEU, reward juga mempertimbangkan kesamaan semantik menggunakan BERTScore, skor METEOR, dan ROUGE-L, serta kelengkapan format tag dan bagian-bagian *reasoning* yang dihasilkan. Jika *output reasoning* atau *answer* kosong, model dikenai penalti besar, dan penalti tambahan diberikan jika kelengkapan bagian *reasoning* tidak terpenuhi secara penuh. Skema *reward* ini dirancang untuk memastikan setiap *output reasoning* memenuhi struktur dan format yang diharapkan dalam konteks hukum. Konfigurasi *training* yang digunakan meliputi *batch size* 2, *gradient accumulation* 4, *num generation* 4, *learning rate* 5e-6, dan *max steps* sebanyak 50.

D. Teknik Evaluasi

Evaluasi kualitas teks yang dihasilkan oleh model dilakukan melalui pendekatan kuantitatif dan kualitatif. Metrik kuantitatif yang digunakan mencakup:

1. BLEU, yang mengukur kesamaan n-gram antara prediksi dan referensi [17]

Rumus BLEU-N:

$$\text{BLEU-N} = \text{BP} \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

di mana p_n adalah *precision* n-gram, w_n adalah bobot (biasanya sama rata), dan BP adalah *brevity penalty*:

$$\text{BP} = \begin{cases} 1, & \text{Jika } c > r \\ e^{1-\frac{r}{c}}, & \text{Jika } c \leq r \end{cases}$$

dengan c panjang prediksi dan r panjang referensi.

2. ROUGE-L, mengukur recall berdasarkan *Longest Common Subsequence* (LCS) antara prediksi dan referensi [18].

Rumus ROUGE-L F1:

$$\text{ROUGE-L} = \frac{(1 + \beta^2) \cdot R_{LCS} \cdot P_{LCS}}{R_{LCS} + \beta^2 \cdot P_{LCS}}$$

Dengan:

$$R_{LCS} = \frac{\text{LCS}(X, Y)}{\text{len}(Y)}, \quad P_{LCS} = \frac{\text{LCS}(X, Y)}{\text{len}(X)}$$

di mana LCS adalah panjang subsekuensi terpanjang yang sama antara prediksi X dan referensi Y , dan β biasanya 1.

3. METEOR, yang mempertimbangkan sinonim dan stemming untuk memberikan skor yang lebih halus [19]

Rumus METEOR:

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Penalty})$$

dengan:

$$F_{\text{mean}} = \frac{10 \cdot P \cdot R}{9P + R}$$

$$\text{Penalty} = 0.5 \cdot \left(\frac{\# \text{chunks}}{\# \text{matches}} \right)^3$$

di mana P adalah *precision*, R adalah *recall*, $\# \text{chunks}$ adalah jumlah fragmen yang tidak cocok, dan $\# \text{matches}$ adalah jumlah kata yang cocok.

4. BERTScore, yang mengukur kesamaan semantik antar kalimat menggunakan representasi vektor dari BERT [20].

Rumus BERTScore F1:

$$\text{Precision} = \frac{1}{|X|} \sum_{x \in X} \max_{y \in Y} \text{sim}(x, y)$$

$$\text{Recall} = \frac{1}{|Y|} \sum_{y \in Y} \max_{x \in X} \text{sim}(y, x)$$

$$\text{BERTScore F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

di mana $\text{sim}(x, y)$ adalah kemiripan kosinus antara embedding token x dan y , X adalah token prediksi, dan Y adalah token referensi.

Sementara itu, evaluasi kualitatif dilakukan dengan menganalisis struktur naratif, penggunaan istilah yudisial, kesesuaian isi dengan konteks, serta konsistensi gaya bahasa hukum. Selain itu, hasil generasi model dibandingkan secara manual terhadap amar putusan asli untuk menilai keselarasan makna dan mendeteksi potensi halusinasi isi.

V. HASIL DAN ANALISIS EKSPERIMEN

A. Hasil dan Analisis Kuantitatif



Gambar 3. Grafik Hasil Pelatihan GRPO Reward vs Step

Berdasarkan grafik *reward* terhadap *step* pada proses *training*, terlihat bahwa nilai *reward* model cenderung berfluktuasi pada rentang 0.2 hingga 0.5 selama proses pelatihan. Meskipun terdapat beberapa penurunan *reward* yang tajam pada step tertentu, secara umum *reward* dapat kembali meningkat pada step-step berikutnya. Pola fluktuasi ini menunjukkan bahwa proses *training* dengan GRPO dan *reward function* berbasis BLEU dan penalti format mendorong model untuk terus mengeksplorasi dan memperbaiki *output reasoning*-nya. *Reward* yang tidak stabil juga mengindikasikan bahwa model masih menghadapi tantangan dalam menghasilkan *reasoning* yang benar-benar konsisten dan sesuai format pada setiap *batch*, namun tetap ada kecenderungan model mampu mencapai *reward* tinggi secara periodik selama training berlangsung.

TABEL 2
HASIL EVALUASI DATA PENGUJIAN

Metrik	Amar	Reasoning
BLEU	0.077	0.051
METEOR	0.227	0.143
ROUGE-L	0.281	0.206
BERTScore F1	0.454	0.469
Format OK	-	0.594

Tabel rata-rata metrik evaluasi menunjukkan bahwa skor BERTScore F1 dan ROUGE-L relatif lebih tinggi dibandingkan BLEU dan METEOR, baik pada bagian <amar> maupun *reasoning*. Hal ini mengindikasikan bahwa model mampu menghasilkan *output* yang secara semantik cukup relevan dan menangkap sebagian besar informasi penting, meskipun variasi redaksi masih cukup besar sehingga skor BLEU dan METEOR cenderung rendah. Nilai Format OK sekitar 0.59 menunjukkan bahwa lebih dari separuh output model sudah memenuhi format dan struktur yang diharapkan, namun masih terdapat ruang perbaikan untuk konsistensi format pada seluruh hasil generasi.

B. Hasil dan Analisis Kualitatif

Analisis kualitatif menunjukkan bahwa model telah berhasil mempelajari struktur dasar dan gaya bahasa yang khas dalam dokumen putusan hukum. Kemampuan ini terbagi menjadi dua aspek utama: penalaran hukum dan generasi teks amar putusan.

TABEL 3
HASIL GENERASI REASONING MODEL

<reasoning>

Bunyi Pasal:

"Setiap orang yang tanpa hak atau melawan hukum memiliki, menyimpan, menguasai, atau menyediakan Narkotika Golongan I bukan tanaman, dipidana dengan pidana penjara paling singkat 4 (empat) tahun dan paling lama 12 (dua belas) tahun dan pidana denda paling sedikit Rp800.000.000,00 (delapan ratus juta rupiah) dan paling banyak Rp8.000.000.000,00 (delapan miliar rupiah)."

"Unsur-unsur:"

"1. Adanya orang."

"2. Tanpa hak atau melawan hukum."

"3. Memiliki, menyimpan, menguasai, atau menyediakan."

"4. Narkotika Golongan I bukan tanaman."

Analisis Unsur:

1. Adanya Orang:

"Terdakwa Sulistyorina adalah orang yang memiliki identitas yang jelas dan valid, sesuai dengan data yang ada."

2. Tanpa Hak atau Melawan Hukum:

"Terdakwa tidak memiliki izin atau hak dari pihak berwenang untuk menguasai sabu-sabu. Hal ini jelas melanggar hukum sebagaimana disebutkan dalam Pasal 6 Ayat 1 Huruf A UU No. 35 Tahun 2009 tentang Narkotika dan Pasal 8 UU No. 35 Tahun 2009 tentang Narkotika."

3. Memiliki, Menyimpan, Menguasai, atau Menyediakan:

"Terdakwa menguasai sabu-sabu seberat 0,90 gram yang disimpan dalam tas selempang warna hitam merk Eiger. Sabu-sabu tersebut ditemukan saat terdakwa ditangkap oleh petugas kepolisian."

4. Narkotika Golongan I Bukan Tanaman:

"Sabu-sabu yang ditemukan adalah Narkotika Golongan I berdasarkan hasil pemeriksaan Laboratorium Forensik Polri Cabang Surabaya Nomor Lab 2793 NNF 2017."

Barang Bukti:

"1. Satu bungkus plastik klip kecil warna putih berisi sabu-sabu seberat 0,90 gram beserta bungkusnya."

"2. Satu buah HP merk Asus Toyo Type Zenfone 5 warna hitam merah dengan casing warna emas."

"3. Satu buah tas selempang warna hitam merk Eiger."

Status Barang Bukti:

"Semua barang bukti tersebut telah dikenali baik oleh para saksi maupun oleh terdakwa. Dalam perkara ini, barang bukti tersebut dikhawatirkan akan dipergunakan untuk mengulangi tindak pidana tersebut, sehingga perlu dirampas untuk dimusnahkan."

Pertimbangan Pidana:

"Terdakwa telah melakukan tindak pidana yang meresahkan masyarakat dan melanggar hukum. Namun, terdakwa juga memiliki beberapa faktor yang meringankan, seperti mengaku terus terang, bersikap sopan di persidangan, menunjukkan penyesalan, dan berjanji tidak akan mengulangi

perbuatannya. Selain itu, terdakwa masih berusia muda dan mengaku belum pernah dihukum.”

Alasan Pemaaf/Pembenar:

“Terdakwa mengaku bahwa ia tidak menyadari bahwa perbuatannya melanggar hukum dan ia tidak memiliki niat untuk meresahkan masyarakat. Selain itu, terdakwa juga menunjukkan penyesalan dan berjanji untuk tidak mengulangi perbuatannya.”

Kesimpulan:

“Dengan mempertimbangkan semua fakta dan alasan yang ada, Majelis Hakim menyimpulkan bahwa terdakwa Sulistyorina telah terbukti secara sah dan meyakinkan melakukan tindak pidana sebagaimana termuat dalam dakwaan kedua, yaitu tanpa hak atau melawan hukum memiliki, menyimpan, menguasai, atau menyediakan Narkotika Golongan I bukan tanaman.”

</reasoning>

Model menunjukkan upaya awal yang menjanjikan dalam mereplikasi struktur penalaran hukum. Pada hasil keluaran model, seringkali terlihat upaya untuk menghasilkan analisis unsur pasal, mirip dengan yang ditemukan pada reasoning referensi yang menjadi target. Misalnya, pada beberapa kasus, model memulai generasinya dengan mengutip bunyi pasal yang relevan (seperti Pasal 363 KUHP atau Pasal 112 UU Narkotika), kemudian mencoba membedah unsur-unsur pasal tersebut satu per satu, seperti "Unsur barangsiapa", "Unsur mengambil sesuatu barang", dan seterusnya.

Kemampuan ini menandakan bahwa model tidak hanya menghafal urutan kata, tetapi juga mulai memahami bahwa sebuah putusan hukum didasarkan pada analisis unsur-unsur yuridis. Meskipun belum sempurna, pola ini menunjukkan bahwa fondasi untuk menghasilkan teks hukum yang terstruktur dan logis telah terbentuk.

TABEL 4
HASIL GENERASI AMAR PUTUSAN DARI MODEL

<answer> Majelis Hakim memutuskan: “1. Terdakwa Sulistyorina dinyatakan terbukti secara sah dan meyakinkan melakukan tindak pidana sebagaimana termuat dalam dakwaan kedua, yaitu tanpa hak atau melawan hukum memiliki, menyimpan, menguasai, atau menyediakan Narkotika Golongan I bukan tanaman, sebagaimana dimaksud dalam Pasal 112 ayat (1) UU RI No. 35 Tahun 2009 tentang Narkotika.” “2. Terdakwa Sulistyorina dijatuhi hukuman penjara selama 7 (tujuh) tahun dan denda sebesar Rp 800.000.000,00 (delapan ratus juta rupiah), subsidiar 3 (tiga) bulan penjara.” “3. Barang bukti berupa 1 (satu) bungkus plastik klip kecil warna putih berisi sabu-sabu seberat 0,90 gram beserta bungkusnya, 1 (satu) buah HP merk Asus Toyo Type Zenfone 5 warna hitam merah dengan casing warna emas, dan 1 (satu) buah tas selempang warna hitam merk Eiger dirampas untuk dimusnahkan.” “4. Terdakwa Sulistyorina dibebani membayar biaya perkara sebesar Rp 2.000.000,00 (dua juta rupiah).” </answer>
--

Dari sisi generasi teks amar putusan, model secara konsisten menunjukkan kemampuannya dalam menghasilkan struktur yang sesuai dengan kaidah bahasa hukum. Hal ini terlihat dari beberapa keberhasilan:

- Struktur Formal: Sebagian besar output yang valid mengikuti format standar, diawali dengan frasa pembuka “MENGADILI” dan dilanjutkan dengan poin-poin keputusan yang dinomori secara sistematis.
- Kelengkapan Unsur: Model secara umum mampu menyusun amar yang lengkap, mencakup elemen-elemen penting seperti pernyataan bersalah, penjatuhan pidana (lama hukuman dan denda), perintah penahanan, penetapan barang bukti, hingga pembebanan biaya perkara.
- Internalisasi Pola: Model juga berhasil menginternalisasi pola naratif hukum, seperti mampu memasukkan detail jenis narkotika atau barang bukti sesuai konteks yang diberikan pada kasus aslinya.

Keberhasilan ini menunjukkan bahwa model telah mampu menangkap esensi dan format dari sebuah amar putusan, sehingga hasil generasinya sudah cukup kuat untuk menjadi draf awal yang dapat disunting oleh praktisi hukum.

Selain struktur, isi amar yang dihasilkan juga menunjukkan internalisasi model terhadap pola naratif hukum. Misalnya, model mampu memasukkan detail seperti jenis narkotika, masa penahanan, denda, serta frasa-frasa penutup standar seperti “membebaskan biaya perkara” atau “menetapkan agar barang bukti dimusnahkan”. Namun, analisis lebih dalam juga mengungkap beberapa jenis kesalahan yang berulang, yang akan dibahas secara sistematis pada bagian Analisis Kesalahan. Meskipun demikian, dalam konteks praktis, hasil ini sudah cukup kuat untuk menjadi draf awal yang kemudian dapat disunting oleh praktisi hukum.

Gambar 3 memperlihatkan hasil prediksi *court view* model *label-conditioned Seq2Seq* dengan *attention* dari penelitian sebelumnya [16]. Pada model sebelumnya, *reasoning* yang dihasilkan cenderung hanya menyoroti aspek utama dari pasal yang didakwakan, namun seringkali kurang mendetail dan tidak selalu menyajikan analisis unsur secara sistematis.

Struktur *reasoning* yang dihasilkan juga masih sederhana, dengan penjelasan yang terbatas pada satu atau dua aspek utama, sehingga informasi pendukung dan narasi hukum yang lebih komprehensif sering terlewatkan.

Sebaliknya, model yang diusulkan dalam penelitian ini menghasilkan *output reasoning* yang lebih lengkap, sistematis, dan terstruktur. Setiap bagian *reasoning*, mulai dari kutipan bunyi pasal, identifikasi dan analisis unsur-unsur, uraian barang bukti, status barang bukti, hingga pertimbangan pidana dan kesimpulan, disajikan secara eksplisit dan berurutan sesuai format hukum yang diharapkan. Model kami juga mampu mengelaborasi setiap unsur pasal dengan mengaitkan fakta spesifik dari kasus, serta memberikan pertimbangan memberatkan/meringankan dan alasan pemaaf/pembenar secara terpisah. Dengan demikian, hasil generasi model tidak hanya lebih informatif, tetapi juga lebih mudah ditelusuri dan siap digunakan sebagai draf awal oleh praktisi hukum.

C. Analisis Kesalahan

Meskipun menunjukkan kemampuan yang menjanjikan, analisis terhadap hasil keluaran model juga mengidentifikasi beberapa jenis kesalahan yang sistematis, baik dalam penalaran maupun pada teks amar putusan. Berikut adalah beberapa jenis kesalahan yang teridentifikasi dalam aspek penalaran hukum (*legal reasoning*):

1. Pada beberapa kasus, model menghasilkan teks dalam bahasa asing atau kata-kata yang sama sekali tidak relevan dengan konteks hukum. Kegagalan ini

	Target	Prediction		Correct the article
Article	Article 234	Article 114	×	Article 234 <i>Corrected</i>
Charge	Crime of Intentional Injury	Crime of Arson	×	Crime of Intentional Injury ✓
Penalty	8-month imprisonment	36-month imprisonment	<i>Error = 28</i>	6-month imprisonment <i>Error = 2</i>
View	The defendant intentionally burned another person with alcohol and caused a minor injury level. His behavior has constituted the crime of intentional injury. After the defendant voluntarily surrendered to the case, he truthfully confessed his criminal facts. His behavior conformed to the legal characteristics of surrendering and was given a lighter punishment in accordance with the law.	The defendant deliberately set fire, endangering public safety, and has not caused serious consequences. His behavior has constituted the crime of arson. The defendant voluntarily surrendered after committing a crime and truthfully confessed his crime, which is surrendering himself and can be punished lightly in accordance with the law.		The defendant intentionally injured the body of another person, causing minor injuries, and his behavior has constituted the crime of intentional injury. The defendant voluntarily surrendered after committing a crime and truthfully confessed his crime, which is surrendering himself and can be punished lightly in accordance with the law.
Content	Article 114: Whoever commits arson, breaches dikes, causes explosions, spreads pathogen of infectious diseases, poisonous or radioactive substances or other substances, or uses other dangerous means to endanger public security, but causes no serious consequences, shall be sentenced to fixed-term imprisonment of no less than three years but no more than ten years. Article 234: Whoever intentionally injures another person is to be sentenced to not more than 3-year imprisonment.			

Gambar 4. Hasil prediksi *court view* model *label-conditioned Seq2Seq* [16]. Output model sebelumnya cenderung sederhana dan kurang sistematis..

menunjukkan bahwa pada input tertentu, model gagal memahami konteks sepenuhnya dan menghasilkan output yang acak.

2. Penalaran yang Tidak Lengkap atau Terpotong: Seringkali, bagian penalaran yang dihasilkan model berhenti di tengah jalan atau tidak lengkap. Model mungkin berhasil mengutip bunyi pasal, namun gagal saat mencoba menganalisis unsur-unsurnya, menghasilkan kalimat yang terpotong atau analisis yang dangkal.
3. Ketidakesesuaian Pasal: Ditemukan kesalahan fatal dalam penalaran hukum di mana model salah menginterpretasikan konteks kasus. Pada salah satu kasus pencurian, kasus yang diberikan dan *reasoning* referensi merujuk pada Pasal 363 KUHP, namun hasil keluaran model justru menganalisis kasus tersebut menggunakan Pasal 112 UU Narkotika.

Sementara itu, pada hasil akhir teks amar putusan, ditemukan beberapa kesalahan sebagai berikut:

1. Kesalahan Faktual/Numerik: Model seringkali tidak akurat dalam menghasilkan detail kuantitatif. Misalnya, pada beberapa kasus, amar putusan referensi menyatakan hukuman 4 tahun penjara, namun model bisa menghasilkan angka yang berbeda. Ketidakakuratan ini konsisten terjadi pada lama hukuman dan jumlah denda.
2. Kegagalan Struktural: Analisis juga menunjukkan adanya kegagalan struktural, di mana model gagal menghasilkan format amar putusan yang valid. Pada kasus-kasus di mana model menghasilkan teks asing atau tidak relevan, hasil keluarannya tidak dapat dikategorikan sebagai amar putusan yang benar secara format.
3. Over-Simplifikasi Logika Hukum: Ketika dihadapkan pada putusan yang kompleks (misalnya, terdakwa bebas dari dakwaan primer namun terbukti bersalah pada dakwaan subsider), model cenderung menyederhanakannya menjadi satu putusan bersalah yang tunggal, sehingga menghilangkan nuansa dan logika hukum yang penting.

VI. KESIMPULAN DAN SARAN

A. Kesimpulan

Penelitian ini mengusulkan model generasi amar putusan berbasis *reinforcement learning* (GRPO) untuk meningkatkan kualitas *reasoning* hukum dan interpretabilitas hasil generasi amar putusan. Berdasarkan hasil eksperimen, model yang dikembangkan mampu menghasilkan *reasoning* hukum yang

lebih terstruktur, lengkap, dan sesuai format hukum dibandingkan model generatif berbasis *supervised learning* yang selama ini banyak digunakan. Evaluasi kuantitatif menunjukkan bahwa model mampu mencapai skor BERTScore F1 dan ROUGE-L yang cukup tinggi, meskipun skor BLEU dan METEOR masih rendah akibat variasi redaksi. Analisis kualitatif memperlihatkan bahwa model telah mempelajari pola naratif dan struktur bahasa hukum, serta mampu mereplikasi analisis unsur pasal dan amar putusan secara sistematis. Namun, masih ditemukan beberapa kelemahan seperti halusinasi, ketidakesesuaian pasal, dan ketidaklengkapan reasoning pada sebagian output. Secara keseluruhan, penelitian ini membuktikan bahwa pendekatan GRPO dapat menjadi solusi untuk mengatasi keterbatasan *reasoning* hukum yang dangkal dan kurang argumentatif pada model generatif konvensional, serta meningkatkan interpretabilitas dan justifikasi hukum pada hasil generasi amar putusan otomatis.

B. Saran dan Penelitian Mendatang

Beberapa temuan penting dari hasil eksperimen membuka peluang untuk penelitian lanjutan. Pertama, untuk meningkatkan kualitas label *reasoning* dan validitas hasil analisis kualitatif, sangat disarankan melibatkan praktisi hukum secara langsung, baik dalam proses anotasi data *reasoning* maupun dalam evaluasi kualitatif hasil generasi model. Keterlibatan praktisi akan memastikan bahwa label *reasoning* yang digunakan benar-benar merepresentasikan logika dan praktik hukum yang berlaku, serta memberikan insight yang lebih tajam terhadap kekuatan dan kelemahan model dalam konteks nyata.

Kedua, *reward function* yang digunakan dalam *training* model dapat dikembangkan lebih lanjut agar lebih sensitif terhadap aspek-aspek penting dalam amar putusan, seperti ketepatan lama hukuman, besaran denda, dan detail numerik lain yang bersifat krusial dalam konteks hukum. Pengembangan *reward function* yang lebih kontekstual dan numerik akan membantu model untuk tidak hanya menghasilkan *reasoning* yang terstruktur, tetapi juga lebih presisi dalam aspek substansi putusan.

Ketiga, untuk mengurangi kesalahan faktual, halusinasi, dan meningkatkan akurasi *reasoning*, dapat dieksplorasi integrasi sistem *retrieval-augmented generation* (RAG) dan *knowledge base* hukum. Dengan adanya RAG dan *knowledge base*, proses pencarian dan penyisipan pasal atau hukum yang relevan dapat diotomatisasi, sehingga model dapat merujuk langsung pada sumber hukum yang valid dan *up-to-date* dalam setiap generasi *reasoning*. Hal ini juga akan memperkuat kemampuan model dalam menangani kasus-kasus yang lebih kompleks dan variatif.

REFERENSI

- [1] H. Ye, X. Jiang, Z. Luo, dan W. Chao, "Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions," arXiv preprint arXiv:1802.08504, 2018.
- [2] Y. Wu, K. Kuang, Y. Zhang, X. Liu, C. Sun, J. Xiao, Y. Zhuang, L. Si, dan F. Wu, "De-Biased Court's View Generation with Causality," Proc. EMNLP, pp. 763–780, 2020, doi: 10.18653/v1/2020.emnlp-main.56.
- [3] Putusan Mahkamah Agung RI, "Pusat Data Putusan Mahkamah Agung RI," [Online]. Tersedia: <https://putusan3.mahkamahagung.go.id/>, tanggal akses: 26-Jul-2025.
- [4] J. Cui, X. Shen, dan S. Wen, "A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges," IEEE Access, vol. 11, pp. 102050–102071, 2023, doi: 10.1109/ACCESS.2023.3317083.
- [5] N. Z. Dina, S. D. Ravana, dan N. Idris, "Legal Judgment Prediction using Natural Language Processing and Machine Learning Methods: A Systematic Literature Review," SAGE Open, vol. 15, pp. 1–25, 2025, doi: 10.1177/21582440251329663.
- [6] Qwen Team, "Qwen3 Technical Report," arXiv preprint arXiv:2505.09388, 2025.
- [7] DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," arXiv preprint, 2025.
- [8] Y. Huang, X. Shen, C. Li, J. Ge, dan B. Luo, "Dependency Learning for Legal Judgment Prediction with a Unified Text-to-Text Transformer," arXiv preprint arXiv:2112.06370, 2021.
- [9] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, 2018.
- [10] D. Silver, A. Huang, C. J. Maddison, et al., "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–489, 2016.
- [11] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, 2010, pp. 1045–1048.
- [12] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
- [13] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y.K. Li, Y. Wu, dan D. Guo, "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models," arXiv preprint arXiv:2402.03300, 2024.
- [14] E. Q. Nuranti, E. Yulianti, dan H. S. Husin, "Predicting the Category and the Length of Punishment in Indonesian Courts Based on Previous Court Decision Documents," *Computers*, vol. 11, no. 6, p. 88, 2022.
- [15] Qwen Team, "Qwen2.5 Technical Report," arXiv preprint arXiv:2412.15115, 2025.
- [16] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, dan W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.
- [17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. ACL*, 2002, pp. 311–318.
- [18] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Workshop*, 2004, pp. 74–81.
- [19] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop*, 2005, pp. 65–72.
- [20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. ICLR*, 2020.

LAMPIRAN

A. DATASET

a. INDOLAW

Dataset ini dapat diperoleh di:
<https://github.com/ir-nlp-csui/indo-law>

b. DATASET PENELITIAN

Dataset yang telah diolah dapat diperoleh melalui tautan:
<https://github.com/Maull09/GRPO-CVG-Indonesia/tree/main/dataset>

B. TOOLS DAN LIBRARY

- Python 3.12
- PyTorch 2.7.1
- HuggingFace Transformers
- Unsloth
- BERTScore
- NLTK
- rouge-score

C. MODEL DAN CHECKPOINT

Qwen2.5-14B-Instruct:
<https://huggingface.co/Qwen/Qwen2.5-14B-Instruct>

Checkpoint Model hasil *training*:
<https://huggingface.co/maull04/checkpoint-vcg-qwen2.5-14b-instruct-grpo>

D. HARDWARE DAN INFRASTRUKTUR

- GPU: NVIDIA H100 SXM 80GB
- CPU: Intel Xeon Platinum 8481C
- RAM: 235,8 GB
- Disk: NVMe 64 GB
- Sistem Operasi: Ubuntu 22.04 LTS
- Cloud Provider: vast.ai

E. KODE SUMBER DAN PIPELINE EKSPERIMEN

Kode *training*:
https://github.com/Maull09/GRPO-CVG-Indonesia/blob/main/final_notebook.ipynb

Kode *generate reasoning*
https://github.com/Maull09/GRPO-CVG-Indonesia/blob/main/generate_reasoning.py

Kode *split dataset*:
https://github.com/Maull09/GRPO-CVG-Indonesia/blob/main/split_dataset.py

F. HASIL EKSPERIMEN

Log *training*:
<https://github.com/Maull09/GRPO-CVG-Indonesia/blob/main/metrics.csv>

output *reasoning* dan amar putusan:
https://github.com/Maull09/GRPO-CVG-Indonesia/blob/main/final_results.csv