

Building Legal Reward Models for Grounding and Abstention

Rilton Franzone*
University of Oxford

Valentin Noël*
Devoteam

Puyu Wang
University of Oxford

Philip Torr
University of Oxford

Fabio J. Fehr
University of Oxford

Abstract

Large language models are increasingly used in high-stakes domains such as law, where systems must ground their reasoning in retrieved evidence and abstain when that evidence is insufficient. However, existing reward models are largely optimised for general preferences rather than contextual grounding, limiting their ability to evaluate these behaviours in retrieval-augmented generation (RAG) settings. We introduce a framework for transforming existing legal QA datasets into contextual preference data and use it to construct LEGALREWARD-BENCH (LRB), a benchmark for evaluating grounded legal generation under noisy and insufficient retrieval conditions.¹ Across general and legal contextual evaluation, we find that contextual DPO improves grounded evaluation, but performance is sensitive to preference-data construction. Length-balanced augmentation substantially improves grounded legal evaluation, with the strongest configuration combining length-balanced legal and general contextual preference data and improving performance by up to +25.6 pp over baseline. We further find evidence of cross-jurisdiction transfer: models contextually refined primarily on Victorian criminal-law data improve grounded evaluation on external US legal benchmarks, including a +16.2 pp improvement on HOUSING STATUTE QA. Together, these results provide a reproducible foundation for constructing and evaluating grounded legal reward models in retrieval-augmented settings.

1 Introduction

Large language models (LLMs) are increasingly deployed in high-stakes domains such as law, medicine, and finance, where generated outputs may influence legal decisions, clinical care, and access to essential services (Siino et al., 2025; Costa-

Gomes et al., 2026; Chen et al., 2024). Unlike open-domain conversational settings, these domains require reasoning grounded in external evidence and constrained by domain-specific rules. In such settings, useful systems must not only produce correct answers, but also remain faithful to the provided context and refuse to answer when the available evidence is insufficient (Pipitone and Alami, 2024; Peng et al., 2025). However, current language models are primarily optimised for fluent and helpful interaction, often over-relying on parametric knowledge and failing to abstain under uncertainty.

Retrieval-augmented generation (RAG) partially addresses these limitations by providing models with access to external documents and domain-specific evidence (Lewis et al., 2020). However, retrieval alone does not guarantee grounded reasoning. Models may ignore retrieved evidence, rely on memorised parametric knowledge, or generate plausible but unsupported conclusions that are not justified by the provided context (Zheng et al., 2025; Huang et al., 2025). The problem becomes particularly severe when retrieved documents are ambiguous, incomplete, or noisy, as models often continue answering rather than abstaining under uncertainty (Feng et al., 2024; Peng et al., 2025; Butler and Butler, 2026). Grounded generation therefore requires mechanisms that evaluate whether responses faithfully use retrieved evidence.

Legal reasoning further intensifies the challenges of grounded generation. Legal systems are jurisdiction-dependent, procedurally constrained, and require reasoning over precedents, statutes, and factual scenarios (Zhong et al., 2020; Katz et al., 2023; Guha et al., 2023). Answers rarely reduce to purely deductive reasoning, instead requiring contextual interpretation and the weighing of competing legal considerations (Guha et al., 2023; Katz et al., 2024; Pipitone and Alami, 2024; Fan et al., 2026). As a result, unsupported or hallucinated conclusions can appear highly convincing despite

*Equal contribution. Correspondence: Rilton Franzone

¹  Benchmark  Code

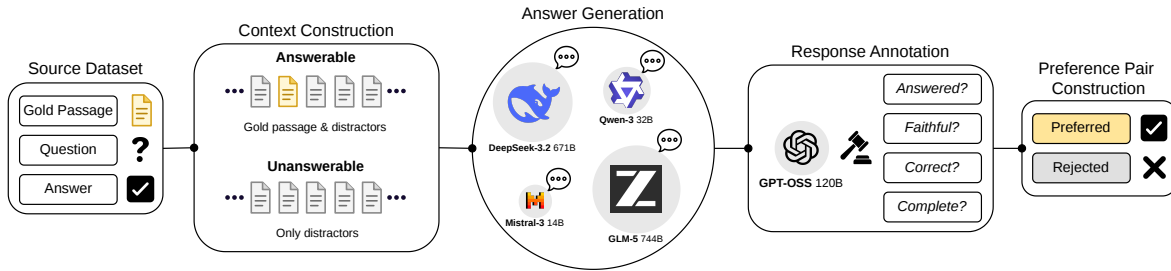


Figure 1: **Legal preference-pair construction framework.** Starting from Legal RAG Bench triples (c_{gold}, q, a) (Section 2.1), we construct six retrieved-context variants by combining the gold passage with distractor passages to control answerability and retrieval noise (Section 2.2). Four LLMs generate candidate responses across all (c, q) instances (Section 2.3). A judge-model assigns pointwise labels for answer behaviour, faithfulness, correctness, and completeness (Section 2.4). A hierarchical preference rule then converts labelled responses into contextual preference pairs (c, q, r^+, r^-) for reward-model training and evaluation (Section 2.5).

lacking justification in the underlying legal sources (Dahl et al., 2024). In such settings, refusal is often preferable to generating confident but unsupported legal reasoning.

Reinforcement learning from human feedback (RLHF) and reward modelling have become central approaches for aligning language models with human preferences (Ouyang et al., 2022; Rafailov et al., 2023; Bai et al., 2022). However, most existing reward models are trained on general-purpose preference data that prioritises helpfulness, fluency, and conversational quality rather than contextual grounding (Wang et al., 2024; Lambert et al., 2025). As a result, they often fail to reliably evaluate whether responses are supported by retrieved evidence or whether a model should abstain when context is insufficient (Xu et al., 2025). Recent work on contextual reward modelling instead evaluates responses relative to retrieved evidence, substantially improving evaluation in retrieval-augmented settings (Xu et al., 2025; Coman et al., 2025). However, reliable reward models and datasets for grounded legal reasoning remain limited.

To address these challenges, we propose a reproducible framework for contextual reward modelling in legal reasoning. We construct a legal reward modelling dataset from existing legal question answering benchmarks by converting $(\text{context}, \text{question}, \text{answer})$ triples into structured preference data capturing answerability, groundedness, completeness, and legal correctness. Rather than treating legal reward modelling as simple domain adaptation, we frame it as the problem of learning contextual grounding under retrieval uncertainty: models must distinguish when retrieved

evidence supports an answer, when it is incomplete or misleading, and when refusal is the preferred behaviour. Across both general and legal retrieval-augmented evaluation settings, we show that contextual reward modelling improves grounded evaluation and refusal behaviour, while also revealing that benchmark construction strongly determines whether these behaviours are learnable and measurable.

We make the following **contributions**:

- **A framework and benchmark for grounded legal reward modelling.** Our framework transforms existing legal QA datasets into contextual preference data, which we use to construct LEGALREWARD BENCH (LRB) for evaluating refusal, faithfulness, completeness, and legal correctness under noisy and insufficient retrieval (Section 2).
- **Legal reward modelling is sensitive to preference-data construction.** We identify response-length asymmetries as a key structural bias and show that length-balanced augmentation and general contextual preference data substantially improve grounded legal evaluation (Sections 4.1–4.2).
- **Contextual grounding can transfer across legal domains and jurisdictions.** Models trained primarily on Victorian criminal law improve grounded evaluation on external US legal tasks, providing evidence that grounding and abstention behaviours are partly transferable (Section 4.3).

2 Methodology

This section describes how we construct LEGALREWARDBENCH, a contextual preference corpus for legal reward models. Starting from a legal QA dataset of (c_{gold}, q, a) triples consisting of a supporting passage, a question, and a reference answer (Section 2.1), we produce preference pairs (c, q, r^+, r^-) where c is a constructed retrieved context and r^+ is more strongly grounded in c than r^- . When c lacks the evidence to answer q , r^+ is an abstention rather than a fabricated answer.

Figure 1 provides an overview of the four-stage framework: Context Construction (Section 2.2), Answer Generation (Section 2.3), Response Annotation (Section 2.4), and Preference Pair Construction (Section 2.5). Together, these stages transform each source example into contextual preference pairs (c, q, r^+, r^-) .

2.1 Source Dataset

We use Legal RAG Bench (Butler and Butler, 2026) as the substrate for dataset construction. Legal RAG Bench is built from the Victorian Criminal Charge Book and contains 100 expert-written questions over a corpus of 4,876 passages. Each question is paired with a supporting passage and a long-form reference answer. The questions were designed to be lexically dissimilar from their supporting evidence, so the task tests semantic retrieval and grounded reasoning rather than keyword matching.

2.2 Context Construction

For each source triple, we construct two kinds of retrieved context. *Answerable* contexts contain the supporting passage alongside nine distractor passages, with the supporting passage placed at a deterministically chosen slot. *Unanswerable* contexts remove the supporting passage and its topic family, replacing them with ten distractors. A passage’s *topic family* consists of passages sharing its leading two-component section number in the Charge Book (e.g. sections 5.7.3 and 5.7.11 share topic family 5.7), and is excluded from distractor sampling in both cases so that distractors cannot reintroduce the supporting evidence indirectly.

We vary the distractor source across three methods: uniform random sampling, BM25 (Trotman et al., 2014), and Nomic semantic retrieval (Nussbaum et al., 2025). Answerable variants test grounded answering under retrieval noise, and unanswerable variants test abstention under insuf-

ficient evidence. This yields six variants (three answerable and three unanswerable), which form the main LEGALREWARDBENCH preference corpus. A distractor ablation extends the set with additional retrieval methods and gold-only baselines (Appendix A.3).

2.3 Answer Generation

We generate candidate responses by prompting four open-weight instruction-tuned models ranging from 14B to 744B parameters² to answer each question using only the retrieved context. Generation uses deterministic decoding with temperature 0.0 and a maximum length of 8,192 tokens, producing four candidate responses per (question, variant) cell, for a total of 2,400 candidate responses across 100 questions and six retrieval variants. Rather than constructing r^- by manually corrupting the reference answer, we let failure modes emerge from the model panel: a candidate may abstain, answer correctly, omit material qualifications, introduce unsupported claims, or reach a conclusion contradicted by the retrieved passages. The reference answer is used only for downstream correctness and completeness labelling, not as a candidate. The generation prompt is in Appendix G.

2.4 Response Annotation

We label each candidate response along four dimensions adapted from CONTEXTUALJUDGE BENCH’s contextual evaluation framework (Xu et al., 2025): answer behaviour, faithfulness, correctness, and completeness. We add correctness because a response can be grounded in the retrieved context yet reach an incorrect legal conclusion, and omit conciseness because it is less directly related to grounding and is often subjective in legal QA.

Annotation proceeds in two stages. We first classify answer behaviour as *attempted* (the response attempts to answer), *abstained* (the response declines due to insufficient or unavailable evidence), or *unusable* (the response is empty, unreadable, incoherent, malformed, or severely repetitive). Only attempted responses are then evaluated for **faithfulness** (*fully supported*, *partially supported*, *unsupported*, or *contradicted*) against the retrieved context, **correctness** (*correct* or *incorrect*) against the reference answer, and **completeness** (*complete* or *incomplete*) based on whether the response con-

²Qwen3 32B (Qwen Team, 2026), DeepSeek V3.2 (Guo et al., 2025), GLM-5 (GLM-5-Team et al., 2026), and Ministral 3 14B (Liu et al., 2026)

Failure Type	Question q	Preferred Response r^+	Rejected Response r^-
Refusal	What sentence was imposed on the accused?	“The provided context is insufficient to answer this question.”	“The accused was sentenced to five years imprisonment.”
Faithfulness	Which Act introduced cyberstalking offences?	“The Crimes (Stalking) Act 2003 introduced cyberstalking offences.”	“The Crimes Amendment Act 2001 introduced cyberstalking offences.”
Correctness	Does Victoria protect against double jeopardy?	“Yes. Victoria protects against double jeopardy.”	“No. Victoria does not recognise double jeopardy protections.”
Completeness	Must jurors be excused after seeing pre-trial publicity?	“No. Judges can usually address prejudice through directions to the jury.”	“No.”

Table 1: **Legal preference pairs.** Representative pairwise preference examples. In the dataset, r^+ and r^- are selected from generated candidate responses after pointwise annotation; they are not manually written reference or corrupted responses. Refusal examples test whether a model abstains when the retrieved evidence is insufficient. Faithfulness, correctness, and completeness examples reflect distinct failures of grounded reasoning. Full examples including retrieved context are provided in Appendix A.4.

tains the essential rule, condition, exception, or qualification needed to justify its conclusion. For faithfulness, the main conclusion governs the label: an unsupported or mischaracterised main conclusion is labelled *unsupported*, while an unsupported secondary claim is labelled *partially supported*.

Labelling is performed by GPT-OSS-120B using structured prompts and deterministic decoding (temperature 0.0, 2,048-token cap), with each dimension evaluated in a separate call. We annotate all 2,400 candidate responses with four independent LLM judges, using GPT-OSS-120B as the primary judge for constructing LEGALREWARD-BENCH. The multi-judge annotations show high agreement: preference pairs reconstructed from the independent annotations preserve 91.7% of GPT-OSS-derived pair directions overall and 94.9% on the test set, with only 1.2% reversing. Full annotation prompts are provided in Appendix G, with dimension-level agreement and preference-level robustness analyses in Appendix B.

2.5 Preference Pair Construction

For each (c, q) instance, we compare every unordered pair of candidate responses using a hierarchical preference rule. A strict preference defines the preferred response r^+ and rejected response r^- , while ties are discarded; representative pairs are shown in Table 1. The hierarchy depends on context answerability. For answerable contexts, responses are compared in order of answer behaviour, faithfulness, correctness, and completeness, with the first differing dimension determining the preference; attempted, fully supported, correct, and complete responses are preferred at each respec-

tive dimension. For unanswerable contexts, only answer behaviour is considered, with abstention preferred over an attempted answer. Applied to 100 questions and six retrieval variants, this yields 3,600 comparisons and 1,220 retained preference pairs. Each retained pair is assigned to the dimension that determines its preference, including answer behaviour, faithfulness, correctness, or completeness, with the resulting distribution reported in Appendix 4.

3 Experimental Setup

Our experiments evaluate grounded legal generation under adversarial retrieval conditions. Rather than benchmarking retrieval quality itself, we study whether models in retrieval-augmented generation (RAG) settings can appropriately abstain when evidence is insufficient, remain faithful when relevant evidence is mixed with distracting context, and avoid unsupported legal conclusions under noisy retrieval conditions.

We evaluate reward models on two contextual preference benchmarks. First, we use CONTEXTUALJUDGE BENCH (CJB) (Xu et al., 2025), a general-domain benchmark for evaluating reward models in retrieval-augmented generation settings. CJB contains 2,000 preference pairs, with 1,600 training pairs and 400 test pairs ($n = 373$ retained after excluding examples exceeding the 4,096-token evaluation limit) covering refusal, faithfulness, completeness, and conciseness under both answerable and unanswerable retrieval conditions. Second, we introduce LEGALREWARD BENCH (LRB), a legal-domain contextual reward modelling benchmark constructed from legal question

answering datasets. LRB contains 1,220 preference pairs split into 940 training, 124 development, and 156 test examples, spanning both general grounding failures and domain-specific Victorian criminal law reasoning. Each example consists of a tuple (c, q, r^+, r^-) , where the rejected response r^- captures failures such as incorrect refusal, hallucination, incompleteness, or unsupported legal conclusions. Full dataset statistics and split details are provided in Appendix A.1.

Reward Model Scoring Let $\mathbf{p} = (p, c, q)$ denote a prompt, retrieved context, and question, and let $r = (r_1, \dots, r_T)$ denote a candidate response of length T . A reward model parameterised by θ assigns a scalar score

$$s_\theta(\mathbf{p}, r) \in \mathbb{R}, \quad (1)$$

where higher values indicate stronger preference for the response. For causal language models, reward scores are commonly computed using the unnormalised sequence log-probability, which sums token log-probabilities across the response. This matches the standard formulation used in DPO and prior reward-model benchmarks (Rafailov et al., 2023; Lambert et al., 2025; Xu et al., 2025). However, because summed log-probabilities scale with response length, this formulation can systematically favour shorter responses when preference pairs exhibit substantial length disparities. We therefore use two complementary scoring functions depending on the benchmark construction.

$$s_\theta(\mathbf{p}, r) = \sum_{t=1}^T \log p_\theta(r_t \mid \mathbf{p}, r_{<t}), \quad (2)$$

where $r_{<t} = (r_1, \dots, r_{t-1})$, matching the convention of prior reward-model benchmarks and the DPO training objective (Rafailov et al., 2023; Lambert et al., 2025; Xu et al., 2025). On LEGALREWARD BENCH (Section 2), where residual length variation in augmented refusal templates would otherwise dominate sum-based comparisons, we use the length-normalised log-probability

$$\ell_\theta(\mathbf{p}, r) = \frac{1}{T} \sum_{t=1}^T \log p_\theta(r_t \mid \mathbf{p}, r_{<t}). \quad (3)$$

Specialist reward models with a scalar reward head use the head output directly in place of s_θ . Given a preference pair (r^+, r^-) , a prediction is considered correct when

$$s_\theta(\mathbf{p}, r^+) > s_\theta(\mathbf{p}, r^-). \quad (4)$$

For an contextual preference pair dataset (c_i, q_i, r_i^+, r_i^-) where $i \in \{1, \dots, N\}$ and prompt p , we report pairwise accuracy

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[s_\theta(\mathbf{p}_i, r_i^+) > s_\theta(\mathbf{p}_i, r_i^-)], \quad (5)$$

where $\mathbf{1}[\cdot]$ denotes the indicator function. Random performance corresponds to 50% accuracy. We report Wilson 95% confidence intervals and overall pairwise accuracy; macro-average across splits is computed in evaluation outputs but omitted from tables for conciseness.

Training Objective We fine-tune models using Direct Preference Optimisation (DPO) (Rafailov et al., 2023), which directly optimises the relative preference margin between preferred r^+ and rejected r^- responses conditioned on the same prompt, context and question. Both training and evaluation rely on the same underlying response scoring function, thus the pairwise accuracy directly measures whether the model assigns higher reward to preferred grounded legal responses. We additionally explored with supervised fine-tuning (SFT) on preference pairs, but found it consistently less effective than DPO (Appendix C.3).

Models We evaluate 14 open-weight instruction-tuned models spanning the 0.5B–27B parameter range, including models from the Qwen (Qwen et al., 2025; Qwen Team, 2026), LLaMA (Grattafiori et al., 2024), Gemma (Team et al., 2024, 2025), Phi (Abdin et al., 2024; Microsoft et al., 2025), Ministral (Liu et al., 2026), DeepSeek-R1-distilled (Guo et al., 2025), SFR-RM-R (Dong et al., 2024), and SmolLM families (Allal et al., 2025). We additionally compare against specialist reward models designed for preference evaluation and LLM-as-a-judge tasks, including Skywork-Reward (Liu et al., 2024), Selene (Alexandru et al., 2025), and SFR-Judge (Wang et al., 2025). For fine-tuning for contextual reward modelling, we fine-tune a representative subset of models spanning different architectures and scales: Phi4-mini (Microsoft et al., 2025), DeepSeek-R1-8B (Guo et al., 2025), Qwen3.5-9B (Qwen Team, 2026), Qwen3.5-27B (Qwen Team, 2026), and Ministral-8B (Liu et al., 2026). All models were evaluated under identical hardware conditions. Full implementation and compute details are provided in Appendix E.1.

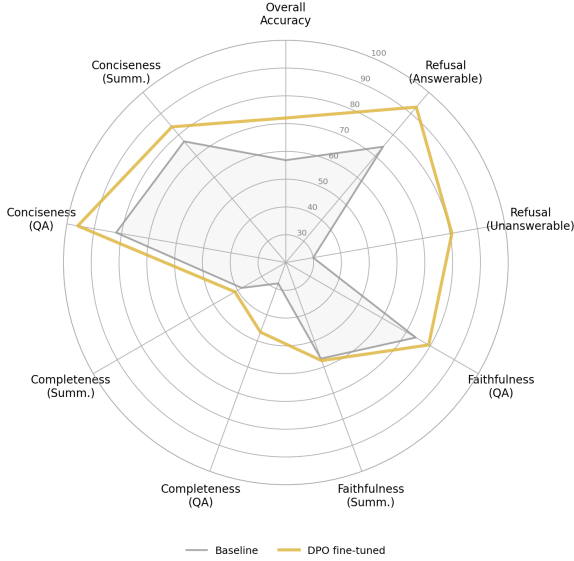


Figure 2: **DPO refinement improves grounded evaluation on CONTEXTUALJUDGE BENCH.** Compared to the Ministral-8B baseline (grey), contextual DPO refinement on CJB (gold) improves performance across the three contextual dimensions we focus on: refusal, faithfulness, and completeness.

4 Results

4.1 Contextual Evaluation

We first evaluate 14 open-weight instruction-tuned models together with specialist reward models on CONTEXTUALJUDGE BENCH (CJB), with full per-dimension results reported in Table 11 (Appendix C.1). Baseline performance is only marginally above random chance and inconsistent across grounding dimensions, with most models favouring shorter responses and consequently performing better on conciseness than on completeness and refusal. We therefore select a representative subset of models for contextual DPO refinement, including Ministral-8B, which achieves the strongest balance across faithfulness, completeness, and refusal among the evaluated baselines, consistent with RAGfreee (Coman et al., 2025). Figure 2 shows that contextual DPO improves grounded evaluation across these dimensions, with the largest improvement in refusal under insufficient retrieval.

4.2 Legal Contextual Reward Modelling

We next evaluate contextual reward modelling in the legal domain using LEGALREWARD BENCH (LRB). Table 2 shows that DPO refinement on CJB improves LRB performance for both Gemma 2-2B and Ministral-8B, despite using no legal-domain preference data during training. CJB-only refine-

Table 2: **Legal contextual reward modelling is sensitive to dataset construction.** Evaluation on LEGAL-REWARD BENCH (LRB). Length-balanced augmentation and general grounded pairs improves grounded legal evaluation, particularly for baseline Ministral-8B. Higher is better. Wilson 95% confidence intervals are provided in Appendix - Table 12. Best results are shown in **bold**; second-best results are underlined.

Training	Gemma 2-2B		Ministral-8B	
	Overall	Δ	Overall	Δ
Random	50.0	–	50.0	–
Baseline	61.8	–	59.3	–
CJB	66.4	+4.6	67.6	+8.3
LRB	63.6	+1.8	61.9	+2.6
CJB+LRB	65.7	+3.9	66.0	+6.7
LRB-Length-Aug	81.1	+19.3	81.8	+22.5
CJB+LRB-Length-Aug	81.4	+19.6	84.9	+25.6

ment also outperforms LRB-only refinement, suggesting that grounded contextual evaluation and refusal behaviour are partly domain-general capabilities that transfer across retrieval-augmented settings.

Table 2 further shows that adding the original legal preference data does not consistently improve over CJB-only refinement, indicating that performance depends on preference-pair structure rather than data volume alone. Preference datasets constructed from naturally generated responses can exhibit systematic response-length asymmetries, particularly in refusal settings where abstentions and substantive answers differ substantially in length. Under sequence log-probability scoring, these asymmetries can systematically favour responses based on length rather than the intended contextual preference, confounding the evaluation of grounded behaviour.

To isolate the effect of preference-pair structure, we introduce a length-balanced augmentation strategy (Appendix G) that rewrites refusal and negative responses to better match the length and structure of preferred grounded answers while preserving the underlying grounding failure. Figure 3 and Table 2 show that length-balanced augmentation reduces response-length asymmetries and substantially improves grounded legal evaluation, with the strongest configuration improving Ministral-8B by +25.6 pp over baseline. These results suggest that legal reward modelling is sensitive to preference-data construction, rather than data volume alone.

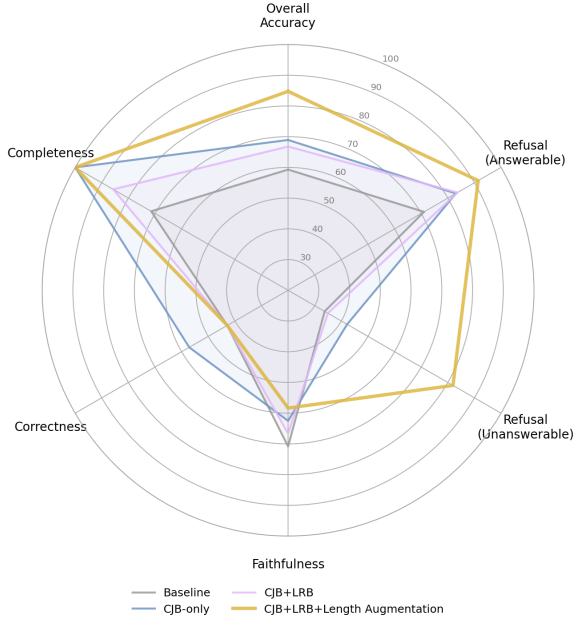


Figure 3: **Length-balanced augmentation improves legal contextual reward modelling.** Compared to the Ministral-8B baseline (grey), DPO refinement on CJB alone transfers strongly to LEGALREWARD BENCH (blue). Joint training on CJB and LRB (pink) is limited by structural response-length artefacts in the legal preference data. After applying length-balanced augmentation (gold), performance improves substantially across evaluation dimensions.

4.3 Cross-Jurisdiction Transfer

We evaluate cross-jurisdiction transfer on two US legal reasoning benchmarks from the Stanford RegLab suite (Zheng et al., 2025): HOUSING STATUTE QA, consisting of yes/no questions grounded in US state housing statutes, and BAR EXAM QA, consisting of Multistate Bar Examination questions paired with supporting US caselaw passages. Both differ substantially in jurisdiction and task from the Victorian criminal-law data used during training, providing an out-of-distribution evaluation of legal contextual reward modelling. Table 3 shows positive but uneven transfer: DPO-refined Ministral-8B improves by +16.2pp on HOUSING STATUTE QA and +4.3pp on BAR EXAM QA.

To better understand cross-jurisdiction transfer, we analyse performance by task. On HOUSING STATUTE QA, baseline Ministral-8B almost always selects “Yes”, achieving 100% accuracy on “Yes” questions but only 0.3% on “No” questions, for 50.2% balanced accuracy. After contextual DPO, accuracy reaches 60.7% and 74.2%, respectively, raising balanced accuracy to 67.4% and indi-

Table 3: **Contextual reward models partially generalise across legal jurisdictions.** Models refined on CJB and LRB improve grounded evaluation on external US legal reasoning QA benchmarks despite training primarily on Victorian criminal law data. Results report pairwise reward accuracy with Wilson 95% confidence intervals. Higher is better. Best results are shown in **bold**; second-best results are underlined.

Model	Bar Exam	Δ	Housing	Δ
Random	50.0	–	50.0	–
Ministral-8B	56.4 [47,65]	–	54.2 [50,58]	–
+DPO	<u>60.7</u> [52,69]	+4.3	70.4 [66,74]	+16.2
Qwen3.5-9B	54.7 [46,63]	–	52.8 [48,57]	–
+DPO	54.7 [46,63]	+0.0	<u>55.4</u> [51,60]	+2.6
Qwen3.5-27B	60.7 [52,69]	–	50.8 [46,55]	–
+DPO	61.5 [52,70]	+0.8	53.0 [49,57]	+2.2

cating better discrimination between legal conclusions. Transfer is more limited on BAR EXAM QA, where contextual DPO produces only modest improvements. Compared with the statute-grounded yes/no setting of HOUSING STATUTE QA, Bar Exam questions involve more complex legal scenarios and competing conclusions, consistent with contextual grounding transferring more readily than jurisdiction-specific legal reasoning.

5 Discussion

Our results suggest that contextual preference optimisation can improve grounded legal evaluation beyond the domains used for training. Across general and legal evaluation settings (Sections 4.1–4.2), contextual refinement improves the ability of reward models to distinguish grounded responses under noisy and insufficient retrieval. Cross-jurisdiction experiments (Section 4.3) provide further evidence of transfer, although gains vary across tasks and models. Stronger gains on statute-grounded HOUSING STATUTE QA, together with more limited transfer on BAR EXAM QA, suggest that contextual grounding may transfer more readily than jurisdiction-specific legal reasoning. This distinction is important in legal settings, where models must distinguish conclusions supported by the available evidence from cases in which abstention is appropriate. The gains observed for relatively small open models further suggest that these behaviours can be strengthened without relying on frontier-scale systems.

Our experiments also suggest that legal reward modelling is sensitive to preference-data construc-

tion (Section 4.2). Naturally generated preference pairs can contain response-length asymmetries: preferred abstentions are often shorter than unsupported answers, while complete legal answers may be longer than incomplete ones. Under sequence log-probability scoring, length can therefore confound the intended preference, obscuring whether models learn grounding and abstention or simply favour responses of particular lengths. Length-balanced augmentation substantially improves grounded legal evaluation, highlighting the importance of benchmarks and reward objectives that disentangle contextual quality from response length.

6 Related Work

Legal Reasoning and Legal Language Models

Recent advances in large language models have increased interest in legal AI (Devlin et al., 2019; OpenAI et al., 2024). Legal reasoning remains challenging due to jurisdiction-specific interpretation, long-context reasoning, and the requirement that conclusions remain grounded in authoritative sources (Katz et al., 2023; Zhong et al., 2020). Legal NLP has evolved from classification and judgement prediction toward large-scale reasoning benchmarks such as LexGLUE, LegalBench, LEXam, and PLAWBench (Chalkidis et al., 2022; Guha et al., 2023; Fan et al., 2026; Shi et al., 2026). However, recent evaluations of frontier systems continue to reveal hallucinated authorities, unsupported conclusions, and unreliable grounding despite strong general reasoning performance (Dahl et al., 2024; Katz et al., 2024).

Retrieval-Augmented Generation in Law

Retrieval-augmented generation (RAG) enables language models to condition responses on retrieved evidence rather than parametric memory alone (Lewis et al., 2020). This paradigm is particularly important in law, where relevant authority is distributed across statutes, regulations, and case law. Recent work has introduced legal-specific retrieval and RAG benchmarks, including LegalBench-RAG and Legal RAG Bench (Pipitone and Alami, 2024; Butler and Butler, 2026). Although retrieval improves access to relevant legal evidence, current systems frequently remain unfaithful to retrieved context, hallucinate unsupported authorities, and fail to abstain under insufficient evidence (Butler and Butler, 2026; Peng et al., 2025).

Reward Modelling and LLM-as-a-Judge Reward modelling and RLHF are widely used to align language models with human preferences (Ouyang et al., 2022; Rafailov et al., 2023). Benchmarks such as RewardBench and JudgeBench evaluate reward models across reasoning and factuality tasks (Lambert et al., 2025; Tan et al., 2024). However, most existing reward models remain largely non-contextual, prioritising helpfulness and stylistic preference rather than evaluating whether responses are grounded in retrieved evidence. Recent work on contextual evaluation begins to address this limitation. CONTEXTUALJUDGE BENCH introduces evaluation for grounded generation under retrieval uncertainty, while RAGferee demonstrates that contextual reward models outperform general-purpose judges in RAG settings (Xu et al., 2025; Coman et al., 2025). Nevertheless, grounded reward modelling for legal reasoning remains comparatively underexplored, particularly for open and compute-efficient models.

7 Conclusion

We introduced a framework for transforming legal QA resources into contextual preference data and used it to construct LEGALREWARD BENCH (LRB) for grounded legal reward modelling under noisy and insufficient retrieval. Across general and legal settings, contextual preference optimisation improves grounded evaluation and appropriate abstention. Our experiments further show that preference-data construction matters: response-length asymmetries can interact with sequence scoring and confound the contextual behaviours reward models are intended to learn. Cross-jurisdiction results provide evidence that contextual grounding can transfer beyond the legal domains used for training, although these gains vary across tasks and models. Together, these findings highlight the need to treat preference-data construction and reward scoring as important considerations in grounded reward modelling, alongside model training itself, we provide a reproducible foundation for developing and evaluating grounded legal reward models.

Future work should extend this framework with direct legal-expert validation and evaluate whether improvements in pairwise reward modelling translate to grounded generation in end-to-end legal RAG systems. Broader evaluation across jurisdictions, languages, and legal tasks will establish how broadly contextual grounding behaviours transfer.

Limitations

Our approach retains legal expertise through expert-curated questions, legal sources, supporting evidence, and reference answers, while using multiple LLM judges to scale response-level annotation. The judges therefore assess generated responses against existing legal supervision rather than establish the underlying legal ground truth. Although our multi-judge analysis shows that preference construction is largely robust across annotators, agreement among LLM judges does not replace direct validation by legal experts. Expert evaluation of a representative subset is therefore an important direction for further validation.

Our experiments focus on pairwise reward evaluation, isolating whether models prefer grounded responses over plausible but unsupported alternatives. We do not evaluate whether the refined models themselves generate better open-ended legal responses, nor do we model the iterative retrieval and reasoning processes of end-to-end legal RAG systems. Extending contextual reward modelling to generation and more complex retrieval workflows is a natural next step.

Finally, LEGALREWARDBENCH is constructed primarily from Victorian criminal law, with cross-jurisdiction transfer evaluated on two US legal tasks. While these experiments provide evidence that contextual grounding can transfer beyond the source jurisdiction, broader evaluation across legal systems, languages, and tasks is needed to establish the extent of this generalisation. Our hierarchical preference rule also represents one operationalisation of grounded legal behaviour; alternative formulations may capture different trade-offs between answering, abstention, faithfulness, correctness, and completeness.

References

- Marah Abidin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Björck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *Preprint*, arxiv:2404.14219 [cs.CL].
- Andrei Alexandru, Antonia Calvi, Henry Broomfield, Jackson Golden, Kyle Dai, Mathias Leys, Maurice Burger, Max Bartolo, Roman Engeler, Sashank Pisupati, Toby Drane, and Young Sun Park. 2025. [Atla selene mini: A general purpose evaluation model](#). *Preprint*, arXiv:2501.17195.
- Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. [SmolLM2: When smol goes big – data-centric training of a small language model](#). *Preprint*, arxiv:2502.02737 [cs.CL]. Version: 1.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Abdur-Rahman Butler and Umar Butler. 2026. [Legal rag bench: an end-to-end benchmark for legal rag](#). *Preprint*, arXiv:2603.01710.
- Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. [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)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.
- Zhiyu Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Ruth Petzold, and William Yang Wang. 2024. [A survey on large language models for critical societal domains: Finance, healthcare, and law](#). *Transactions on Machine Learning Research*. Survey Certification.
- Andrei C. Coman, Ionut-Teodor Sorodoc, Leonardo F. R. Ribeiro, James Henderson, Bill Byrne, and Adrià de Gispert. 2025. [RAGfree: Building contextual reward models for retrieval-augmented generation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8164–8211, Suzhou, China. Association for Computational Linguistics.
- Beatriz Costa-Gomes, Pavel Tolmachev, Eloise Taysom, Viknesh Sounderajah, Hannah Richardson, Philipp Schoenegger, Xiaoxuan Liu, Matthew M Nour, Seth Spielman, Samuel F Way, and 1 others. 2026. Public use of a generalist llm chatbot for health queries. *Nature Health*, pages 1–8.
- Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. [Large legal fictions: Profiling legal hallucinations in large language models](#). *Journal of Legal Analysis*, 16(1):64–93.
- Tim Dettmers, Artidoro Pagnoni, Ari Fanush, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. [Rlhf workflow: From reward modeling to online rlhf](#). *Preprint*, arXiv:2405.07863.
- Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, Daniel Brunner, Markus Leippold, Mrinmaya Sachan, Alexander Stremitzer, Christoph Engel, Elliott Ash, and Joel Niklaus. 2026. [LEXam: Benchmarking legal reasoning on 340 law exams](#). In *The Fourteenth International Conference on Learning Representations*.
- Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. [Don’t hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.
- GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, and 168 others. 2026. [Glm-5: from vibe coding to agentic engineering](#). *Preprint*, arXiv:2602.15763.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Re, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. [Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- International Energy Agency. 2023. [World Energy Outlook 2023](#). Technical report, International Energy Agency, Paris.
- Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. [Gpt-4 passes the bar exam](#). *Philosophical Transactions of the Royal Society A*, 382(2276).
- Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J. Bommarito II. 2023. [Natural language processing in the legal domain](#). *Preprint*, arXiv:2302.12039.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. [RewardBench: Evaluating reward models for language modeling](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyachchi, Baptiste Bout, and 101 others. 2026. [Ministral 3](#). *Preprint*, arxiv:2601.08584 [cs.CL].
- Chris Yuhao Liu, Liang Zeng, Jiakai Liu, Yangguang Li, Dehao Zhang, Yuxiang Liu, Peng Liu, and Xiaoming Shi. 2024. Skywork-Reward: Bag of tricks for reward modeling in LLMs. *arXiv preprint arXiv:2410.18451*.
- Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen,

- Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yiling Chen, Qi Dai, Xiyang Dai, and 56 others. 2025. [Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs](#). *Preprint*, arxiv:2503.01743 [cs.CL].
- Zach Nussbaum, John Xavier Morris, Andriy Mulyar, and Brandon Duderstadt. 2025. [Nomic embed: Training a reproducible long context text embedder](#). *Transactions on Machine Learning Research*. Reproducibility Certification.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Xiangyu Peng, Prafulla Kumar Choubey, Caiming Xiong, and Chien-Sheng Wu. 2025. [Unanswerability evaluation for retrieval augmented generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8452–8472, Vienna, Austria. Association for Computational Linguistics.
- Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arxiv:2412.15115 [cs.CL].
- Qwen Team. 2026. [Qwen3.5: Towards native multimodal agents](#).
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Yuzhen Shi, Huanghai Liu, Yiran Hu, Gaojie Song, Xinran Xu, Yubo Ma, Tianyi Tang, Li Zhang, Qingjing Chen, Di Feng, and 1 others. 2026. Plaw-bench: A rubric-based benchmark for evaluating llms in real-world legal practice. *arXiv preprint arXiv:2601.16669*.
- Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*, 13:18253–18276.
- Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. 2024. [Judgebench: A benchmark for evaluating llm-based judges](#). In *The Thirteenth International Conference on Learning Representations*.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arxiv:2503.19786 [cs.CL].
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arxiv:2408.00118 [cs.CL].
- Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. [Improvements to bm25 and language models examined](#). *Proceedings of the 19th Australasian Document Computing Symposium*.
- PeiFeng Wang, Austin Xu, Yilun Zhou, Caiming Xiong, and Shafiq Joty. 2025. [Direct judgement preference optimization](#).
- Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024. [Helpsteer 2: Open-source dataset for training top-performing reward models](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Austin Xu, Srikanth Bansal, Yifei Ming, Semih Yavuz, and Shafiq Joty. 2025. [Does context matter? Contextual-JudgeBench for evaluating LLM-based judges in contextual settings](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9541–9564, Vienna, Austria. Association for Computational Linguistics.
- Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. 2025. [A reasoning-focused legal retrieval benchmark](#). In *Proceedings of the*

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. [How does NLP benefit legal system: A summary of legal artificial intelligence](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

Contributions

Rilton Franzone and Valentin Noël contributed equally to this work. Rilton Franzone led the dataset development and methodological design, while Valentin Noël led the experimental implementation and evaluation. Puyu Wang assisted with experimental setup, manuscript preparation, and figure design. Fabio J. Fehr conceived and led the project, overseeing its design, execution, and overall direction.

A Dataset and Corpus Construction

A.1 Dataset Statistics

LegalRewardBench Construction

The source set contains 100 Legal RAG Bench questions (Butler and Butler, 2026) over 4,876 Victorian Criminal Charge Book passages. For the main training corpus we use six retrieval-context variants: three answerable variants containing the gold passage plus distractors (random, BM25, and Nomic), and three unanswerable variants containing only distractors after excluding the gold passage and its topic family (random, BM25, and Nomic). With four candidate generators, this gives $100 \times 6 \times 4 = 2,400$ candidate responses and 3,600 within-cell pair comparisons.

The distractor ablation uses the same 100 questions and four generators but extends the context set to twelve variants by adding the gold-only base context, gold-surrounding context, Contriever and GloVe hard negatives, and gold-plus-Contriever/GloVe distractor contexts. This gives $100 \times 12 \times 4 = 4,800$ labelled candidate responses.

ContextualJudgeBench

CJB contains 2,000 preference pairs across refusal, faithfulness, completeness, and conciseness dimensions. We use the official 80/20 split: 1,600 training pairs and 400 test pairs, with 373 test pairs retained after excluding examples exceeding the

Table 4: Main LEGALREWARDBENCH preference pair counts by split and first distinguishing hierarchy dimension. The 80/10/10 split is performed at question level.

Dimension	Train	Val	Test	Total
refusal_answerable	451	68	91	610
refusal_unanswerable	264	36	47	347
faithfulness	145	12	12	169
correctness	49	3	4	56
completeness	31	5	2	38
Total	940	124	156	1,220

Table 5: Strict pair yield by corpus. “Main six” is the preference corpus used for training; “all variants” is the context-variation ablation.

First distinguishing dimension	Main six	All variants
refusal_answerable	610	1,443
refusal_unanswerable	347	564
faithfulness	169	330
correctness	56	147
completeness	38	113
Strict pairs	1,220	2,597
Labelled responses	2,400	4,800

4,096-token evaluation limit. CJB is not legal-domain data; it supplies a complementary general contextual-grounding signal.

Transfer Benchmarks

Bar Exam QA. We use 117 Multistate Bar Exam multiple-choice questions from the Stanford Reglab suite (Zheng et al., 2025), each paired with a gold US-caselaw passage. For pairwise evaluation, the chosen response is the correct answer and the rejected response is a randomly sampled incorrect answer.

Housing Statute QA. We use 500 yes/no questions on US state housing law from the Stanford Reglab suite (Zheng et al., 2025), each paired with a gold statute excerpt. The chosen response is the conclusion supported by the statute; the rejected response gives the opposite conclusion.

A.2 Corpus Construction Details

Source dataset. We use Legal RAG Bench (Butler and Butler, 2026), a grounded legal QA benchmark built from the Victorian Criminal Charge Book containing 100 expert-authored triples (c, q, a) over a corpus of 4,876 legal passages. Questions are intentionally lexically dissimilar from their supporting evidence, stressing semantic understanding and grounded reasoning.

Retrieval-context variants. All constructed contexts contain ten passages except the base gold-only condition, which contains the original supporting passage alone. The main corpus uses six variants: gold_plus_random_distractors, gold_plus_bm25_distractors, gold_plus_nomic_distractors, random_context, bm25_hard_negative, and nomic_hard_negative. The context-variation ablation adds base, gold_surrounding, gold_plus_contriever_distractors, gold_plus_glove_distractors, contriever_hard_negative, and glove_hard_negative.

Pair construction. Each candidate response is labelled under the hierarchy answer behaviour $\succ$ faithfulness $\succ$ correctness $\succ$ completeness. Within each question-variant cell, all six unordered pairs among the four model responses are compared. Ties under the hierarchy are discarded; strict comparisons become preference pairs.

A.3 Context-Variant Ablation

We construct twelve retrieval-context variants beyond the main six to isolate how retrieval conditions affect grounded answering. The variants fall into four families:

- **Base gold:** only the supporting passage is shown.
- **Gold + surrounding:** the supporting passage plus its neighbouring Charge Book sections.
- **Gold + distractors:** the supporting passage plus nine distractors retrieved by random sampling, BM25, Contriever, GloVe, or Nomic.
- **Hard negatives:** the supporting passage and its topic family are removed; ten passages are retrieved by random sampling, BM25, Contriever, GloVe, or Nomic.

The first two families isolate generation quality when all relevant evidence is present; the gold-plus-distractor family tests answering under noise; the hard-negative family tests abstention under plausible-but-insufficient evidence.

Answerable variants. Table 6 reports the seven answerable variants. The target behaviour is to attempt an answer grounded in the retrieved passages, so we report *ideal answer rate* (attempted,

fully supported, correct, and complete) alongside *attempted rate* for context.

Table 6: Answerable context-variant ablation. Each row has 400 labelled responses. Ideal answer means attempted, fully supported, correct, and complete.

Variant	Ideal	Attempted
Base gold	67.2	81.2
Gold + surrounding	64.5	79.0
Gold + random	64.0	78.5
Gold + BM25	61.8	75.0
Gold + Contriever	57.5	72.8
Gold + GloVe	54.5	70.2
Gold + Nomic	57.5	75.0

Base gold sets the ceiling at 67.2% ideal. Surrounding context and random distractors cost only two to three points (64.5% and 64.0%); both add material the model can largely ignore. BM25 distractors drop the rate further (61.8%), and semantic distractors are hardest — Contriever and Nomic both land at 57.5%, GloVe at 54.5%. Attempted rate tracks ideal rate closely across all variants, so most of the degradation is in the quality of attempted answers rather than in whether the model attempts at all.

Unanswerable variants. Table 7 reports the five unanswerable variants. Here the supporting passage and its topic family have been removed; the target behaviour is to recognise insufficient evidence and refuse.

Table 7: Unanswerable context-variant ablation. Each row has 400 labelled responses.

Variant	Correct refusal
Random context	95.0
BM25 hard negative	74.0
Contriever hard negative	76.8
GloVe hard negative	90.0
Nomic hard negative	73.8

Refusal rates split sharply by retrieval method. Off-topic distractors are easy to refuse (random contexts 95.0%, GloVe hard negatives 90.0%) because the retrieved passages are obviously unrelated. The hard cases are BM25 (74.0%), Nomic (73.8%), and Contriever (76.8%): these methods surface passages topically related to the question but lacking the supporting evidence, drawing models into unsupported attempts.

Overall. Across all twelve variants, the four generators produce an ideal-response rate of 69.7%: 61.0% on answerable contexts and 81.9% correct refusal on unanswerable contexts. Two patterns hold across the ablation. On answerable contexts,

degradation is gradual and increases with distractor semanticity: random and surrounding distractors preserve quality, lexical distractors drop it modestly, semantic distractors most. On unanswerable contexts, refusal is largely determined by whether the hard negatives are topically plausible: off-topic random or GloVe negatives are well-rejected, while BM25 and dense semantic negatives are confusable.

Length asymmetry in strict pairs. As a side observation, Table 8 reports per-pair token counts for strict preference pairs, broken down by behaviour transition. Refusal pairs carry a large template-length asymmetry, since abstentions are short formulaic refusals while attempted answers contain reasoning. Attempted-vs-attempted pairs from answerable contexts are length-balanced.

Table 8: Length diagnostics for the 1,220 strict preference pairs in the main corpus. Token counts use the c1100k_base tokenizer. The answerable attempted-attempted row is length-balanced, while refusal pairs (in both directions) carry a large template-length asymmetry.

Answerability	Transition	n	r^+ tok.	r^- tok.
answerable	attempted > abstained	610	93.9	10.6
answerable	attempted > attempted	263	76.6	78.4
unanswerable	abstained > attempted	347	11.9	104.3

A.4 Examples

Preference pairs are selected from model-generated candidates within the same retrieved context and question. For example, on the base context for legal-rag-bench-0001, the question asks whether a judge is required to excuse a juror who knows the accused. The retrieved passage states that the court *may* excuse a potential juror if satisfied that the person cannot consider the case impartially. The hierarchy therefore prefers the generated answer “No. The provided context states that the court may excuse a potential juror ... but it does not state that the court is required to do so” over the generated refusal “The provided context is insufficient to answer this question.” This is a refusal-answerable pair: both candidates see the same evidence, but the preferred response correctly attempts a grounded answer.

For an unanswerable variant, the gold passage and topic-family passages are removed before generation. In that setting, an abstention is preferred over a substantive answer even when the answer is fluent, because the response is unsupported by the

retrieved context shown to the model. Faithfulness, correctness, and completeness pairs are created only among attempted answers after higher-priority labels are tied.

B Annotation Validation

B.1 Multi-Judge Agreement

To assess whether annotations depend on the choice of GPT-OSS-120B as the primary judge, we re-annotate all 2,400 candidate responses using three additional independent judges: Llama4 Maverick³, Gemma4 31B⁴, and Nemotron3 120B⁵. We measure agreement across the four judges for answer behaviour, faithfulness, correctness, and completeness.

Dimension	N	Chance	Raw	Fleiss’ κ
Answer behaviour	2,400	50.1%	99.9%	0.998
Faithfulness	1,139	74.7%	84.3%	0.379
Correctness	1,139	74.7%	94.2%	0.771
Completeness	1,139	69.4%	91.5%	0.722

Table 9: **Agreement across LLM judges.** Chance and raw agreement are reported as percentages. Chance agreement reflects expected agreement from the marginal label distributions, while raw agreement is the average observed agreement across all six judge pairs.

The semantic dimensions are evaluated on the 1,139 responses that attempted an answer. Agreement is nearly perfect for answer behaviour and substantial for correctness and completeness. Faithfulness has lower Fleiss’ κ despite 84.3% raw agreement. Faithfulness requires distinguishing whether response claims are fully supported, partially supported, unsupported, or contradicted by the retrieved context, and 86% of responses are labelled fully supported, increasing expected chance agreement and reducing κ .

We additionally propagate the independent annotations through the same hierarchical preference-construction procedure described in Section 2.5 to assess whether annotation disagreements alter the resulting preference supervision.

Most annotation disagreements do not change the preference supervision used for training and evaluation: 91.7% of pair directions are preserved overall and 94.9% on the test set, while only 1.2% of pairs reverse direction. This indicates that the

³Llama-4-Maverick-17B-128E-Instruct

⁴Gemma-4-31B-it

⁵NVIDIA-Nemotron-3-Super-120B-A12B-NVFP4

Preference reconstruction	Result
Overall pair directions preserved	1,119 / 1,220 (91.7%)
Test-set pair directions preserved	148 / 156 (94.9%)
Pair reversals	15 / 1,220 (1.2%)
Answer-behaviour pair directions preserved	99.8%

Table 10: **Robustness of preference construction across LLM judges.** Agreement is measured after propagating independent annotations through the same hierarchical preference-construction procedure used to construct LEGAL-REWARD BENCH.

resulting preference pairs are largely robust to the choice of a single LLM judge.

C Additional Results and Analyses

C.1 Detailed CONTEXTUALJUDGE BENCH Results

Table 11 reports per-dimension pairwise reward accuracy on the CONTEXTUALJUDGE BENCH (CJB) test set for all evaluated models. We compare three categories of systems: (i) open-source models evaluated using sequence log-probability scoring, (ii) specialist reward models with dedicated scalar reward heads, and (iii) DPO fine-tuned contextual judge models trained on the combined LEGALREWARD BENCH and CJB preference corpus. All models were evaluated in full bfloat16 precision on NVIDIA H100 80GB GPUs. Results are reported across refusal, faithfulness, completeness, and conciseness dimensions for both question answering and summarisation settings. Random performance corresponds to 50% pairwise accuracy.

Takeaway. Most open-weight models and specialist reward models perform only marginally above random chance overall. In particular, many models exhibit a strong bias toward shorter responses, leading to high performance on conciseness dimensions but substantially weaker performance on completeness. We further observe that models frequently prefer attempting an answer over abstaining (refusal), even when the provided context is insufficient. In contrast, DPO fine-tuning consistently improves grounded evaluation performance across most categories, with the fine-tuned Ministral-8B model achieving the strongest overall results.

C.2 Detailed LegalRewardBench Results

C.3 Supervised Fine-Tuning vs. DPO

A natural question is whether the gains in the main results come from contextual exposure to the legal preference data alone, which a simpler super-

vised fine-tuning (SFT) recipe would also provide, or specifically from the contrastive DPO objective that optimises the margin between preferred and rejected responses. To answer this we train a matched SFT baseline on identical data, hyperparameters, LoRA configuration, optimiser, and evaluation pipeline to the DPO run. The only difference is the training objective.

Training procedure. The same 2,540-pair combined corpus is used for both adapters, CONTEXTUALJUDGE BENCH (1,600 pairs) plus length-augmented LEGALREWARD BENCH (940 pairs). Each preference triple $(\mathbf{p}, r^+, r^-)$ is reduced to a single causal-language-model example by concatenating the prompt with the chosen response, $\mathbf{x} = \mathbf{p} \oplus r^+$; the rejected response r^- is discarded. We train with TRL’s SFTTrainer under standard token-level cross-entropy

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{|\mathbf{x}|} \sum_{t=1}^{|\mathbf{x}|} \log p_{\theta}(x_t | x_{<t}),$$

with no reward modelling, no preference margin, and no reference model. Hyperparameters (LoRA $r=32$, $\alpha=64$ on the four attention projections; learning rate 2×10^{-5} ; warmup ratio 0.1; 3 epochs; effective batch size 16 in bfloat16) are identical to the DPO run.

Result. Table 13 reports both adapters on the held-out evaluation sets used for SFT comparability ($n = 373$ for CJB, $n = 156$ for LRB-v2). SFT lifts the raw model by only +3.5 pp on CJB and +1.9 pp on the augmented LRB-v2 test, while DPO under the identical compute budget achieves +15.9 pp and +35.3 pp respectively. The contrastive signal accounts for an order-of-magnitude larger improvement than imitation of the preferred response alone.

Mechanism. Reward accuracy at evaluation is the indicator $\mathbf{1}[\log p_{\theta}(r^+ | \mathbf{p}) > \log p_{\theta}(r^- |$

Table 11: **Open-weight and specialist reward models perform poorly, while DPO fine-tuning improves grounded evaluation.** Per-dimension pairwise reward accuracy on the CONTEXTUALJUDGE BENCH (CJB) test set (~ 400 preference pairs). Results are grouped into open-weight models, specialist reward models with scalar reward heads, and DPO fine-tuned contextual judge models. Dimensions evaluate completeness (C), conciseness (Cc), faithfulness (F), and refusal under answerable (R-A) and unanswerable (R-U) retrieval settings. QA denotes question answering subsets and S denotes summarisation subsets. Random performance corresponds to 50% pairwise accuracy. Higher is better. Best scores in each column are shown in **bold**; second-best scores are underlined.

Model	Overall	C-QA	Cc-QA	F-QA	C-S	Cc-S	F-S	R-A	R-U
Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
<i>Baselines</i>									
Qwen2.5-0.5B	55.4	24.0	88.0	66.0	36.2	71.8	45.5	84.1	32.0
SmolLM2-1.7B	57.1	30.0	86.0	68.0	34.0	74.4	47.7	88.4	34.0
Gemma3-1B	58.1	30.0	90.0	72.0	38.3	65.9	50.0	86.0	36.0
Llama3.2-1B	55.9	24.0	88.0	68.0	29.8	79.5	47.7	84.1	32.0
Gemma2-2B	<u>61.9</u>	34.0	92.0	78.0	38.3	75.6	56.8	93.0	32.0
Qwen3.5-2B	<u>58.6</u>	34.0	86.0	68.0	36.2	76.9	56.8	86.4	30.0
Llama3.2-3B	58.0	30.0	90.0	74.0	38.3	66.7	52.3	84.1	32.0
Qwen2.5-3B	58.6	32.0	86.0	68.0	38.3	76.9	52.3	84.1	36.0
Phi4-mini	58.2	32.0	84.0	72.0	48.9	63.4	56.8	81.8	30.0
Phi3.5-mini	58.8	30.0	88.0	74.0	40.4	74.4	53.5	85.7	30.0
Ministral-8B	56.8	28.0	82.0	74.0	38.3	76.9	56.8	74.4	30.0
Llama3.1-8B	59.6	40.0	88.0	72.0	40.4	64.1	59.1	88.6	28.0
DeepSeek-R1-8B	58.3	38.0	92.0	66.0	36.2	71.8	45.5	90.5	32.0
SFR-8B	62.0	40.0	88.0	74.0	42.6	76.9	59.1	88.6	32.0
Qwen3.5-9B	54.3	35.3	78.6	70.5	34.4	92.6	45.8	81.8	20.0
<i>Specialist reward models</i>									
Skywork-Reward-8B	46.3	50.0	37.3	66.0	46.0	38.8	60.0	70.0	2.0
Selene-1-Mini-8B	<u>60.4</u>	42.0	88.0	76.0	38.3	64.1	56.8	88.6	32.0
SFR-Judge-8B	60.8	58.0	98.0	56.0	31.9	85.0	52.3	56.8	50.0
<i>Our fine-tuned models</i>									
DeepSeek-R1-8B	61.8	40.0	96.0	74.0	34.0	79.5	40.9	92.9	42.0
Phi4-mini	65.4	36.0	92.0	74.0	46.8	73.2	59.1	90.9	54.0
Qwen3.5-9B	67.9	46.0	94.0	72.0	36.2	84.6	61.4	93.2	60.0
Qwen3.5-27B	67.4	40.0	92.0	76.0	44.7	82.1	56.8	93.2	58.0
Gemma 2-2B	69.1	42.0	96.0	76.0	38.3	82.9	59.1	93.0	68.0
SFR-8B	<u>70.9</u>	44.0	96.0	78.0	42.6	87.2	61.4	93.2	68.0
Ministral-8B	72.7	50.0	96.0	80.0	44.7	84.6	56.8	93.0	78.0

Table 12: **Wilson 95% confidence intervals for Table 2.** Pairwise reward accuracy on LEGALREWARD BENCH-V2 val+test ($n=280$). Gemma 2-2B values are from a single training seed; Ministral-8B CIs correspond to the seed=42 run (main table reports 3-seed means).

Training	Gemma 2-2B		Ministral-8B	
	Overall	95% CI	Overall	95% CI
Baseline	61.8	[56.0, 67.3]	59.3	[53.4, 64.9]
CJB	66.4	[60.7, 71.7]	66.8	[61.1, 72.0]
LRB	63.6	[57.8, 69.0]	60.7	[54.9, 66.3]
CJB+LRB	65.7	[60.0, 71.0]	65.4	[59.6, 70.7]
LRB-Length-Aug	81.1	[76.1, 85.2]	81.8	[76.8, 85.9]
CJB+LRB-Length-Aug	81.4	[76.5, 85.6]	85.0	[80.4, 88.7]

$\mathbf{p})$], a function of the *margin* between two log-probabilities. SFT maximises $\log p_{\theta}(r^{+} | \mathbf{p})$ in isolation; nothing in its objective discourages the model from also raising $\log p_{\theta}(r^{-} | \mathbf{p})$ on closely-related (e.g. hallucinated, ungrounded, or template-style refusal) text it has been exposed to during pre-

training. DPO directly optimises the log-margin, the same quantity scored at evaluation, and so cleanly aligns the training objective with the metric. This alignment, rather than additional exposure to in-domain data, is what we believe accounts for the gap.

Table 13: SFT vs. DPO on Ministral-8B, identical combined corpus and training budget. Evaluated on the CJB test split ($n = 373$) and the LRB-v2 test split ($n = 156$).

Eval set	Raw	SFT	DPO	$\Delta(\text{DPO}-\text{SFT})$
CJB ($n = 373$)	56.8	60.3	72.7	+12.4 pp
LRB-v2 test ($n = 156$)	56.4	58.3	91.7	+33.4 pp

C.4 Multi-Seed Per-Run Detail

Table 14: Ministral-8B DPO-Combined-v2 (CJB+LRB-v2) across three random seeds on the CJB test split ($n = 373$). Wilson 95% CI in brackets. Unnormalised sum log-probability scoring per the convention used in the main results table.

Seed	CJB reward accuracy (95% CI)
42	72.7 [67.9, 76.9]
1	71.6 [66.8, 75.9]
2	71.9 [67.1, 76.2]
Mean $\pm \sigma$	72.0 $\pm$ 0.6

C.5 Qualitative Error Analysis

To characterise remaining failure modes after DPO fine-tuning, we manually reviewed 40 pairs where the best-performing adapter (Ministral-8B DPO-Combined) incorrectly preferred the rejected response on the CJB test set. We identify three dominant failure categories.

Category 1: Completeness failures (55% of errors). The most frequent error occurs on completeness pairs where the chosen response enumerates multiple legal sub-issues (statutes, exceptions, qualifications) while the rejected response states only the top-level conclusion. Even after DPO, the model assigns higher log-probability to the shorter, less complete response. This is consistent with the persistent below-chance completeness accuracy (0.40–0.52) in Table 11: the DPO gradient on completeness pairs is diluted by the 25:1 imbalance between refusal and completeness pairs in the combined corpus. A targeted upsampling strategy or a separate completeness-specific reward head would likely address this.

Category 2: Long-context faithfulness (28% of errors). On faithfulness pairs where both chosen and rejected responses are long (> 200 tokens), DPO-trained models occasionally prefer the unfaithful response. Inspection reveals that the unfaithful response is often fluent and closely mirrors

the structure of the retrieved context, making the log-probability gap small. The mean reward margin on these pairs is close to zero (margin < 5 nats), indicating low-confidence predictions rather than systematic miscalibration.

Category 3: Cross-dimension confounds (17% of errors). Some pairs that are labelled as one dimension (e.g., faithfulness) also exhibit a secondary signal from another dimension (e.g., the unfaithful response happens to be shorter). The model’s log-probability is influenced by both signals simultaneously, producing an error on the primary label. These cases suggest that future preference pair construction should explicitly control for cross-dimension confounds at sampling time.

D Training and Evaluation Details

D.1 Training and Reproducibility Details

Training hyperparameters. All DPO and SFT runs use LoRA with rank $r = 32$, $\alpha = 64$, dropout 0.05 on the four attention projections (Q,K,V,O). Hyperparameters: $\beta = 0.1$ (DPO), learning rate 2×10^{-5} , 3 epochs, batch 4 with gradient accumulation 4 (effective batch 16), default max_length 1,024 tokens. Qwen3.5-27B uses 4-bit NF4 QLoRA (Dettmers et al., 2023) to fit single-GPU memory; all other models train in full bfloat16. All runs use a single NVIDIA H100 80GB HBM3.

Statistical methodology. Wilson 95% CIs use the standard score-test interval. McNemar’s test operates on 2×2 contingency tables of per-pair correctness vectors (raw vs DPO), reported as the two-sided exact p -value. Bootstrap CIs use 10,000 resamples with seed 42.

Released artefacts. We release: (1) the augmentation prompts (Appendix G) and the LEGALREWARD BENCH-v2 training and test files; (2) all trained DPO adapters on HuggingFace Hub; (3) the evaluation harness with per-pair JSONL outputs supporting Wilson CI, mean margin, macro and length-controlled metrics; (4) seed/run scripts to reproduce every table in this paper. Full software and hardware configuration details are in Appendix E.2; compute budget and CO₂ estimates are in Appendix E.1; corpus statistics and evaluation set composition in Appendix A.1; hyperparameter sensitivity analysis in Appendix D.2; statistical methodology details in Appendix D.3; qualitative

error analysis in Appendix C.5; and broader impact considerations in Appendix F.

D.2 Hyperparameter Sensitivity

Table 15 reports the hyperparameter sensitivity sweep performed on the initial DPO-Combined (CJB+LRB) corpus. The optimal defaults (shaded row) are inherited unchanged for the CJB+LRB-v2 headline run; we did not re-run the sweep on the augmented corpus because the training procedure is identical. All ablations use seed 42; all other hyperparameters held at their defaults.

Table 15: Hyperparameter sensitivity, Ministral-8B DPO-Combined on the CJB test split. Shaded row = default configuration carried forward to the DPO-Combined-v2 model. At these defaults, DPO-Combined-v2 reaches 72.7 at seed 42 and 72.0 ± 0.6 across three seeds (cf. Table 14).

Hyperparameter	Value	CJB Acc.
DPO β	0.01	70.1
	0.05	71.8
	0.10	73.7
	0.20	72.9
LoRA rank r	8	71.0
	16	72.2
	32	73.7
Learning rate	5×10^{-6}	70.5
	2×10^{-5}	73.7
	5×10^{-5}	72.1
	1×10^{-4}	69.4
Training epochs	1	69.8
	2	72.5
	3	73.7

Observations. $\beta = 0.1$ is the standard DPO default and performs best; lower β (0.01) produces excessive forgetting of the base model’s instruction-following abilities, visible as a drop on conciseness dimensions. LoRA rank $r = 32$ provides the best trade-off between adapter capacity and the risk of overfitting on 2,540 pairs; higher ranks ($r = 64$, not shown) offer no improvement. The learning rate is sensitive: 5×10^{-5} causes moderate gradient instability on refusal pairs (grad-norm spikes), while 5×10^{-6} converges too slowly to reach a good solution in 3 epochs. Three epochs is optimal; a fourth epoch shows marginal over-fitting (-0.5 pp) on a held-out 10% validation split.

D.3 Statistical Methodology Details

Wilson score confidence interval. For a proportion estimate $\hat{p} = k/n$ (where k is the number of

correct predictions and n is the total pairs), the Wilson 95% CI bounds are:

$$p_{lo} = \frac{\hat{p} + \frac{z^2}{2n} - z \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}},$$

$$p_{hi} = \frac{\hat{p} + \frac{z^2}{2n} + z \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}},$$

where $z = 1.96$ for a 95% interval. Unlike the Wald interval, the Wilson interval has correct coverage for p near 0 or 1 and small n , which matters for sub-splits with as few as 47 pairs (refusal_unanswerable on LRB test).

McNemar’s test. Given per-pair binary correctness vectors **a** (model A) and **b** (model B), McNemar’s test uses the discordant counts b_{01} (A wrong, B right) and b_{10} (A right, B wrong). The exact two-sided p -value is computed via the binomial distribution with $n = b_{01} + b_{10}$ trials and $p_0 = 0.5$. We use the exact test rather than the χ^2 approximation since $b_{01} + b_{10}$ can be small on sub-splits. All reported p -values are two-sided.

Bootstrap confidence intervals. Bootstrap CIs for the mean reward margin are constructed by re-sampling with replacement from the per-pair margins, computing the mean of each resample, and reporting the 2.5th and 97.5th percentiles of the bootstrap distribution ($B = 10,000$ resamples, seed 42).

Statistical power. On the CJB test set ($n = 373$), a McNemar test for detecting a 10 pp improvement (from $p = 0.50$ to $p = 0.60$ at the pair level) has power $> 99\%$ at $\alpha = 0.05$, computed via the normal approximation to the discordant-cell binomial. On the LEGALREWARD BENCH test ($n = 156$), power for the same effect size is $\approx 90\%$. For the length-controlled CJB subset ($n = 280$), power is $> 95\%$. These figures confirm that our observed gains are not limited by sample size; the reported McNemar $p < 10^{-4}$ margins are well within the detectable range.

E Compute and Environment

E.1 Compute Budget and Environmental Impact

All experiments were conducted on a single NVIDIA H100 80 GB HBM3 SXM5 GPU (TDP 700 W peak). Table 16 summarises wall-clock time and estimated GPU-hours per experimental phase.

Energy and carbon. Total energy consumption: $88\text{ h} \times 0.70\text{ kW} = 61.6\text{ kWh}$. At the IEA 2023 global average grid carbon intensity of $0.49\text{ kg CO}_2/\text{kWh}$ (International Energy Agency, 2023), total estimated emissions are $\approx 30\text{ kg CO}_2\text{e}$, roughly equivalent to a 200 km passenger-car journey. No cloud pre-emption or repeated failed runs occurred; the above counts completed runs only. All training used bfloat16 mixed precision; QLoRA further reduces the 27B model’s memory and therefore active compute by approximately $4\times$ vs. full bfloat16.

Inference cost at deployment. The trained LoRA adapters add no additional memory overhead during evaluation beyond the base model ($<0.1\%$ of parameters for $r=32$). Evaluation throughput on the H100: Ministral-8B scores approximately 52 preference pairs per minute at `max_length = 4,096`; Qwen3.5-27B QLoRA scores approximately 14 pairs per minute at the same limit. For a 500-pair benchmark (Housing Statute QA), Ministral-8B evaluation completes in under 10 minutes.

E.2 Software and Hardware Configuration

Hardware. Training and evaluation ran inside a RunPod container with one NVIDIA H100 SXM 80 GB HBM3 GPU, 20 vCPUs (Intel Xeon Platinum 8468), 188 GB RAM, a 20 GB container disk, and a 100 GB persistent network drive.

Software dependencies. Table 17 lists the primary Python packages and their exact versions used in this work.

Determinism. All Python, NumPy, and PyTorch random seeds are fixed to the run seed (0, 1, or 2) at the start of each training script. DataLoader worker seeds are pinned via a `worker_init_fn`. CUDA deterministic mode (`torch.use_deterministic_algorithms(True)`) is enabled for all evaluation runs and is disabled for training to preserve throughput, which is why three seeds are reported.

F Broader Impact and Ethical Considerations

Intended use and scope. This work develops contextual reward models for evaluating legal RAG responses, not for providing legal advice. The systems studied are designed to *score* responses against a retrieval context, a task distinct from legal reasoning or legal decision-making. The trained adapters are released for research purposes under the terms of their respective base model licences.

Jurisdictional and domain bias. The LEGAL-REWARD BENCH training data is derived from Legal RAG Bench and is drawn entirely from Victorian (Australian) criminal law. Preference pairs reflect the structure, terminology, and precedent hierarchy of that jurisdiction. Transfer to other jurisdictions (US, EU, Asia-Pacific) may be impaired by differences in legal reasoning style, statutory drafting conventions, and which sources count as authoritative. The positive but limited transfer gains on Bar Exam QA (+0–4 pp) and the architecture-dependent gap on Housing Statute QA (+16 pp for Ministral-8B vs. +2–3 pp for Qwen) are consistent with these limitations. Users deploying these models in non-Victorian legal contexts should evaluate on in-domain held-out data before trusting the reward signal.

Risk of misuse in high-stakes settings. Legal reward models could be misused to automate legal assessments in high-stakes settings without appropriate human oversight. We emphasise that reward accuracy on the benchmarks studied remains well below human expert agreement levels, and that the models studied are explicitly *not* calibrated to serve as definitive legal evaluators. Any deployment in real legal workflows should require expert human review of all outputs.

Data and model licences. Legal RAG Bench (Butler and Butler, 2026) and CONTEXTUAL-JUDGE BENCH (Xu et al., 2025) are released for research use; Bar Exam QA and Housing Statute QA (Zheng et al., 2025) are from the Stanford Reglab and used under their published terms. Ministral-8B is released under the Mistral Research Licence (non-commercial research only). Qwen3.5-9B and Qwen3.5-27B are released under Apache 2.0. All trained adapters released by this work inherit the licence of their respective base model.

Table 16: Estimated GPU-hours by experimental phase. Wall-clock times are rounded to the nearest 5 minutes.

Phase	Wall-clock (h)	GPU-hours
<i>Raw model benchmark (14 models $\times$ CJB)</i>		
0.5B–3B models ($\times 8$)	8.0	8.0
8B models ($\times 6$)	9.0	9.0
<i>Specialist model evaluation ($\times 3$)</i>		
Skywork / Selene / SFR-Judge	1.5	1.5
<i>DPO and SFT training (all corpora and models)</i>		
Ministral-8B ($\times 8$ runs incl. 3-seed)	12.0	12.0
Qwen3.5-9B ($\times 4$ runs)	8.0	8.0
Qwen3.5-27B QLoRA ($\times 4$ runs)	17.0	17.0
<i>Adapter evaluation (≈ 20 adapters $\times 4$ benchmarks)</i>		
CJB and Bar Exam (max_length=4096)	14.0	14.0
LEGALREWARDBENCH (max_length=8192)	10.0	10.0
Housing Statute QA ($n=500$)	5.0	5.0
<i>Augmentation (inference only, Ministral-8B)</i>		
LRB-v2 training pairs (940 pairs)	3.0	3.0
LRB-v2 test pairs (156 pairs)	0.5	0.5
Total	88.0	88.0

Table 17: Software dependencies. Versions correspond to the locked environment used for all reported experiments.

Package	Version	Role
Python	3.11.15	Runtime
PyTorch	2.4.1+cu124	Tensor operations, autograd
Transformers	4.45.2	Model loading, tokenisation
PEFT	0.12.0	LoRA and QLoRA adapters
TRL	0.11.4	DPO / SFT training loop
BitsAndBytes	0.49.2	4-bit NF4 quantisation (QLoRA)
Datasets	4.8.5	Preference data loading
Accelerate	0.34.2	Single-GPU training harness
NumPy	1.26.3	Numerical computation
HuggingFace Hub	0.36.2	Model and dataset upload

Privacy. No personally identifiable information was used in any stage of this work. The Victorian Criminal Charge Book is a public document.

AI Assistance AI assistants were used during the preparation of this work for limited writing, editing, and coding support, including proofreading, clarity suggestions, formatting assistance, and debugging. All scientific ideas, experimental design, analyses, interpretations, and conclusions were developed and verified by the authors. AI systems were not used to autonomously generate research contributions or make scientific decisions.

G Prompts

The following prompts are reproduced from the implementation used to generate and annotate the corpus. Placeholders such as {context} and {question} are filled by the pipeline at runtime.

Answer Generation

You are answering a legal question using only the provided legal context.

Rules:

- Provide the final answer only. Do not include hidden reasoning, analysis notes, or <think> tags.
- Use only the legal context below.
- Do not rely on outside knowledge or assumptions.
- If the context does not contain enough information to answer the question, say exactly:
"The provided context is insufficient to answer this question."
- Do not give a yes/no answer, legal conclusion, or likely answer when the context is insufficient.
- For yes/no questions, start with "Yes" or "No" when the context supports it, then briefly explain.
- Mention cases, statutes, or legal authorities only if they appear in the context.
- Keep the answer concise, neutral, and legally careful.

Legal context:

{context}

Question:

{question}

Answer:

Answer Behaviour Annotation

You are judging whether a model response actually answers a legal question or refuses to answer it.

Classify only the response behavior. Do not evaluate correctness, grounding, or completeness.

Return one JSON object that exactly follows the provided schema. Do not include markdown, comments, explanations, or extra keys.

Choose exactly one:

- "attempted": the response tries to answer the question.
- "abstained": the response does not answer because it says the provided information is insufficient, unavailable, or uncertain.
- "unusable": the response is empty, unreadable, incoherent, malformed, or severely repetitive.

Rules:

- Do not judge whether the answer is legally correct.
- Do not judge whether the answer is supported by context.
- If the response is readable, do not label it "unusable".
- If the response says the context is insufficient but still gives a substantive answer, classify it as "attempted".
- If the response gives a yes/no answer, legal conclusion, or practical legal outcome, classify it as "attempted", even if it includes uncertainty or caveats.
- Short readable answers, including legal terms, acronyms, and one-word answers, are "attempted" if they try to answer the question.
- If the response only refuses or explains why it cannot answer, classify it as "abstained".

Examples:

- Response: "No. The statute does not allow that remedy."
Output: {"answer_behavior": "attempted"}
- Response: "The context does not provide enough information to answer this question."
Output: {"answer_behavior": "abstained"}
- Response: "The context is limited, but the answer is likely no because the rule requires notice."
Output: {"answer_behavior": "attempted"}
- Response: ""
Output: {"answer_behavior": "unusable"}

Question:

```
{}  
{question}  
{}  

```

Candidate response:

```
{}  
{candidate_response}  
{}  

```

Faithfulness Annotation

You are judging whether a candidate legal answer is supported by the legal context that was shown to the model.

Return one JSON object that exactly follows the provided schema. Do not include markdown, comments, explanations, or extra keys.

Use only the provided legal context. Do not use the authoritative answer, outside legal knowledge, or assumptions about what the law should be.

Choose exactly one:

- "fully_supported": the context supports the answer's main conclusion and all material claims made in the answer.
- "partially_supported": the context supports the answer's main conclusion, but one or more other material claims made in the answer are not supported.
- "unsupported": the context does not support the answer's main conclusion, even if it supports some surrounding or related claims.
- "contradicted": the answer's main conclusion or material claims conflict with the context.

Topical overlap alone is not support. Invented or mischaracterized material claims are not supported.

If an invented or mischaracterized claim is the main conclusion, use "unsupported"; if it is a secondary material claim, use "partially_supported".

Examples:

- Context: "A tenant may terminate the lease only after giving written notice."
Candidate answer: "The tenant may terminate the lease after giving written notice."
Output: {"faithfulness": "fully_supported"}
- Context: "A tenant may terminate the lease only after giving written notice."
Candidate answer: "The tenant may terminate the lease after giving written notice, and the landlord must pay damages."
Output: {"faithfulness": "partially_supported"}
- Context: "A tenant may terminate the lease only after giving written notice."
Candidate answer: "The tenant may terminate the lease without notice."
Output: {"faithfulness": "contradicted"}
- Context: "The court discusses filing deadlines for appeals."
Candidate answer: "The tenant may terminate the lease after giving written notice."
Output: {"faithfulness": "unsupported"}
- Context: "A license is required before operating the business."
Candidate answer: "The context does not establish whether the defendant had a license."
Output: {"faithfulness": "fully_supported"}

Question:

{question}

Legal context shown to the model:

{candidate_context}

Candidate answer:

{candidate_response}

Correctness Annotation

You are judging whether a candidate legal answer reaches the correct bottom-line answer.

Return one JSON object that exactly follows the provided schema. Do not include markdown, comments, explanations, or extra keys.

Compare the candidate answer to the authoritative answer. Do not judge whether the candidate answer is grounded in any retrieved context.

Choose exactly one:

- "correct": the candidate answer reaches the same bottom-line legal answer as the authoritative answer.
- "incorrect": the candidate answer reaches a different bottom-line legal answer, gives no clear bottom-line answer, or materially changes the legal conclusion.

Examples:

- Authoritative answer: "No. The statute requires written notice before termination."
Candidate answer: "No, termination is not allowed without written notice."
Output: {"correctness": "correct"}
- Authoritative answer: "No. The statute requires written notice before termination."
Candidate answer: "Yes, the tenant can terminate immediately."
Output: {"correctness": "incorrect"}
- Authoritative answer: "Yes, if the claimant files within thirty days."
Candidate answer: "It depends on the facts."
Output: {"correctness": "incorrect"}

Question:

{question}

Authoritative answer:

{reference_answer}

Candidate answer:

{candidate_response}

Completeness Annotation

You are judging whether a candidate legal answer includes the key legal point needed to answer the question.

Return one JSON object that exactly follows the provided schema. Do not include markdown, comments, explanations, or extra keys.

Compare the candidate answer to the authoritative answer. Do not judge whether the candidate answer is grounded in any retrieved context.

Choose exactly one:

- "complete": the candidate answer includes the core legal point, condition, exception, or qualification needed to justify the authoritative answer.
- "incomplete": the candidate answer misses the core legal point, gives only a bare conclusion, or omits an essential condition, exception, or qualification.

Examples:

- Authoritative answer: "No. The tenant cannot terminate unless they first give written notice."
Candidate answer: "No. The tenant must first give written notice."
Output: {"completeness": "complete"}
- Authoritative answer: "No. The tenant cannot terminate unless they first give written notice."
Candidate answer: "No."
Output: {"completeness": "incomplete"}
- Authoritative answer: "Yes, but only if the appeal is filed within thirty days."
Candidate answer: "Yes, an appeal is allowed."
Output: {"completeness": "incomplete"}
- Authoritative answer: "Yes, but only if the appeal is filed within thirty days."
Candidate answer: "Yes, an appeal is allowed if it is filed within thirty days."
Output: {"completeness": "complete"}

Question:

{question}

Authoritative answer:

{reference_answer}

Candidate answer:

{candidate_response}

Length-Balanced Augmentation Prompts

The following three prompts are used by the length-balanced augmentation procedure (Section 2). They are applied to a frozen Ministral-8B-Instruct-2410 model with temperature = 0.7, top- p = 0.9, and max_new_tokens = 128 to regenerate the short-template side of length-asymmetric pairs. Each prompt is wrapped with the Ministral chat template; the legal question q is extracted from the original benchmark prompt by regex.

P1. Hallucinated Answer (replaces 9-word rejected on refusal_answerable)

You are answering legal questions for a study on RAG faithfulness. Ignore any provided retrieval context: answer the following question from general legal world knowledge alone. Be confident, specific, and give 2-3 sentences of reasoning. Do NOT refuse and do NOT cite the context.

Question:
{q}

Answer:

P2. Elaborate Refusal (replaces 9-word chosen on refusal_unanswerable)

You are a careful legal RAG assistant. The retrieved context does not contain enough information to answer the following question. Write a professional 2-3 sentence refusal: acknowledge the question, briefly explain why the available context is insufficient, and indicate what additional sources would resolve it. Do NOT speculate or invent law.

Question:
{q}

Refusal:

P3. Terse-but-Incomplete Answer (replaces “No.”-style rejected on completeness)

Answer the following legal question correctly but tersely: give the bottom-line conclusion in one short sentence, omitting the legal rule, statutory authority, exceptions, and qualifications that would normally be required for a complete answer. Do NOT cite cases.

Question:
{q}

Brief answer:

Model Repository References

All models were accessed via the Hugging Face Hub on 25 May 2026. Table 18 lists the repository identifier for every model evaluated in this work.

Table 18: Hugging Face repository identifiers for all evaluated models.

Model name	Hugging Face repository
<i>Instruction-tuned baselines</i>	
SmolLM2-1.7B	HuggingFaceTB/SmolLM2-1.7B-Instruct
Qwen2.5-0.5B	Qwen/Qwen2.5-0.5B-Instruct
Qwen3.5-2B	Qwen/Qwen3.5-2B
Llama-3.2-1B	meta-llama/Llama-3.2-1B-Instruct
Llama-3.2-3B	meta-llama/Llama-3.2-3B-Instruct
Gemma 2-2B	google/gemma-2-2b-it
Qwen2.5-3B	Qwen/Qwen2.5-3B-Instruct
Phi-4-mini	microsoft/Phi-4-mini-instruct
Phi-3.5-mini	microsoft/Phi-3.5-mini-instruct
Qwen2.5-7B	Qwen/Qwen2.5-7B-Instruct
Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct
Minstral-8B	mistralai/Minstral-8B-Instruct-2410
DeepSeek-R1-8B	deepseek-ai/DeepSeek-R1-Distill-Llama-8B
SFR-8B	Salesforce/LLaMA-3-8B-SFR-Iterative-DPO-R
Qwen3.5-9B	Qwen/Qwen3.5-9B
Qwen3.5-27B	Qwen/Qwen3.5-27B
<i>Specialist reward / judge models</i>	
Skywork-Reward-8B	Skywork/Skywork-Reward-Llama-3.1-8B-v0.2
Selene-1-Mini-8B	AtlaAI/Selene-1-Mini-Llama-3.1-8B
SFR-Judge-8B	Salesforce/LLaMA-3-8B-SFR-RM-R
<i>DPO fine-tuning base models</i>	
Gemma 2-2B (DPO)	google/gemma-2-2b-it
SFR-8B (DPO)	Salesforce/LLaMA-3-8B-SFR-Iterative-DPO-R
DeepSeek-R1-8B (DPO)	deepseek-ai/DeepSeek-R1-Distill-Llama-8B
Qwen2.5-7B (DPO)	Qwen/Qwen2.5-7B-Instruct
Minstral-8B (DPO)	mistralai/Minstral-8B-Instruct-2410
SFR-Judge-8B (DPO)	Salesforce/LLaMA-3-8B-SFR-RM-R