LEXREWARD: A TAXONOMY-DRIVEN REWARD FRAMEWORK FOR LEGAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Legal language models require reward signals that capture not only answer correctness but also the multidimensional quality of legal responses. Existing reward methods, however, often rely on coarse-grained holistic judgments, providing limited domain specificity and interpretability. We introduce **LexReward**, a taxonomy-driven framework for legal reward modeling. **LexReward** characterizes legal response quality along three complementary dimensions: *Style*, covering lexical and syntactic quality; *Element*, assessing legal subjects, facts, statutes, and decisions; and *Chain*, evaluating the order, completeness, correctness, and non-redundancy of legal reasoning. For each dimension, we develop rubrics that specify evaluation criteria and quality levels. The resulting rewards are used to construct pairwise preference data for Direct Preference Optimization (DPO) and reward-model training. Experiments show that the rubric-based rewards reliably distinguish legal responses of different quality and that DPO training on the preference data improves performance across all three dimensions. The learned reward models, **LexRM**, also support effective downstream optimization: each dimension-specific reward model improves policy performance in its corresponding dimension through reinforcement learning, without requiring reference answers at reward time. Dimension-wise analyses further support the effectiveness of the proposed taxonomy and reward construction. All code and data will be made available.

1 INTRODUCTION

Legal large language models have demonstrated growing potential across a wide range of legal generation tasks, such as legal question answering and legal reasoning (Fei et al., 2024; Xie et al., 2026; Yao et al., 2025; Li et al., 2024). Nevertheless, producing a high-quality legal response requires considerably more than arriving at the correct final answer. A legally sound response should identify the relevant subjects and facts, accurately invoke applicable statutes, construct a complete and coherent reasoning chain, and express its conclusions in precise and objective legal language. Models may reach a correct conclusion through incomplete reasoning, omit legally decisive facts, cite an inappropriate statutory basis, or produce fluent but legally unreliable explanations. These failures cannot be adequately captured by final-answer accuracy alone.

As reinforcement learning (RL) becomes increasingly important for improving model reasoning, reward design plays a central role in specifying which aspects of legal response quality models are encouraged to improve. Existing reward approaches in legal AI often rely on general-purpose reward models or reward functions that assess only selected aspects of legal response quality, such as judgment-outcome accuracy (Cai et al., 2025; Zhang et al., 2025). Although useful in some settings, these approaches may overlook other aspects of a response, making it difficult to determine whether improvements reflect better legal response quality or superficial features such as response length and fluency. More fundamentally, without a structured definition of legal response quality, it remains unclear what a legal reward should assess.

To clarify which aspects of legal response quality should be rewarded, as shown in Fig. 1, we introduce **LexReward**, a taxonomy-driven reward framework for legal language models. Drawing on legal experts' domain knowledge, we first develop a taxonomy of legal response quality with three complementary dimensions: *Style*, which captures lexical and syntactic properties of legal writing;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

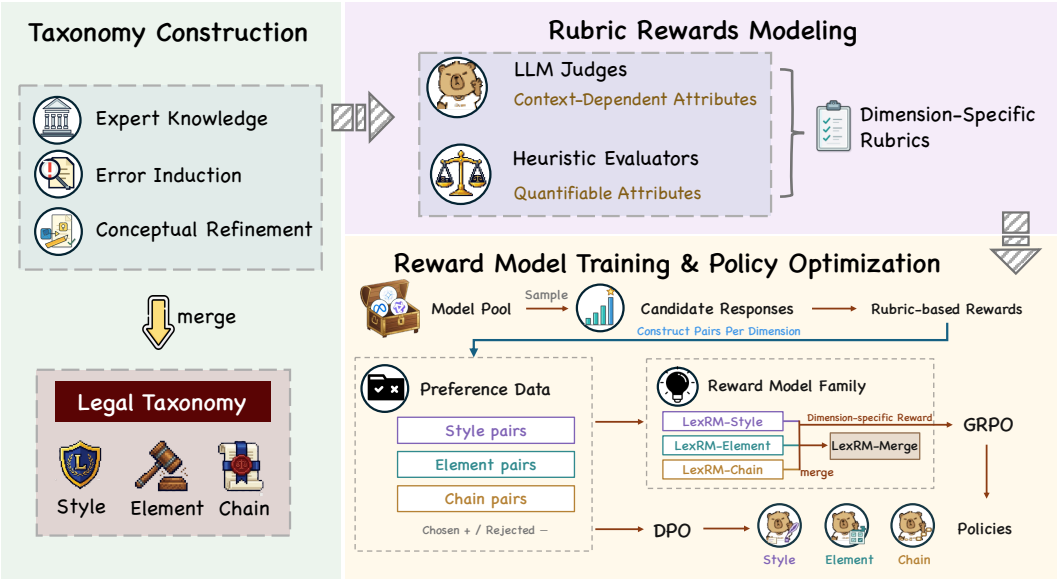

Figure 1: Overview of the LexReward framework. LexReward first constructs a legal taxonomy through expert knowledge, error induction, and conceptual refinement. It then operationalizes the taxonomy into dimension-specific rubric-based rewards, which construct preference data for training a reward model and guide reinforcement learning.

Element, which assesses the identification and treatment of legal subjects, facts, statutes, and decisions; and *Chain*, which evaluates the order, completeness, correctness, and non-redundancy of legal reasoning. Legal experts further refine the taxonomy through an empirical analysis comparing existing models' responses with reference answers, yielding fine-grained criteria for each dimension. We design a scoring rubric for each fine-grained criterion and use either rule-based or LLM-as-a-judge evaluators, depending on the context. Experiments show that the resulting rewards reliably distinguish responses of different quality along their corresponding dimensions.

To provide dense reward signals and evaluate the practical utility of **LexReward**, we score diverse responses from a pool of models using rubric-based rewards and construct pairwise preference data both to evaluate their utility for Direct Preference Optimization (DPO), and to train **LexRM**, a family of Chinese legal reward models. To the best of our knowledge, **LexRM** is the first collection of reward models developed for the Chinese legal context. We evaluate **LexRM** through Test-Time Scaling (TTS), using it to select the highest-scoring response from candidates generated by multiple models and comparing against random selection from the same candidate pool. We further assess its effectiveness for Group Relative Policy Optimization (GRPO). Our experiments show that DPO improves generation performance across all three dimensions, while **LexRM**-guided selection outperforms random selection. When used for GRPO, **LexRM** also yields gains across all dimensions and outperforms rule-based outcome rewards (Dai et al., 2026; Cai et al., 2025; Zhang et al., 2025).

Together, these results establish **LexReward** as a framework for defining and rewarding legal response quality, enabling legal domain knowledge to guide both reward construction and model optimization. Our main contributions are as follows:

- We introduce a structured taxonomy of legal response quality that decomposes the requirements of a high-quality legal response into three complementary dimensions (*Style*, *Element*, and *Chain*) and fine-grained criteria, providing a principled foundation for legal reward modeling.
- We develop interpretable, rubric-based rewards for each fine-grained criterion in the taxonomy using rule-based and LLM-based evaluation, and demonstrate that these rewards effectively distinguish response quality in Best-of-5 selection tasks.
- We construct preference data using the rubric-based rewards and demonstrate that DPO training on these data improves legal response generation.

- We train **LexRM**, to the best of our knowledge the first family of Chinese legal reward models, on rubric-derived preference data and demonstrate its effectiveness in Test-Time Scaling (TTS).
- We further integrate the learned reward models into a Group Relative Policy Optimization (GRPO) pipeline and demonstrate that their supervision improves the generation quality of the policy.

2 Related Work

2.1 Legal Language Models and Domain-Specific Evaluation

Large language models (LLMs) have been increasingly adapted to the legal domain, with prior work improving their legal knowledge and reasoning abilities through domain-specific training and alignment methods (Chalkidis et al., 2020; Xu et al., 2025; Dai et al., 2026). Meanwhile, benchmarks such as LegalBench (Guha et al., 2023) and LawBench (Fei et al., 2024) provide standardized testbeds for evaluating legal knowledge, rule understanding, and reasoning ability across different legal tasks (Zhong et al., 2020). However, existing legal evaluations often emphasize task-level performance, which does not fully capture the multidimensional quality of a legal response. A reliable legal response should also exhibit appropriate legal expression, sufficient coverage of relevant legal elements, faithful grounding in applicable statutes, and coherent reasoning. This motivates a structured reward framework that decomposes legal response quality into interpretable dimensions and translates them into training and optimization signals.

2.2 Rubric-based Reward Modeling

Reward modeling learns to assess response quality for candidate selection and policy optimization. Early approaches learn scalar rewards from holistic human preferences (Ouyang et al., 2022), while subsequent work introduces finer-grained supervision over error types, text segments, and multiple quality objectives (Wu et al., 2023; Wang et al., 2024). Complementing this decomposition, rubric-based evaluation (Kim et al., 2024) makes judgment criteria and quality levels explicit, enabling evaluators to assess responses against specified standards. In the legal domain, existing work (Chen et al., 2026) learns process rewards for criminal-law knowledge-graph reasoning, but focuses on a specific task. Building on the broader shift from holistic preferences toward structured quality supervision, **LexReward** systematically defines reward dimensions and evaluation criteria through a legal quality taxonomy, then operationalizes it into rubrics and cross-task preference data, connecting domain-specific quality definitions with reward model training and reinforcement learning.

3 The LexReward Taxonomy

Taxonomy construction. Legal response quality encompasses multiple requirements that a single holistic criterion leaves implicit. We therefore construct a taxonomy that decomposes legal response quality into explicit, assessable dimensions. With the assistance of legal experts, we combine top-down specification with bottom-up error analysis. In the top-down component, legal experts draw on their domain knowledge to identify the core requirements of a high-quality legal response (Goodrich, 1990; Osbeck, 2011; Maley, 2014). In the bottom-up component, they examine responses generated by multiple language models (Qwen Team, 2025; Llama Team, 2024) across legal AI benchmarks (Fei et al., 2024; Ma et al., 2026; Li et al., 2024), comparing them with reference answers to identify errors and unmet quality requirements. The requirements identified through both components are then integrated and refined into a unified taxonomy.

As illustrated in Fig. 2, the resulting taxonomy comprises three complementary dimensions: *Style*, *Element*, and *Chain*. These dimensions characterize how legal content is expressed, what legally relevant information is included, and how that information is connected to support a conclusion. Each dimension is further decomposed into finer-grained criteria for rubric-based assessment.

Style assesses the linguistic quality of legal responses through *lexical features* and *syntactic features*. *Lexical features* capture the precise use of legal terminology (*word specificity*) and the use of objective language without unwarranted subjective judgments (*subjective word control*). *Syntactic features* capture cohesion across sentences (*sentence cohesion*), clarity of sentence structure

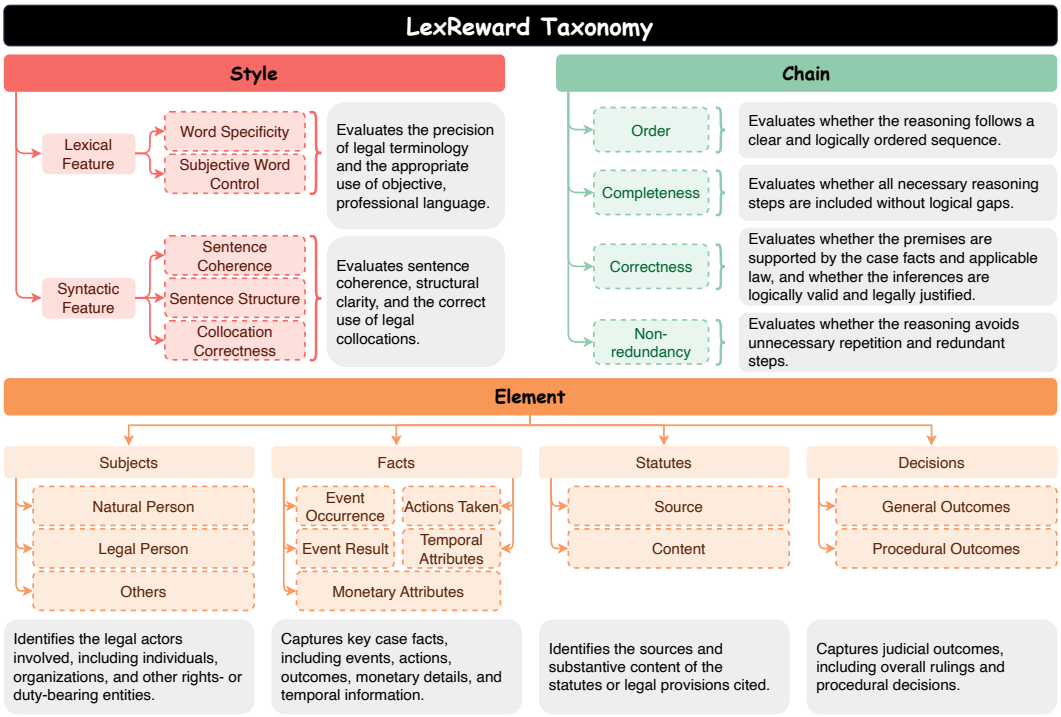

Figure 2: The LexReward taxonomy and its rubric-oriented operationalization. Legal response quality is decomposed into three complementary dimensions: *Style* (red), *Element* (orange), and *Chain* (green). Colored boxes show the dimensions and their constituent criteria, while gray boxes summarize the corresponding evaluation objectives. This structured decomposition supports the development of fine-grained rubrics for legal reward modeling.

(*sentence structure*), and appropriate combinations of words in legal expressions (*collocation correctness*). Together, these criteria assess the precision, objectivity, and clarity of legal expressions.

Element assesses whether a response includes the legally relevant information needed to address a legal problem. It comprises four categories: *subjects*, *facts*, *statutes*, and *decisions*, covering the relevant entities, case circumstances, statutory provisions, and legal conclusions, respectively.

Chain assesses the reasoning that connects case information to a legal conclusion. It comprises four categories: *order*, *completeness*, *correctness*, and *non-redundancy*. These criteria assess whether the reasoning steps are logically arranged, sufficiently developed, legally valid, and free from unnecessary repetition. Whereas *Element* assesses coverage of relevant information, *Chain* assesses the inferential connections among that information.

Together, the three dimensions organize legal response quality in terms of stylistic expression, substantive content, and reasoning. The taxonomy provides a structured basis for translating expert-defined quality requirements into explicit rubric criteria, which support fine-grained assessment and rubric-based reward modeling.

4 Taxonomy-Driven Rewards

4.1 Rubric-based Reward Operationalization

We operationalize the fine-grained criteria in our taxonomy using two complementary scoring strategies, selected according to whether evaluating a criterion requires the input context. For *context-independent* criteria, the score can be computed from intrinsic properties of the generated response. We therefore use rule-based reward functions calibrated against a corpus of high-quality legal texts,

providing deterministic and interpretable scores. For *context-dependent* criteria, evaluation requires determining whether the response appropriately addresses the legal problem specified by the input. We use an LLM-as-a-judge evaluator for these criteria, leveraging its ability to assess the relationship between the input and the generated response under a criterion-specific rubric.

For a context-independent criterion k, we define the reward as:

$$r_k(y) = \mathrm{Eval}_k^{\mathrm{rule}}\left(y; \theta_k, \mathcal{C}\right), \tag{1}$$

where y denotes the generated response, θ_k denotes the criterion-specific scoring parameters, and $\mathcal{C}$ is an optional corpus of high-quality legal texts used to calibrate the evaluator.

For a context-dependent criterion k, we define the reward as:

$$r_k(x, y) = \mathrm{Eval}_k^{\mathrm{LLM}}\left(x, y; \rho_k\right), \tag{2}$$

where x denotes the input context, y denotes the generated response, and ρ_k specifies the scoring rubric supplied to the LLM judge.

We assign an evaluation strategy to each taxonomy criterion based on its dependence on the input context. *Style* primarily concerns intrinsic linguistic properties of the response and is therefore evaluated using a rule-based evaluator. *Element* requires assessing whether the response identifies and treats the legally relevant information in the input and is therefore evaluated using an LLM judge. The evaluation of *Chain* is task-dependent. When a task provides an explicit reasoning structure, we use rule-based functions to assess conformity to the prescribed steps and order. When the appropriate reasoning path must be inferred from the legal context, we instead use an LLM judge. We next describe the rubric and scoring procedure for each fine-grained criterion:

Style. The *Style* reward measures how closely a response conforms to the lexical and syntactic patterns of authentic Chinese judicial documents. We evaluate five attributes: $\mathcal{K}_{\mathrm{Style}} = \{\mathrm{ws}, \mathrm{sw}, \mathrm{coh}, \mathrm{str}, \mathrm{col}\}$, corresponding to *word specificity*, *subjective word control*, *sentence cohesion*, *sentence structure*, and *collocation correctness*, respectively.

For each attribute k, we extract a feature representation $f_k(y)$ from response y and measure its discrepancy from the corresponding reference representation f_k^{ref}, estimated from a corpus of authentic judicial documents:

$$d_k(y) = D_k\big(f_k(y), f_k^{\mathrm{ref}}\big), \qquad k \in \mathcal{K}_{\mathrm{Style}}, \tag{3}$$

where D_k is a non-negative discrepancy measure appropriate to the feature type. *Word specificity* and *subjective word control* use KL divergence to compare legal-term and sentiment-word distributions, respectively. *Sentence cohesion* uses the absolute difference in conjunction frequency. *Sentence structure* and *collocation correctness* use Euclidean distance to compare sentence-length statistics and 3- to 6-gram coverage vectors, respectively.

To account for scale differences, we divide each discrepancy by its mean over calibration corpus $\mathcal{C}$:

$$s_k = \frac{1}{|\mathcal{C}|} \sum_{y' \in \mathcal{C}} d_k(y'), \qquad \widetilde{d}_k(y) = \frac{d_k(y)}{s_k + \epsilon}, \tag{4}$$

where $\epsilon > 0$ ensures stability. The *Style* reward is the negated mean normalized discrepancy:

$$R_{\mathrm{Style}}(y) = -\frac{1}{|\mathcal{K}_{\mathrm{Style}}|} \sum_{k \in \mathcal{K}_{\mathrm{Style}}} \widetilde{d}_k(y). \tag{5}$$

Higher rewards indicate greater similarity with the reference writing style across the five attributes. Feature extraction, reference estimation, and calibration procedures are detailed in Appendix A.1.

Element. The *Element* rubric evaluates whether a legal response matches the task-required legal components. We instantiate this rubric with an LLM-as-a-Judge evaluator, which scores the candidate response along four sub-dimensions: *subjects*, *facts*, *statutes*, and *decisions*. *Subjects* capture legal actors and their roles; *facts* capture case-relevant factual information; *statutes* capture statutory grounds; and *decisions* capture the final legal conclusion. Given a query and a candidate response, the judge assigns each sub-dimension a score from 0 to 1 without access to the reference answer, and the *Element* reward is computed as the average of the four scores. The judge is instructed to assess whether the candidate answer matches the query requirements in each *Element* sub-dimension,

considering element accuracy, coverage, specificity, and unsupported or fabricated content. The full prompt is provided in Appendix A.2.

Chain. The *Chain* rubric evaluates the structure of legal reasoning along four sub-dimensions: *order*, *completeness*, *correctness*, and *non-redundancy*. *Order* assesses the logical sequence of reasoning steps; *completeness* measures coverage of required steps; *correctness* checks whether each step serves an appropriate reasoning function; and *non-redundancy* assesses unnecessary repetition or conflicting content across repeated steps. When a task specifies predefined reasoning steps, a rule-based evaluator scores these dimensions against the prescribed step types, required groups, and ordering constraints. Otherwise, an LLM-as-a-Judge evaluates the response along the same dimensions based on the query requirements. The dimension-level scores are aggregated into the final *Chain* reward. Detailed scoring rules and the judge prompt are provided in Appendix A.3.

4.2 Reward Model Training from Rubric-Guided Preferences

Rubric-based rewards can be sparse, often assigning identical scores to responses of differing quality. Such ties limit their ability to distinguish between candidate responses and can reduce their effectiveness in downstream applications. We therefore construct preference data from rubric-derived scores and train **LexRM**, a family of Chinese legal reward models, to learn legal preference functions that generalize beyond the discrete distinctions captured by the rubrics.

Dimension-specific reward models. To diversify response sources and reduce reliance on stylistic cues specific to a single generator, we maintain a pool of n models that generate candidate responses for each query x. For each quality dimension d, we score these candidates using the corresponding rubric and construct preference pairs within the same query. To capture quality differences across the score range, we organize the pairs into three categories: (i) the highest-scoring response versus the nearest lower-scoring response in the high-score range; (ii) a high-scoring response versus a medium-scoring response; and (iii) a medium-scoring response versus a substantially lower-scoring response. These categories provide supervision for both fine-grained distinctions among strong responses and broader differences across quality levels.

The resulting preference dataset $\mathcal{P}_d$ contains tuples (x, y^+, y^-), where y^+ receives a higher rubric score than y^-. We train a separate reward model R_{θ_d} per dimension using the Bradley–Terry objective (Bradley & Terry, 1952):

$$\mathcal{L}_{\mathrm{BT}}^{(d)} = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{P}_d}\left[\log\sigma\big(R_{\theta_d}(x,y^+) - R_{\theta_d}(x,y^-)\big)\right], \tag{6}$$

where σ denotes the sigmoid function. This objective encourages the model to assign higher scalar scores to preferred responses. The resulting dimension-specific reward models support separate assessment of each taxonomy dimension, enabling analysis of where legal response quality improves or remains deficient.

Multi-dimensional reward models. The dimension-specific reward models, one per dimension $d \in \mathcal{D} = \{\textit{element}, \textit{style}, \textit{chain}\}$, share one pretrained backbone θ_0 and differ only in the preference datasets used for fine-tuning. For each dimension, we represent the parameter update as a task vector $\tau_d = \theta_d - \theta_0$. Across the backbone's weight matrices, which carry virtually all of its parameters, the pairwise cosine similarities between these task vectors are small ($|\cos| \leq 0.03$ on average, never above 0.14), with the little overlap that exists confined to the value and output projections of the topmost layers. These low pairwise cosine similarities motivate exploring additive composition of the dimension-specific updates. We form a single multi-dimensional reward model using task arithmetic (Ilharco et al., 2023):

$$\theta_{\mathrm{merged}} = \theta_0 + \textstyle\sum_{d\in\mathcal{D}} \lambda_d \tau_d, \qquad \lambda_d = 1, \tag{7}$$

where λ_d scales the update contributed by dimension d. The merged model is a single reward model that scores all dimensions, removing the need to run three backbones at inference time.

5 Experiments and Results

We evaluate taxonomy-driven rewards at three levels. First, we assess the taxonomy-derived rubric-based rewards on pairwise selection tasks and Best-of-5 selection for generation tasks. Second, we

Table 1: Performance of rubric-based rewards and reward models on datasets evaluating three dimensions. Bold denotes the best result in each column.

Style: CLASE		**Element: LegalΔ**						
Method	Accuracy (%)	Method	SPP-F	CCP	SLP	CAS	CAC	Avg.
Random	50.00	Five-model average	53.37	46.34	28.46	35.88	84.12	49.63
Rubric	79.00	Rubric	**79.04**	57.03	33.10	39.40	**92.80**	60.27
Skywork-Qwen	78.50	Skywork-Qwen	75.97	57.84	**49.80**	**64.20**	92.40	**68.04**
Skywork-Llama	52.50	Skywork-Llama	61.15	54.35	29.90	51.80	91.00	57.64
Lawformer	68.50	Lawformer	50.95	44.92	39.00	30.40	88.60	50.77
Legal-BERT	41.00	Legal-BERT	31.19	39.27	20.10	29.00	74.80	38.87
LexRM-Style	**81.75**	LexRM-Element	77.72	57.80	44.70	59.20	92.20	66.32
LexRM-Merge	69.75	LexRM-Merge	75.57	**58.27**	43.80	58.80	92.40	65.77

Chain: LexChain

Method	Plaintiff	Defendant	Dispute	Statute	Liability	Damages	Judgment	Overall
Five-model average	96.30	87.47	18.01	21.68	22.82	23.07	24.24	45.41
Rubric	96.00	87.45	21.20	27.25	25.75	25.85	27.82	47.80
Skywork-Qwen	96.75	88.60	27.00	29.95	27.95	**28.35**	**32.06**	50.19
Skywork-Llama	96.55	87.75	26.10	29.80	28.20	28.25	30.73	49.83
Lawformer	96.20	88.15	18.20	23.65	23.05	23.15	24.53	45.93
Legal-BERT	95.85	85.80	13.50	16.95	19.85	21.10	19.84	42.70
LexRM-Chain	**96.80**	**89.15**	**27.40**	**30.65**	**28.75**	28.00	31.46	**50.46**
LexRM-Merge	96.25	88.00	19.30	23.45	24.15	26.15	26.81	46.84

use rubric-derived preference data to train reward models and evaluate their effectiveness through MoE-based TTS, selecting the highest-scoring response from candidates generated by multiple models and comparing against random selection from the same pool. Third, we assess whether this supervision improves generation quality through DPO on the preference data and reinforcement learning from an SFT checkpoint using the learned reward models.

5.1 Settings

5.1.1 Datasets

For *Style*, we use CLASE (Ma et al., 2026), with 4,000 training instances and an official test set of 1,000 instances. Reward scorers are evaluated by pairwise accuracy in selecting the gold response over a model-generated negative; policies are evaluated by the CLASE-Mix score. For *Element*, we draw data from the criminal questions of JEC-QA (Zhong et al., 2020) and the civil judgments of LexChain (Xie et al., 2026), using 2,244 preference pairs built from 2,736 instances for training. Evaluation follows the in-domain protocol of LegalΔ (Dai et al., 2026) on its 3,000-instance test set, which reports F1 for statutory-article and charge prediction and accuracy for sentence-length prediction, case analysis, and financial calculation; the overall score is the unweighted mean of the five values. For *Chain*, we use LexChain (Xie et al., 2026), with 9,550 training instances and an official test set of 1,000 instances, and adopt its native LLM-based evaluation, which scores seven aspects of a judgment and reports an overall score. Dataset statistics and the construction of preference data are detailed in Appendix B.1.

5.1.2 Models

We use DeepSeek-V4-Flash (DeepSeek-AI, 2026) for all LLM-based components of the rubric evaluators and Qwen3-8B (Qwen Team, 2025) as the backbone for all trained models. To generate candidate responses for preference data construction, we use a pool of five models: Qwen3-4B (Qwen Team, 2025), Qwen3.5-4B, Qwen3.5-9B (Qwen Team, 2026), Llama-3.1-8B-Instruct (Llama Team, 2024), and gemma-4-12B-it (Gemma Team, 2026). To evaluate reward model performance, we com-

Table 2: Downstream policy performance across the three **LexReward** dimensions. SPP-F and CCP report entity-level F1; SLP, CAS and CAC report exact-match accuracy; Avg. is the mean of these five Element metrics. Bold denotes the best result in each column, including ties.

Method	**Style: CLASE**	**Element: LegalΔ**					
	CLASE-Mix (/10)	SPP-F	CCP	SLP	CAS	CAC	Avg.
Qwen3-8B	3.08	76.80	51.89	45.00	35.00	80.20	57.78
SFT	7.32	75.45	49.85	46.80	49.20	81.80	60.62
DPO	5.21	75.67	52.99	44.20	36.20	81.60	58.13
GRPO (rule)	8.43	**80.45**	48.67	47.60	51.80	75.80	60.86
GRPO (RM)	**8.54**	78.64	**54.77**	**56.70**	**53.40**	**89.80**	**66.66**

Chain: LexChain

Method	Plaintiff	Defendant	Dispute	Statute	Liability	Damages	Judgment	Overall
Qwen3-8B	97.25	88.90	34.20	42.10	34.30	28.45	26.46	53.55
SFT	97.95	**90.85**	**39.00**	39.40	35.05	35.15	36.06	55.99
DPO	97.50	89.05	36.20	**42.75**	35.40	30.10	27.79	54.47
GRPO (rule)	97.40	89.80	36.50	37.90	**35.55**	34.00	37.64	55.29
GRPO (RM)	**98.25**	90.80	37.30	39.55	35.20	**36.70**	**38.49**	**56.40**

pare **LexRM** against Skywork-Reward-V2-Llama-3.1-8B, Skywork-Reward-V2-Qwen3-8B (Liu et al., 2025), as well as Lawformer (Xiao et al., 2021) and LegalBERT (Chalkidis et al., 2020). Training and inference configurations are provided in Appendix B.2.

5.2 Results

5.2.1 Rubric and RM Evaluation

In this section, we evaluate the rubric-derived rewards, together with **LexRM** trained on the preference data we construct, on selecting among candidate responses from multiple models.

As shown in Table 1, on all three datasets the rubric-based rewards (**Rubric**) select better responses than the five-model average or random selection, and the reward models trained on the preference pairs achieve further improvements: despite using substantially less training data than the Skywork reward models, **LexRM-Style** and **LexRM-Chain** achieve the best overall results on their respective datasets, while **LexRM-Element** ranks second. These results suggest that the reward models successfully internalize the rubric criteria as continuous scoring functions, enabling them to distinguish candidates that the rubrics may score equally. **LexRM-Merge** combines the three dimension-specific reward models into one. On LegalΔ, it matches the element expert and obtains the highest charge-prediction score, while on CLASE and LexChain it falls below the respective expert models.

5.2.2 DPO and RL Evaluation

In this section, we evaluate how rewards derived from the **LexReward** taxonomy transfer to downstream policies by examining performance after DPO and comparing RL with **LexRM** against RL with an outcome reward.

As shown in Table 2, DPO improves over the vanilla model across all three dimensions, indicating that the rubric-derived preference data provide useful reward information for policy optimization. Comparing the vanilla model, SFT, and GRPO initialized from SFT further demonstrates the benefits of reinforcement learning: GRPO (RM) achieves the highest score on every dimension, with the largest gain on Element, where it improves all five metrics and raises their average by 6.04 points over SFT. These results show that **LexRM** provides effective supervision for improving legal generation beyond supervised fine-tuning.

Legal RL is currently driven almost entirely by outcome rewards defined on a verifiable final answer, such as a predicted charge, statute, or monetary amount (Dai et al., 2026; Cai et al., 2025; Zhang

Table 3: Ablation study of the *Style*, *Element*, and *Chain* rubric dimensions. Bold denotes the best result in each column.

Style: CLASE		**Element: LegalΔ**						
Rubric	Accuracy (%)	Rubric	SPP-F	CCP	SLP	CAS	CAC	Avg.
Lexical	76.50	Subjects	65.70	48.76	7.60	32.80	**93.00**	49.57
Syntactic	77.50	Facts	68.13	50.73	11.20	33.60	**93.00**	51.33
Overall	**79.00**	Statutes	79.52	**57.64**	30.60	39.20	92.80	59.95
		Disposition	**79.54**	56.76	17.50	31.80	**93.00**	55.72
		Overall	79.04	57.03	**33.10**	**39.40**	92.80	**60.27**

Chain: LexChain								
Rubric	Plaintiff	Defendant	Dispute	Statute	Liability	Damages	Judgment	Overall
Order	95.70	87.30	19.80	24.65	23.65	23.35	25.94	46.25
Completeness	96.00	87.20	**21.80**	**27.95**	24.60	24.25	27.33	47.43
Correctness	95.95	87.15	20.00	25.85	23.95	24.40	26.60	46.77
Non-redundancy	**96.05**	**88.15**	18.10	22.10	24.55	25.30	26.79	46.43
Overall	96.00	87.45	21.20	27.25	**25.75**	**25.85**	**27.82**	**47.80**

et al., 2025). Following this practice, we use outcome-based rewards for *Element* and *Chain*, and ROUGE against the reference text for *Style*. As shown in Table 2, for *Style*, GRPO (rule) improves performance, but the gain is smaller than that achieved by GRPO (RM), suggesting that lexical overlap provides less effective supervision than the learned reward. GRPO (rule) leaves the Element and Chain averages largely unchanged: its gains are confined to the final answer, whereas the dimensions that ground that answer, such as Legal Basis, stagnate or decline. An outcome reward is indifferent to whether the conclusion it credits was reached through a complete and correctly grounded chain. However, rubric-derived rewards are subject to neither restriction. They are equally applicable to open-ended legal scenarios and do not depend on reference labels. This shows that the benefit comes from making the reward dimensional: what the taxonomy supplies is supervision on the reasoning that leads to an answer, not a better estimate of the answer itself.

5.2.3 Dimension Ablation Analysis

In this section, we examine whether the sub-dimensions within each taxonomy dimension carry complementary signal by deriving a reward from each sub-dimension in isolation and comparing it against their aggregation.

As shown in Table 3, aggregation gives the best overall score in all three dimensions of the taxonomy. On individual tasks it is occasionally beaten by a single criterion, but always by a small margin, whereas no single criterion consistently performs best across tasks. Within *Element*, for instance, a rubric using subjects alone never checks which provision is cited, and statute prediction drops well below the aggregate, which is restored by adding statutes; within *Chain*, the aggregate improves over every single criterion, with the gain concentrated in the later stages of the chain, liability, loss and judgment, whose conclusions must be reasoned to rather than read off the case description, while the earlier fact-identification stages are already saturated. This shows that the value of the taxonomy is not that one criterion suffices, but that scoring along several at once keeps a task-irrelevant criterion from deciding the outcome.

6 Conclusion

We introduce **LexReward**, a taxonomy-driven framework that characterizes legal response quality along three complementary dimensions: *Style*, *Element*, and *Chain*. **LexReward** operationalizes these dimensions as fine-grained scoring rubrics that effectively distinguish responses of different quality. The rubric-derived preference data improve legal response generation through DPO and support the training of **LexRM**, a family of Chinese legal reward models. **LexRM** improves multi-model response selection and provides effective reinforcement-learning rewards for separately optimizing policies in *Style*, *Element*, and *Chain*. These results establish taxonomy-driven rewards as effective supervision for improving legal language models.

AI USE STATEMENT

In this work, we used generative AI tools for language refinement, partial code implementation, the generation of preference data, and assistance with interpreting results. We did not use these tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, or provide essential components of mathematical proofs; the remaining disclosure categories are not applicable to this work. The authors reviewed and validated all AI-assisted outputs. Specifically, we checked AI-refined text for accuracy, tested AI-generated code and data for correctness, and independently verified AI-assisted interpretations of the results. The authors take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

The construction of the legal quality taxonomy is described in Section 3, while Section 4 specifies the rubric operationalization, preference data construction, and reward model training objectives. Section 5 presents the datasets, models, and evaluation protocols used in our experiments. Appendix A provides implementation details for the rubric-based rewards, while Appendix B documents dataset statistics, preference data construction, and training/inference/evaluation configurations.

REFERENCES

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 1952.

Hua Cai, Shuang Zhao, Liang Zhang, Xuli Shen, Qing Xu, Weilin Shen, Zihao Wen, and Tianke Ban. Unilaw-r1: A large language model for legal reasoning with reinforcement learning and iterative inference. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18117–18131, 2025.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2898–2904, 2020.

Jiujiu Chen, Yazheng Liu, Sihong Xie, and Hui Xiong. SCPRM: A schema-aware cumulative process reward model for knowledge graph question answering. *arXiv preprint arXiv:2605.02819*, 2026.

China Judgments Online. China Judgments Online, 2013. URL `https://wenshu.court.gov.cn`.

Xin Dai, Buqiang Xu, Zhenghao Liu, Yukun Yan, Huiyuan Xie, Xiaoyuan Yi, Shuo Wang, and Ge Yu. LegalΔ: Enhancing legal reasoning in LLMs via reinforcement learning with chain-of-thought guided information gain. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026. URL `https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf`.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7933–7962, 2024.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Peter Goodrich. *Legal discourse: Studies in linguistics, rhetoric and legal analysis*. Springer, 1990.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2023.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *arXiv preprint arXiv:2409.20288*, 2024.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*, 2025.

Llama Team. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yiran Rex Ma, Yuxiao Ye, and Huiyuan Xie. Clase: A hybrid method for chinese legalese stylistic evaluation. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pp. 642–653, 2026.

Yon Maley. The language of the law. In *Language and the Law*, pp. 11–50. Routledge, 2014.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Mark K Osbeck. What is" good legal writing" and why does it matter? *Drexel L. Rev.*, 2011.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Qiang Dong, Maosong Sun, and Zhendong Dong. Openhownet: An open sememe-based lexical knowledge base. *arXiv preprint arXiv:1901.09957*, 2019.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Jian Sun. jieba, 2012. URL `https://github.com/fxsjy/jieba`.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 2023.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. Lawformer: A pre-trained language model for chinese legal long documents. *arXiv preprint arXiv:2105.03887*, 2021.

Huiyuan Xie, Chenyang Li, Huining Zhu, Chubin Zhang, Yuxiao Ye, Zhenghao Liu, and Zhiyuan Liu. Lexchain: Modeling legal reasoning chains for chinese tort case analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 35913–35921, 2026.

Buqiang Xu, Xin Dai, Zhenghao Liu, Huiyuan Xie, Xiaoyuan Yi, Shuo Wang, Yukun Yan, Liner Yang, Yu Gu, and Ge Yu. LegalDuet: Learning fine-grained representations for legal judgment prediction via a dual-view contrastive learning. In *International Conference on Advanced Data Mining and Applications*, pp. 337–352, 2025.

Rujing Yao, Yang Wu, Chenghao Wang, Jingwei Xiong, Fang Wang, and Xiaozhong Liu. Elevating legal LLM responses: Harnessing trainable logical structures and semantic knowledge with legal reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025.

Kepu Zhang, Guofu Xie, Weijie Yu, Mingyue Xu, Xu Tang, Yaxin Li, and Jun Xu. Legal mathematical reasoning with LLMs: Procedural alignment through two-stage reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1586–1598, 2025.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. JEC-QA: A legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9701–9708, 2020.

A Implementation Details of Rubric-Based Rewards

A.1 Style

Reference and Calibration Corpora. The lexical distributions and linguistic reference statistics are constructed from 1,000 authentic Chinese judgments in the CJO ms corpus (China Judgments Online, 2013). The normalization scales are estimated separately from the LY fields of 986 CJO ms documents. This separation ensures that each raw metric is normalized to a comparable numerical range.

Word Specificity. Legal terms are extracted using a domain-specific terminology lexicon and forward maximum matching (FMM). The maximum and minimum matching lengths are eight and two Chinese characters, respectively. For each term t in the legal vocabulary V_{legal}, the response distribution is calculated as

$$Q_y^{\text{legal}}(t) = \frac{N_y(t)}{\sum_{u \in V_{\text{legal}}} N_y(u)}, \tag{8}$$

where $N_y(t)$ is the number of occurrences of term t in y. Terms absent from the response are assigned $\epsilon = 10^{-10}$ before normalization to avoid undefined KL-divergence values. The complete divergence is

$$d_{\text{ws}}(y) = \sum_{t \in V_{\text{legal}}} P^{\text{legal}}(t) \log \frac{P^{\text{legal}}(t)}{Q_y^{\text{legal}}(t)}. \tag{9}$$

Here, $P^{\text{legal}}(t)$ is the normalized frequency of t in the reference corpus. The normalization scale for this dimension is $s_{\text{ws}} = 12.3071$.

Subjective Word Control. Subjective expressions are identified using the Chinese HowNet sentiment lexicon(Qi et al., 2019) and the same FMM procedure. The reference and response distributions, P^{subj} and Q_y^{subj}, are constructed in the same manner as the legal-term distributions. The resulting KL divergence is normalized using $s_{\text{sw}} = 12.6957$.

Sentence Cohesion. The response is tokenized and POS-tagged using jieba.posseg (Sun, 2012). Tokens with the POS tag c are treated as conjunctions. The reference conjunction rate is

$$\rho^*_{\text{conj}} = 0.0191. \tag{10}$$

Dimension	Raw deviation	Scale s_k
WS	KL divergence	12.307
SW	KL divergence	12.696
COH	Conjunction-rate deviation	0.012
STR	Sentence-statistic distance	26.057
COL	n-gram coverage distance	0.282

Table 4: Normalization scales for the Style reward. WS: Word Specificity; SW: Subjective-Word Control; COH: Sentence Cohesion; STR: Sentence Structure; COL: Collocation Correctness.

Here, ρ^*_{conj} is the mean conjunction-token proportion estimated from the reference documents. The absolute deviation from this value is normalized using $s_{\text{coh}} = 0.011573$.

Pronoun frequency is not included because preliminary experiments showed that it provided little discrimination between authentic and generated responses.

Sentence Structure. Sentences are segmented using Chinese punctuation delimiters and newline characters. Sentence length is measured in Chinese characters. The reference statistics are

$$\mu^* = 64.7, \qquad \sigma^* = 45.3, \tag{11}$$

where μ^* and σ^* are the reference mean and standard deviation of sentence lengths. The Euclidean distance between the response and reference statistics is normalized using $s_{\text{str}} = 26.0567$.

Collocation Correctness. Both the reference documents and generated responses are tokenized using jieba, with punctuation tokens removed before n-gram extraction. A reference n-gram is retained only if it occurs in at least ten reference documents.

For each $n \in \{3, 4, 5, 6\}$, let $G_n(y)$ denote the set of n-grams extracted from response y, and let G^*_n denote the retained reference set. The coverage statistic is

$$c_n(y) = \frac{|G_n(y) \cap G^*_n|}{|G_n(y)|}, \qquad n \in \{3, 4, 5, 6\}. \tag{12}$$

Here, $|G_n(y) \cap G^*_n|$ is the number of response n-grams found in the reference set, while $|G_n(y)|$ is the total number of response n-grams.

The collocation deviation is calculated as

$$d_{\text{col}}(y) = \sqrt{\sum_{n=3}^{6} \left(c_n(y) - \bar{c}^*_n\right)^2}, \tag{13}$$

where $\bar{c}^*_n$ is the mean reference coverage for order n. We use 3-grams through 6-grams in the final metric. Bigrams are excluded because their high frequency and short length result in many generic or non-legal combinations, reducing their ability to characterize professional legal collocations.

Normalization Parameters. The complete set of normalization scales is summarized below:

For every dimension, the normalized reward is $r_k(y) = -d_k(y)/s_k$. The five normalized rewards are then combined using an equal-weight arithmetic mean.

A.2 Element

Fig. 3 presents the prompt template used by the LLM-based *Element* evaluator. The prompt provides the task context, candidate response, and element-specific rubric criteria, and requires the judge to produce a structured assessment of the relevant legal elements. For readability, the prompt shown

Prompt for Element Dimension

```
You are a strict legal answer-matching evaluator. Given a legal question and a candidate answer, assess how well the answer satisfies the question across four Element dimensions: subjects, facts, statutes, and disposition. Use only the question, choices, output requirements, and candidate answer. You cannot see the reference answer. Output valid JSON only, without Markdown.
Core rules:
Score the match between the candidate answer and the question—not merely the presence of legal terms.
Score all four dimensions separately from 0 to 1. Never omit a dimension or output null.
For each dimension, assess accuracy, completeness, specificity, and unsupported content.
If little useful information is available for a dimension, assign 0 or a low score; do not give full credit merely because it is not tested.
For multiple-choice questions, check the answer against both the stem and choices. Penalize missing choices, extra choices, and mutually exclusive conclusions.
Compute final_score as a weighted combination of the four dimensions, giving greater weight to dimensions central to the question.
Dimensions:
subjects
Identify people, organizations, courts, procuratorates, companies, and roles such as plaintiff, defendant, accused, victim, and applicant. Roles without specific subjects score low; specific subjects with unclear roles score moderately; correct subject–role matches score high. Penalize invented subjects.
facts
Identify amounts, dates, places, conduct, outcomes, losses, and expert findings. Exact details score highest. Amounts should receive progressively lower scores for deviations of about 1%, 5%, 10%, 25%, and 50%. Dates should be scored by precision, such as exact date, year-month, or year only. Penalize unsupported numbers, places, and events.
statutes
Identify law names, article numbers, paragraphs/items, and rule content. “According to relevant law” scores very low; a law name alone scores low to medium; an article number alone scores moderately; law name plus article number scores high; law name, article number, and rule content score highest. Apply only a minor penalty when the question requires external legal citations.
disposition
Identify offenses, causes of action, liability, granting or rejection, compensation, sentences, fines, forfeiture, and other outcomes. A label such as “Answer: A” alone scores low; a general conclusion scores moderately; a specific offense, liability finding, or penalty scores high. Penalize contradictory conclusions and over-selection.
Output exactly one JSON object. Keep every reason to one short sentence. All scores and final_score must be numbers from 0 to 1.
{
"subjects": {"score": 0.0, "reason": "Short reason"},
"facts": {"score": 0.0, "reason": "Short reason"},
"statutes": {"score": 0.0, "reason": "Short reason"},
"disposition": {"score": 0.0, "reason": "Short reason"},
"dimension_weights": {
"subjects": 0.0,
"facts": 0.0,
"statutes": 0.0,
"disposition": 0.0
},
"final_score": 0.0,
"extracted_answer": "Core answer extracted from the candidate answer",
"final_reason": "Short explanation of the overall score"
}
Question:
{query}
Candidate Answer:
{answer}
```

Figure 3: English translation of the rubric-guided prompt used for *Element* evaluation. The original experimental prompt was written in Chinese, while its evaluation criteria, scoring procedure, and output structure are preserved in the translation.

in the figure has been translated into English, and the original prompt used in all experiments was written in Chinese.

A.3 CHAIN

The *Chain* reward uses heuristic scoring when task-specific scoring logic is available. We evaluate four criteria: *Order*, which detects violations of required step precedence; *Completeness*, which measures coverage of necessary reasoning steps; *Correctness*, which checks compliance with step-level reasoning requirements; and *Non-redundancy*, which identifies repeated or conflicting steps. Detected omissions or violations are mapped to discrete criterion-level rewards, which are averaged to obtain the final Chain reward.

When task-specific scoring logic is unavailable, we use an LLM judge to score the response along the same four criteria using the prompt in Fig. 4. The criterion-level scores are averaged to obtain

Prompt for Chain Dimension

You are a rigorous reviewer of legal reasoning chains. Your task is to assess the quality of a reasoning chain against the source text and the reasoning steps extracted from it.

Evaluation Dimensions

Evaluate the following four dimensions independently. Report only defects that can be clearly located:

C1 Logical Order

Premises should precede inferences that depend on them, and reasons should precede conclusions. Differences in writing style alone do not constitute an order error. If Step B depends on the conclusion of Step A, A should appear before B.

C2 Reasoning Completeness

Determine whether any essential reasoning step is missing between the available premises and the final conclusion. Require only steps relevant to the current conclusion, not unrelated template fields.

C3 Step Correctness

Check whether each step name accurately summarizes the corresponding source text and whether each inference is supported by facts or legal grounds in the source. A summary that distorts the source or an unsupported inference is a defect.

C4 Non-redundancy

Check for pointless repetition or substantive contradictions between steps. Reasonable restatements, such as recalling a fact when later citing it, and progressive analysis are not repetitions.

Important Rules

1. Report each issue only once, under the most directly relevant dimension. Do not duplicate it across dimensions.
2. Step names are free-form summaries. Do not penalize a name merely because it differs from a standard legal-document template.
3. Use an empty array for a dimension with no defects. Do not invent defects.

Output Format

Output JSON only, with no other text. Use this structure:

```
{
  "issues": {
    "C1": [],
    "C2": [],
    "C3": [],
    "C4": []
  }
}
```

For each defect, use {"positions": [step numbers], "reason": "specific issue"}. Example:

```
{"issues": {
  "C1": [{"positions": [3, 5], "reason": "Step 3 relies on Step 5, but appears before it"}],
  "C2": [{"positions": [4], "reason": "The causal link between the breach and the loss is missing"}],
  "C3": [{"positions": [2], "reason": "The step says fraud was established, but the source only suggests it"}],
  "C4": [{"positions": [1, 6], "reason": "Steps 1 and 6 are substantively identical"}]
}}
```

Figure 4: English translation of the reasoning-step extraction prompt used for *Chain* evaluation when predefined reasoning steps are unavailable. The LLM converts free-form reasoning into an ordered sequence of labeled steps, which is subsequently scored by heuristic rubrics. The original experimental prompt was written in Chinese.

the final reward. The displayed prompt is translated into English; the original experimental prompt was written in Chinese.

B Experimental Settings

B.1 Datasets

For Style, we use the official CLASE test set of 1,000 response pairs and construct preference data from all 4,000 training queries using our five-model generation procedure. We reserve 2,000 training instances as the reference corpus for computing the objective component of CLASE-Mix and split the remaining 2,000 queries equally between SFT and GRPO. No model is trained on

authentic judicial texts: SFT targets and preferred responses in preference pairs are model-generated candidates selected by their rubric scores, while GRPO uses the learned reward model to score online generations.

For Element, we use all 2,736 JEC-QA and LexChain instances for preference construction, yielding 2,244 preference pairs, and designate separate subsets for SFT and GRPO. Evaluation uses the 3,000 instances in the LegalΔ in-domain test set, comprising 500 instances each for statutory-article prediction, charge prediction, case analysis, and financial calculation, and 1,000 for sentence-length prediction.

For Chain, we allocate 1,000 of the 9,550 LexChain training instances to SFT, 1,000 to GRPO, and 4,000 to pairwise preference training, and evaluate on the official test set of 1,000 instances.

Across all three dimensions, we construct preference data by sampling one response per query from each model in the pool, scoring the five candidates with the corresponding rubric, and forming pairs from responses with different scores. Queries for which all candidates receive identical scores are discarded.

B.2 Hyper-parameters

Inference. For LLM-based rubric evaluation we use the DeepSeek-V4-Flash(DeepSeek-AI, 2026) API with the default reasoning effort and a maximum input length of 8,192 tokens; all inputs fit within this limit, requiring no truncation. Candidate responses are sampled with temperature 0.8 and top-p 0.95, one response per model.

Evaluation. CLASE-Mix combines objective and subjective scores with equal weights. The objective component compares textual features of generated responses with those of authentic judicial texts, while the subjective component uses GPT-4o-mini to score responses with the benchmark-provided prompt. For LexChain, we use GPT-4o-2024-05-13 (OpenAI, 2024) to score each evaluation dimension following the benchmark protocol. For LegalΔ, we report F1 scores for statutory-article and charge prediction, and accuracy for case analysis, financial calculation, and sentence-length prediction. We follow the evaluation procedures specified in the respective benchmark papers; further details can be found therein.

Training. Reward models are trained with the Bradley–Terry objective (Bradley & Terry, 1952) for two epochs using DeepSpeed ZeRO-3, a learning rate of 1×10^{-7} and a batch size of 32; DPO uses the same configuration. We train policies separately for Style, Element, and Chain. For each dimension, we first perform SFT on 1,000 queries paired with their highest-scoring candidate responses under the corresponding rubric, then apply GRPO to another 1,000 queries. Both GRPO variants start from the same dimension-specific SFT checkpoint: GRPO (RM) uses the corresponding dimension-specific LexRM as its reward function, whereas GRPO (rule) uses the baseline reward described in Section 5.2.2. Each dimension block in Table 2 reports results from the respective policies. DPO is initialized directly from Qwen3-8B. SFT, DPO and GRPO all use low-rank adaptation (LoRA).

C Limitations

Our taxonomy is grounded in Chinese legal materials input from the Chinese legal context, and our evaluation is therefore limited to Chinese-language datasets. Its applicability to other languages and legal systems remains untested. Future work will extend the framework to these settings and examine how the taxonomy and its reward criteria should be adapted.

This work primarily studies dimension-specific rewards. Combining multiple dimensions into a unified rubric reward or reward model remains an open challenge, including the choice of aggregation weights, the composition of preference data, and the training strategy. Developing and systematically evaluating such integration methods is a central direction for future work.