Beyond Pointwise Scores: Decomposed Criteria-Based Evaluation of LLM Responses

Fangyi Yu[‡], **Nabeel Seedat**[‡], **Dasha Herrmannova**[§],
Frank Schilder[§], **Jonathan Richard Schwarz**[‡]
[‡]Thomson Reuters Foundational Research [§]Thomson Reuters Labs
firstname.lastname@thomsonreuters.com

Abstract

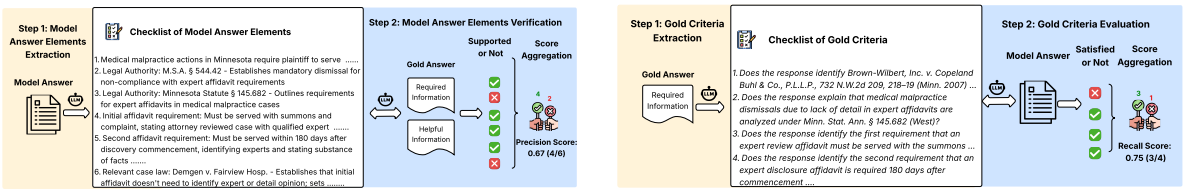
Evaluating long-form answers in high-stakes domains such as law or medicine remains a fundamental challenge. Standard metrics like BLEU and ROUGE fail to capture semantic correctness, and current LLM-based evaluators often reduce nuanced aspects of answer quality into a single undifferentiated score. We introduce **DeCE**, a decomposed LLM evaluation framework that separates *precision* (factual accuracy and relevance) and *recall* (coverage of required concepts), using instance-specific criteria automatically extracted from gold answer requirements. DeCE is model-agnostic and domain-general, requiring no predefined taxonomies or handcrafted rubrics. We instantiate DeCE to evaluate different LLMs on a real-world legal QA task involving multi-jurisdictional reasoning and citation grounding. DeCE achieves substantially stronger correlation with expert judgments (r=0.78), compared to traditional metrics (r=0.12), pointwise LLM scoring (r=0.35), and modern multidimensional evaluators (r=0.48). It also reveals interpretable trade-offs: generalist models favor recall, while specialized models favor precision. Importantly, only 11.95% of LLM-generated criteria required expert revision, underscoring DeCE's scalability. DeCE offers an interpretable and actionable LLM evaluation framework in expert domains.

1 Introduction

As large language models (LLMs) are increasingly deployed in high-stakes, expert settings, such as law, medicine, and finance; their outputs must satisfy demanding requirements: factual accuracy, citation support, and coverage of domain-specific obligations (Lai et al., 2024a; Wang et al., 2024; Zhang et al., 2024a). However, evaluating such complex, long-form responses remains a fundamental challenge (Liang et al., 2023; Chang et al., 2024). Evaluation failures in these domains carry real consequences, including tangible harm, legal liability and erosion of trust in AI systems. To anchor ideas, consider legal question answering, where a lawyer might ask: "Does a $2B acquisition of a competitor trigger antitrust filing obligations in California?" An effective LLM answer must synthesize statutes, regulations, and case law. Subsequent evaluation is hence a fundamentally open-ended, multi-dimensional challenge.

Unfortunately, existing evaluation paradigms fail to address this challenge. ***Human expert reviews***, though considered the gold standard, are costly and unscalable (e.g., it can take a legal expert around 45 minutes to evaluate one model response by thoroughly checking citations, verifying legal reasoning, and assessing applicability). This approach becomes prohibitively expensive for large-scale evaluation across multiple models and tasks. ***Lexical metrics*** like ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) are cheap but correlate poorly with human judgment on complex, knowledge-intensive tasks. ***LLM-as-a-judge*** has emerged as a promising alternative (Zheng et al., 2023; Li et al., 2025; Gu et al., 2025), yet many implementations reduce evaluation to a single pointwise score, obscuring actionable insights. Recent multidimensional variants (e.g., GPTScore (Fu et al., 2024a) , G-Eval (Liu et al., 2023a)) add labeled axes (e.g., accuracy, completeness) and show improved alignment in general domains; however, they typically rely on generic, task-agnostic criteria that miss domain-specific obligations and hierarchies central to expert settings like law. ***Checklist-based approaches*** (e.g., LLM-Rubric (Hashemi et al., 2024), CheckEval (Lee et al., 2024)) increase granularity but require manual rubrics, taxonomies, or seed questions. Hence, while these methods provide greater granularity than naive LLM-as-a-judge, they require substantial manual effort or lack systematic decomposition into dimensions such as precision (factual accuracy, rele-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1931–1954
November 4-9, 2025



(a) Precision Workflow

(b) Recall Workflow

Figure 1: Overview of the DeCE evaluation pipeline. **(a)** The precision workflow decomposes the model-generated answer into factual elements, which are then individually verified for factual correctness and relevance against the gold answer. **(b)** The recall workflow extracts evaluation criteria from the gold answer Required Information and checks whether each criterion is satisfied in the model response. Together, these workflows yield decomposed scores that provide interpretable evaluation signals for expert-domain model evaluation.

vance) and recall (coverage), which are essential in expert domains to diagnose specific model behaviors. RAGCHECKER (Ru et al., 2025) advances claim-level precision/recall via bidirectional entailment over extracted claims but treats claims uniformly, overlooking requirement hierarchies. Beyond claim overlap, AQuAECHR (Weidinger et al., 2025) and ALCE (Gao et al., 2023) measure precision and recall of citations to assess citation accuracy and coverage in legal contexts, but are limited to citation verification. A comparison of our work vs related works is illustrated in Table 1.

To address these gaps, we propose DeCE, an LLM-based evaluation framework that decomposes evaluation into two interpretable dimensions: precision (factual accuracy and relevance) and recall (coverage of Required Information). DeCE leverages gold-standard answers, typically available in high-stakes domains. Rather than collapsing evaluation into a single opaque score, DeCE automatically extracts instance-specific, domain-aware criteria and uses them to perform a structured precision–recall decomposition that verifies both factual grounding and requirement satisfaction beyond citations.

We include direct comparisons to lexical metrics, pointwise LLM-as-a-judge scoring, GPTScore, G-Eval, and RAGCHECKER in our experiments, and find they align worse with legal expert judgments than DeCE, indicating reduced effectiveness in highly specialized, high-stakes evaluation. See Appendix A for extended related work.

Contributions. (1) We propose **DeCE**, a decomposed criteria-based evaluation framework for LLM evaluation. (2) We show that DeCE aligns significantly better with human expert judgments than standard lexical metrics, holistic LLM-judge scores and multidimensional LLM-as-a-judge baselines, achieving correlation scores up to $r = 0.78$. (3) We evaluate five diverse frontier LLMs and reveal precision-recall trade-offs. (4) We demonstrate that DeCE enables fine-grained model behavior analysis, identifying systematic weaknesses of frontier LLMs across jurisdictions and query types. (5) We validate the reliability of DeCE criteria, finding that only 11.95% require revision, demonstrating DeCE is scalable and deployable with minimal expert supervision.

2 Decomposed Criteria-Based Evaluation

We introduce **Decomposed Criteria-Based Evaluation (DeCE)**, a structured LLM-based evaluation framework. Unlike scalar pointwise scoring methods, DeCE decomposes evaluation into two orthogonal, interpretable dimensions: *Precision*: the factual accuracy and relevance of claims made in the model-generated answer and *Recall*: the completeness of the model answer with respect to the Required Information in a gold reference answer.

2.1 Problem Formulation

Let each evaluation instance be a tuple (q, a_g, a_m), where $q \in \mathcal{Q}$ is a question, $a_g \in \mathcal{A}_g$ is a gold standard answer with Required Information a_{gr} and supportive case laws in Helpful Information a_{gh}, and $a_m \in \mathcal{A}$ is the model-generated answer.

DeCE computes a decomposed evaluation score:

$$\text{DeCE}(q, a_g, a_m) = (P(q, a_g, a_m), R(q, a_{gr}, a_m))$$

where P and R denote precision and recall, derived through automated workflows using an LLM judge comparing a_m elements against criteria from a_g.

Motivation and practicality. Legal analysis prioritizes authorities by hierarchy and direct applicability (e.g., constitutions/statutes > regulations > cases; controlling > persuasive). Expert review therefore distinguishes between requirements

Table 1: Comparison of evaluation frameworks across key dimensions. ✓ indicates full support, △ indicates partial support, and × indicates no support.

Method	Auto Criteria	Domain-Aware	Inst.-Level Adapt.	Decomp. Eval	Prec/Recall	Manual Taxo. Free	Crit. Interpret.	Mod. Diagnostics	Hierarchy-Aware	Scalable
Traditional Metrics										
ROUGE / BLEU	×	×	×	×	×	✓	△	×	×	✓
BERTScore / MoverScore	×	×	×	×	×	✓	△	×	×	✓
LLM-as-a-Judge										
G-Eval/GPTScore	×	×	×	✓	×	✓	△	×	×	✓
Pointwise Judge (Likert)	×	×	×	×	×	✓	×	×	×	✓
Criteria-Based LLM Judge										
CheckEval	△	✓	×	✓	×	×	✓	×	×	△
LLM-Rubric	×	✓	×	✓	×	×	△	×	×	×
Human Expert Review	×	✓	✓	△	×	△	✓	✓	×	×
RAGCHECKER	✓	△	✓	△	✓	✓	✓	✓	×	×
DeCE (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

that must be satisfied and supporting material that strengthens interpretation. This motivates separating gold answers into "Required" (directly governing) and "Helpful" (supportive/persuasive), which our recall and precision workflows leverage: recall evaluates satisfaction of criteria extracted from a_{gr}, while precision verifies the factual support of elements in a_m against a_g.

2.2 DeCE Evaluation Pipeline

DeCE consists of two self-contained workflows - *Precision Scoring* and *Recall Scoring* - as illustrated in Fig. 1. Prompts for each step are detailed in Appendix C.4, and hyperparameter settings are provided in Appendix C.5.

1. Precision Scoring

(a) Answer Element Extraction. The model answer a_m is decomposed into factual elements:

$$\mathcal{E}_m = \texttt{ExtractElements}(a_m) = \{e_1, e_2, \ldots, e_l\},$$

where each e_j denotes a requirement, principle, or legal authority in a_m.

(b) Element Verification. Each element e_j is verified against the gold answer a_g. The precision score is defined as:

$$P(q, a_g, a_m) = \frac{1}{|\mathcal{E}_m|} \sum_{j=1}^{|\mathcal{E}_m|} \mathbb{I}\left[\text{supported}(e_j, a_g)\right],$$

where $\mathbb{I}[\text{supported}(e_j, a_g)] = 1$ if e_j is supported by a_g, 0 otherwise. This reflects the proportion of model claims grounded in the gold answer.

2. Recall Scoring

(a) Criteria Extraction. We extract evaluation criteria from Required Information a_{gr} (excluding additional supportive case laws in a_{gh} as non-essential for completeness measurement):

$$\mathcal{C}_g = \texttt{ExtractCriteria}(a_{gr}) = \{c_1, c_2, \ldots, c_k\},$$

where c_i represents a query-specific requirement.

(b) Criteria Satisfaction. The recall score is:

$$R(q, a_g, a_m) = \frac{1}{|\mathcal{C}_g|} \sum_{i=1}^{|\mathcal{C}_g|} \mathbb{I}\left[\text{satisfies}(a_m, c_i)\right],$$

where $\mathbb{I}[\text{satisfies}(a_m, c_i)] = 1$ if a_m satisfies criterion c_i, 0 otherwise. This quantifies how fully the model answer covers the essential concepts of the gold answer requirements.

Remark. This decomposition enables interpretable evaluation and precise failure attribution, valuable in expert domains.

3 Experiments

We evaluate DeCE across multiple dimensions: alignment with expert judgment, model-specific diagnostic insights, and reliability of criteria extraction. We instantiate DeCE on legal question

answering, a high-stakes domain where LLMs increasingly assist professionals (Lai et al., 2024b). DeCE is applicable to other domains with structured, expert-authored outputs (clinical QA, financial compliance, scientific summarization, etc.).

Dataset. To assess real-world applicability, we use a professionally curated dataset of 224 English legal QA pairs spanning diverse U.S. state and federal jurisdictions[1]. While 224 examples may appear modest compared to general-domain QA datasets, each instance reflects high annotation cost in expert domains—often requiring up to an hour of expert time for curation or evaluation. Hence, this scale is representative of real-world resource constraints in expert domains. Each question is written by expert Attorney Editors and paired with a gold-standard answer that delineates *Required Information* (essential content for recall) and *Helpful Information* (supportive authorities). Relevant legal documents are retrieved using a retrieval system and held constant across all models to mitigate the influence of retrieval. Dataset details are provided in Appendix B.

LLMs evaluated. We assess five LLMs generating legal answers under standardized retrieval. These include: general-purpose GPT-4o (OpenAI); reasoning models Gemini-2.5-Pro (Comanici et al., 2025) and DeepSeek-R1 (DeepSeek-AI et al., 2025); open-source model Llama-3.1-405B (Grattafiori et al., 2024); and domain-specific model, Legal Llama-3.1-70B, which we fine-tuned using supervised learning (Zhang et al., 2024b) and Direct Preference Optimization (Rafailov et al., 2023) on legal corpora. All models have over 70 billion parameters[2].

Evaluation method baselines. We compare DeCE (using Claude 3.5 Sonnet (Anthropic, 2024) as the backbone LLM) against four categories of evaluation methods: (1) **Lexical overlap metrics.** ROUGE-L and BLEU: standard overlap metrics that are computationally cheap but correlate weakly with semantic correctness on complex generation tasks. (2) **Pointwise LLM-as-a-judge.** Following prior work (Zheng et al., 2023; Li et al., 2025), we prompt an LLM (Claude 3.5 Sonnet) to assign 0–4 Likert scores relative to the gold answer using a rubric-driven prompt with chain-of-thought (Wei et al., 2023); scores are normalized to a GPA-style 0–4 metric (see prompt in Appendix C.3). (3) **Multidimensional LLM-as-a-judge.** GPTScore (Fu et al., 2024b) and G-Eval (Liu et al., 2023a) serve as representative multi-axis judges. For comparability, we: (i) use the same backbone LLM (Claude 3.5 Sonnet), (ii) provide the gold answer as reference , and (iii) report the same dimensions—precision and recall (as in DeCE). (4) **Claim-level**. RAGCHECKER (Ru et al., 2025) which computes claim-level precision/recall via bidirectional entailment. We use author-recommended settings and provide the same gold answers and retrieved materials for consistency. We use Claude 3.5 Sonnet for claim extraction for a fair comparison with DeCE.

Extended implementation details and prompts are provided in Appendix C.

3.1 Does DeCE Align with Human Experts?

Goal. Does DeCE's decomposed metrics better align with legal expert judgments compared to the baseline automatic metrics?

Setup. Four U.S. legal experts with 10+ years of practice/academic experience evaluated model responses. For pointwise evaluation, experts assessed all 224 GPT-4o responses using our LLM judge rubric. For decomposed evaluation, experts assessed responses from all five models on 20 randomly selected queries (100 annotations total), evaluating recall (criterion satisfaction) and precision (factual accuracy/relevance). We compute Pearson and Spearman correlations between human judgments and automated metrics. Following established practices in specialized domain evaluation where expert annotation is costly, we employed single-annotator protocol to maximize coverage breadth over agreement assessment.

We use F2 as our primary correlation benchmark, weighting recall over precision for two empirically-motivated reasons: (1) DeCE recall shows stronger correlation with human recall ($r = 0.80$) than DeCE precision with human precision ($r = 0.69$) (see Appendix C.6), and (2) legal queries often have multiple valid citations beyond gold references, making precision inherently noisy—F2's recall emphasis better captures comprehensive legal reasoning quality while mitigating false negatives from incomplete gold standards.

[1] Due to proprietary constraints, the data cannot be publicly released.

[2] GPT-4o, DeepSeek-R1, and Llama-3.1-405B were accessed via AWS Bedrock API; Gemini-2.5-Pro via Google Vertex AI API, in accordance with each provider's terms of service and consistent with their intended use. We fine-tuned Llama-3.1-70B under Meta's Llama 3 Community License. The finetuning specifics of the Legal Llama-3.1-70B model are out of scope for this paper.

Table 2: Correlation coefficients between automated metrics and human expert F2. A full correlation comparison is provided in Appendix C.6.

Metric Pair	Pearson	Spearman	P-Value
ROUGE-L vs Human	0.11	0.15	0.29
BLEU vs Human	0.12	0.13	0.13
Point. Judge vs Human	0.35	0.37	< 0.05
GPTScore F2 vs Human	0.48	0.39	< 0.05
G-Eval F2 vs Human	0.42	0.34	< 0.05
RAGCHECKER vs Human	0.38	0.31	< 0.05
DeCE F2 vs Human	**0.78**	**0.76**	< 0.05

Analysis. Table 2 shows that ROUGE-L and BLEU exhibit weak correlation with expert assessments (Pearson $r = 0.11$ and 0.12), confirming their inadequacy for legal responses where semantic correctness and legal reasoning matter more than surface similarity. Pointwise LLM-as-a-judge improves alignment moderately ($r = 0.35$). Multidimensional LLM-as-a-judge baselines further increase correlation (GPTScore F2: $r = 0.48$; G-Eval F2: $r = 0.42$), and RAGCHECKER attains $r = 0.38$. However, DeCE achieves substantially higher alignment (F2: $r = 0.78$), indicating that instance-specific, domain-aware criteria and explicit precision–recall decomposition capture expert signals that generic multi-axis rubrics and claim-overlap methods miss.

Takeaway. DeCE significantly outperforms existing metrics in aligning with expert judgment, validating its utility as a reliable and low-cost proxy for human evaluation in high-stakes domains.

3.2 What trade-offs exist across LLMs?

Goal. We evaluate five different LLMs and aim to highlight insights into their legal answer generation capabilities on the basis of the evaluations.

Setup. We first examine pointwise GPA scores assigned to the generations from the five frontier LLMs. This allows comparison of holistic quality as measured by Likert ratings. We then analyze the decomposed precision and recall distributions produced by DeCE for the same models to investigate trade-offs in factual accuracy and coverage. Note we exclude traditional metrics (ROUGE-L and BLEU) from performance analysis due to their weak correlation with human judgments.

Analysis. Fig. 2 shows Gemini-2.5-Pro achieved the highest GPA (3.56) with 71.3% "Excellent" ratings, followed by DeepSeek-R1 (3.42 GPA, 60.7% excellent). GPT-4o was balanced (3.21 GPA), while Legal Llama-3.1-70B trailed with only 17.0% excellent responses, suggesting scale may outweigh domain specialization.

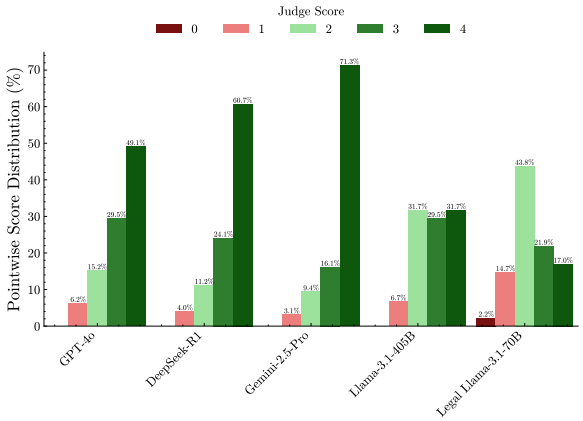


Figure 2: Distribution of pointwise scores (0–4) for each model, judged by Claude 3.5 using rubric-based Likert evaluation. Gemini-2.5-Pro achieves the highest proportion of top-rated responses (71.3%), while legally fine-tuned Llama-3.1-70B shows lower scores, suggesting model scale may outweigh domain specialization for complex legal reasoning.

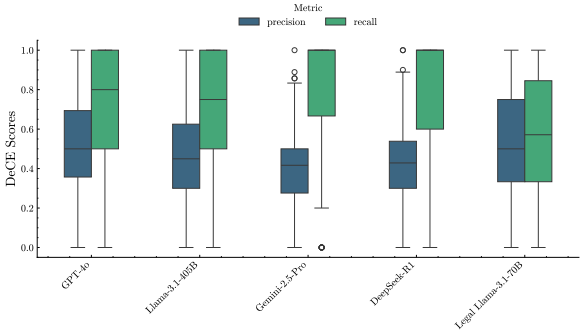


Figure 3: DeCE scores (precision and recall) for each evaluated model. Larger generalist models (e.g., Gemini, GPT-4o) demonstrate stronger recall, while legally fine-tuned models exhibit higher precision, highlighting complementary strengths.

Fig. 3 reveals insights into precision-recall trade-offs surfaced by our decomposed approach: Gemini-2.5-Pro and DeepSeek-R1 excel at recall (median $\sim$1), indicating strong comprehensiveness, but have lower precision (median $\sim$0.42). In contrast, Legal Llama-3.1-70B achieves the highest precision (median $\sim$0.50) but at the cost of recall. GPT-4o and Llama-3.1-405B perform moderately across both axes.

Takeaway. While GPA scores suggest general performance rankings, DeCE reveals a precision and recall trade-off. Larger models tend to provide more comprehensive (higher recall) but occasionally inaccurate answers (lower precision), whereas smaller or specialized models are more precise but less complete. DeCE permits to diagnose such trade-offs, which holistic metrics obscure.

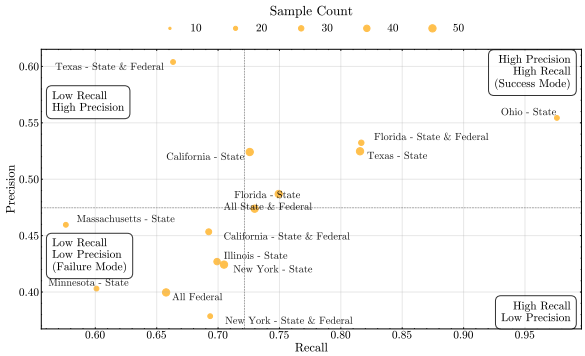


Figure 4: Model performance across jurisdictions insights (precision vs. recall). Ohio State achieves high performance (recall: 0.98, precision: 0.55), while Texas State and Florida State & Federal show strong balanced performance. New York State & Federal exhibits low precision (0.38) despite moderate recall, and Minnesota State falls into the failure quadrant with both low precision and recall.

3.3 What Insights does DeCE Reveal About Model Behavior?

3.3.1 LLM-Generated Evaluation Criteria Reliability

Goal. Determine whether an LLM judge can reliably extract evaluation criteria from expert gold answers with minimal human revision.

Analysis. Expert review of all 979 criteria for the 224-question corpus shows strong reliability. For 54.5% of queries the criteria were accepted verbatim. At the granular level only 11.95% of individual criteria were modified, 0.7% were discarded, and 2.0% were added to capture overlooked nuances. Revisions fell into three refinement patterns: *specificity calibration* (tightening imprecise wording), *legal authority differentiation* (separating conflated statutory and case-law requirements), and *case-law flexibility* (allowing alternative valid precedents). These low intervention rates indicate that LLM-generated criteria provide a sound foundation for decomposed evaluation while substantially reducing expert labour. A detailed criteria evaluation analysis is provided in Appendix D.

Takeaway. LLM-driven criterion extraction is sufficiently accurate for large-scale use, relegating human experts to a light-touch validation role.

3.3.2 Jurisdictional Performance Patterns

Goal. Identify strengths and weaknesses across U.S. jurisdictions using DeCE scores.

Analysis. We analyze model performance, observing variation across jurisdiction (Fig. 4). Ohio-State queries are handled exceptionally well (average recall 0.98, precision 0.55), while Texas-State and Florida-State show similarly balanced results. New York-State shows a different failure mode: recall remains moderate but precision drops sharply, signalling unsupported or outdated citations. Minnesota-State is the most challenging, with deficits on both axes that place all models in the failure quadrant.

Takeaway. Jurisdiction materially affects model behavior. Outputs concerning New York and Minnesota warrant scrutiny, whereas Ohio-related queries can be addressed with greater confidence.

3.3.3 Query-Type Performance Patterns

Goal. Assess which categories of legal questions are well-handled and which remain problematic.

Analysis. Model performance varies by query type (Fig. 5). Basic concept inquries achieve near-optimal recall (0.87) with respectable precision (0.55). In contrast, source-specific requests and those requiring legal reasoning capabilities remain difficult: recall ($\pm$0.57) and precision (<0.4).

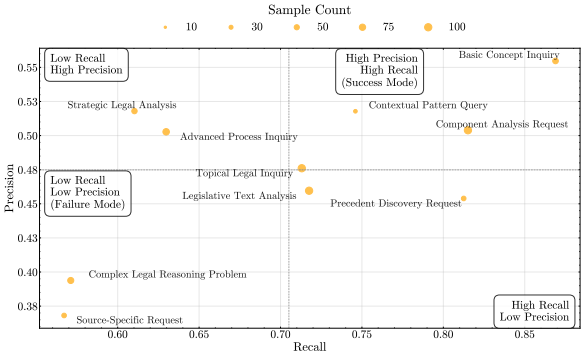


Figure 5: Model performance across query types (precision vs. recall). Basic concept inquiries achieve optimal performance (recall: 0.87, precision: 0.55), while source-specific requests show the poorest results (recall: 0.57, precision: 0.37). Complex legal reasoning problems consistently challenge all models, highlighting fundamental gaps in legal reasoning capabilities.

Takeaway. Source-specific and legal reasoning queries expose persistent weaknesses in current systems and should be prioritized for data augmentation and methodological improvement.

3.3.4 Cross-Model Challenge Analysis

Goal. Highlight challenges across models.

Analysis. Aggregating error slices for GPT-4o, Gemini 2.5-Pro, DeepSeek-R1, and both Llama variants reveals a shared pattern of difficulty. Massachusetts and Minnesota consistently yield low precision and recall across models, suggesting fundamental challenges rather than architectural limitations. The same holds for source-specific

requests, and complex legal reasoning problems, which reduce both metrics across models.

Remark: These findings suggest several practical implications: (i) *complementary deployment* strategies could route queries based on jurisdiction and query type to leverage model-specific strengths; (ii) *targeted improvement* efforts could focus on consistently challenging areas like Massachusetts/Minnesota jurisdictions and source-specific requests; and (iii) *confidence calibration* systems could provide uncertainty indicators for known challenging categories.

Takeaway. The convergence of failure modes across five diverse models indicates systemic limitations in current LLMs for legal QA. Targeted corpus expansion and human-in-the-loop routing for these challenging cases will likely yield greater benefits than further model scaling alone.

4 Discussion

We presented DeCE, a decomposed criteria-based evaluation framework tailored to high-stakes domains like law. By separately assessing *precision* (factual accuracy and relevance) and *recall* (coverage of required concepts) based on automatically extracted criteria from gold answer requirements and elements from model answers, DeCE offers a scalable and interpretable alternative to existing evaluation methods.

Our results demonstrate that DeCE achieves substantially stronger alignment with human expert judgment than traditional lexical metrics ($r = 0.12$) and pointwise LLM-based scoring ($r = 0.35$), with decomposed precision, recall, and F2 reaching correlations of $r = 0.69$, $r = 0.80$, and $r = 0.78$, respectively. Applying DeCE across five frontier LLMs reveals distinct precision–recall trade-offs: larger general models favor recall over accuracy, while legally fine-tuned models yield higher precision but lower completeness. DeCE enables diagnostic insights across jurisdictions (e.g., underperformance in Minnesota and Massachusetts) and query types (e.g., source-specific and legal reasoning requests), revealing challenges common across all LLMs.

Overall, by exposing these nuanced strengths and weaknesses, DeCE enables targeted improvements. The effectiveness of decomposed evaluation establishes a foundation for more sophisticated, domain-sensitive and nuanced evaluation.

5 Limitations

Dataset Scale and Scope. Our evaluation dataset, while carefully curated by legal experts, comprises 224 question-answer pairs focused primarily on U.S. jurisdictions. This scale, though sufficient for demonstrating evaluation framework effectiveness and achieving statistically significant correlations with human expert judgments, represents opportunities for broader validation across international legal systems and specialized practice areas.

Gold Answer Exhaustiveness Assumption. Our precision evaluation assumes that gold answers sufficiently cover the space of correct support, particularly with respect to supportive authorities and case law. In practice, legal questions often admit multiple valid lines of reasoning and alternative authorities, and even expert-authored references are not exhaustive. Consequently, precision may underestimate model capability when a response relies on doctrinally valid but unlisted authorities or argument paths. To mitigate this, we outline two complementary extensions that retain DeCE's decomposed design while better accommodating legitimate multiplicity: 1) Flexible authority matching. During precision verification, treat doctrinally equivalent precedents as supporting when they establish the same controlling rule or holding, even if not explicitly listed in the gold answer (subject to jurisdiction and temporal validity). This reduces false negatives arising from alternative but correct citations. 2) Human-in-the-loop promotion. Introduce a lightweight adjudication pass that reviews frequently recurring, reasonable "false positive" authorities surfaced by DeCE. When validated, these are promoted into the Helpful pool for subsequent runs, expanding coverage without conflating Required content. These extensions preserve interpretability of the precision–recall decomposition—recall continues to target Required criteria, while precision remains a claim-grounding check—yet improve robustness where the gold reference is incomplete.

Evaluation Framework Scope. Our decomposed approach focuses on precision and recall dimensions, which capture completeness and factual accuracy—the primary concerns identified by legal experts in our validation study. While other aspects such as argumentation quality and legal writing style are valuable, our framework addresses the core evaluation challenges that correlate most strongly with expert assessment, providing a solid

foundation for future extensions.

Model Selection Constraints. Our evaluation includes representative models across different types (standard, reasoning-optimized, domain-specialized), though broader inclusion of legal-specific larger models with reasoning capability would strengthen generalizability.

References

Anthropic. 2024. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku. https://www.anthropic.com/news/3-5-models-and-computer-use. Accessed: 2025-03-27.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024a. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024b. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

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.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. 2024a. Large language models in law: A survey. *AI Open*.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2024b. Large language models in law: A survey. *AI Open*, 5:181–196.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, and Pilsung Kang. 2024. Checkeval: Robust evaluation framework using large language model via checklist. *arXiv preprint arXiv:2403.18771*.

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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *Preprint*, arXiv:2411.16594.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2025. Ragchecker: a fine-grained framework for diagnosing retrieval-augmented generation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Jiaqi Wang, Huan Zhao, Zhenyuan Yang, Peng Shu, Junhao Chen, Haobo Sun, Ruixi Liang, Shixin Li, Pengcheng Shi, Longjun Ma, Zongjia Liu, Zhengliang Liu, Tianyang Zhong, Yutong Zhang, Chong Ma, Xin Zhang, Tuo Zhang, Tianli Ding, Yudan Ren, and 3 others. 2024. Legal evalutions and challenges of large language models. *Preprint*, arXiv:2411.10137.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Korbinian Q. Weidinger, Santosh T.y.s.s, Oana Ichim, and Matthias Grabmair. 2025. AQuAECHR: Attributed question answering for European court of human rights. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1418–1447, Vienna, Austria. Association for Computational Linguistics.

Ruizhe Zhang, Haitao Li, Yueyue Wu, Qingyao Ai, Yiqun Liu, Min Zhang, and Shaoping Ma. 2024a. Evaluation ethics of llms in legal domain. *arXiv preprint arXiv:2403.11152*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024b. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Appendix

A Extended Related Work

A.1 Human Expert Review

Human expert evaluation remains the gold standard for specialized domains, providing deep domain knowledge and the ability to adapt evaluation criteria to specific question requirements (Lai et al., 2024a; Wang et al., 2024). Expert reviewers naturally perform instance-level adaptation, adjusting their evaluation focus based on the specific legal question, jurisdiction, and context. They can provide interpretable feedback and implicitly assess both precision (accuracy of provided information) and recall (completeness relative to expert expectations). However, human expert review faces fundamental scalability limitations due to cost and expertise requirements (Liang et al., 2023).

A.2 Traditional Evaluation Paradigms

Early text generation evaluation relied heavily on lexical overlap metrics such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002), which measure surface-level similarity between generated and reference texts. These approaches fundamentally fail in expert domains because lexical overlap correlates poorly with human judgments about text quality (Chang et al., 2024). Human evaluators prioritize semantic meaning, factual accuracy, and domain-specific correctness over word-level similarity.

A.3 LLM-as-a-Judge

The emergence of large language models as evaluators represented a paradigm shift (Zheng et al., 2023; Li et al., 2025). Approaches like G-Eval (Liu et al., 2023b) and GPTScore (Fu et al., 2024b) demonstrated that LLMs could achieve better correlation with human judgments than traditional metrics when prompted with evaluation criteria and chain-of-thought reasoning. However, these approaches rely on generic, manually-designed criteria (e.g., relevance, usefulness, clarity) that cannot adapt to domain-specific nuances or individual question characteristics. They treat all evaluation instances identically, missing opportunities for instance-level adaptation and providing minimal interpretability for targeted system improvements.

A.4 Criteria-Based LLM Evaluation Frameworks

Recent work has recognized that complex evaluation tasks benefit from decomposition into finer-grained criteria. LLM-RUBRIC (Hashemi et al., 2024) introduced personalized calibration networks that account for individual judge preferences through multi-dimensional frameworks. However, it relies entirely on manually authored rubrics and requires substantial human annotation data to train calibration networks, limiting scalability.

CheckEval (Lee et al., 2024) addresses this by using LLMs to generate rubrics, but still requires humans to define evaluation taxonomies and provide seed questions. Critically, CheckEval uses fixed criteria across all instances within a domain, missing instance-level adaptation opportunities, and ultimately collapses scores into single values, preventing systematic decomposition and error analysis.

A.5 Claim-Based Evaluation: RAGCHECKER

RAGCHECKER (Ru et al., 2025) proposes a fine-grained evaluation framework for retrieval-augmented generation (RAG) (Lewis et al., 2020) that computes precision and recall over extracted atomic claims. It extracts factual claims from both model output and gold answers, then applies bidirectional entailment to determine which claims are supported or missing.

However, RAGCHECKER has two key limitations for expert domains. First, it treats all factual claims as equally important, overlooking task-specific hierarchies critical in high-stakes fields. In legal reasoning, different information types carry distinct importance levels: legal analysis involves navigating complex hierarchies between statutes, regulations, and case law, while requiring careful selection of the most relevant precedents from potentially hundreds of applicable cases. For instance, finding statutes that directly govern a legal issue is more critical than citing tangentially related case law, and among relevant

precedents, those most favorable to a client's position hold greater strategic value. RAGCHECKER's flat claim structure cannot capture these domain-specific priorities.

Second, RAGCHECKER relies on rich, free-form human-authored gold answers as reference standards. While appropriate for open-domain QA, such answers are time-consuming to create and difficult to standardize in high-stakes settings, limiting scalability.

A.6 DeCE: Decomposed Criteria Evaluation

Our DeCE framework addresses these fundamental limitations through several key innovations:

Automatic Domain-Specific Criteria Generation: Unlike CheckEval's manual dimension definition (Lee et al., 2024) or LLM-RUBRIC's manual rubrics (Hashemi et al., 2024), DeCE automatically extracts evaluation criteria from expert-authored gold answer requirements, which are usually available in expert domains. This approach captures domain expertise while requiring modification in only 11.95% of cases, enabling scalable deployment across different legal domains without extensive manual taxonomy construction.

Instance-Level Adaptive Evaluation: DeCE generates question-specific evaluation criteria rather than applying fixed criteria across all instances. This addresses a critical gap in existing automated approaches and achieves the instance-level adaptation that only human expert review previously provided, but with systematic scalability.

Systematic Precision-Recall Decomposition: DeCE leverages structured gold answer requirements that specify core informational goals (e.g., "cite controlling statute", "identify most relevant precedent"), inherently capturing domain-specific hierarchies by distinguishing between different information types. This structured approach evaluates whether models address the complex hierarchies inherent in legal reasoning—such as identifying governing statutes versus selecting strategically relevant precedents from hundreds of potential cases—with each requirement weighted according to its role in the overall legal analysis. DeCE thus shifts evaluation from claim overlap toward goal fulfillment and task grounding.

Interpretable Performance Analysis: Unlike holistic LLM judges that provide opaque scores (Zheng et al., 2023), DeCE offers detailed criterion-level feedback that enables identification of systematic challenges across jurisdictions and query types, providing concrete guidance for legal AI improvement.

B Dataset Details

B.1 Jurisdictional Distribution

Our dataset of 224 legal question-answer pairs spans diverse U.S. jurisdictions with representation proportional to the population of legal practitioners in each region. Consequently, jurisdictions with larger legal communities—such as New York, California, and Texas—appear with greater frequency than less populous states like Wyoming, Montana, or Nebraska. This distribution reflects the natural concentration of legal activity and ensures that evaluated models encounter jurisdictional diversity necessary for broad applicability while maintaining relevance to real-world usage patterns.

B.2 Data Example

Data Example

Query:
What constitutes good cause or excusable neglect to be given more time to timely serve complaint in Florida?
Search Results:

```
<document>
[...]
</document>
[other related documents truncated]
```

Gold Answer:
Required Information: Circumstances that may constitute good cause or excusable neglect for failure to timely serve are expansive, and it is [other Required Information truncated]
Helpful Information: There is ample case law interpreting what constitutes "good cause or excusable neglect" under Rule 1.070(j). Some examples include:

[other related caselaws truncated]

B.3 Taxonomic Categories

The legal queries have been systematically categorized by Attorney Editors according to a comprehensive taxonomy capturing the breadth of legal information needs. Our classification framework encompasses various distinct categories organized hierarchically based on query complexity and subject matter. This taxonomic structure enables fine-grained analysis of model performance across different question types and complexity levels.

C Evaluation Framework Details

C.1 Detailed Pointwise Evaluation Rubric

Our pointwise evaluation employs the following detailed criteria:

- **Irrelevant (0):** Completely incorrect or unrelated to the input and request. The response demonstrates no understanding of the legal question or provides information that is factually wrong or legally inapplicable.
- **Poor (1):** Mostly incorrect or largely fails to address the specific request. The response may contain some relevant information but is predominantly inaccurate or misses the core legal issues.
- **Fair (2):** Partially correct with noticeable gaps or minor inaccuracies. The response addresses some aspects of the legal question but omits important elements or contains minor factual errors.
- **Good (3):** Correct and adequately addresses the request, but lacks some nuanced information. The response covers the main legal points accurately but may miss subtle distinctions or comprehensive coverage.
- **Excellent (4):** Fully correct, comprehensive, and directly addresses the specific request. The response demonstrates thorough understanding of the legal issues and provides complete, accurate information.

C.2 Chain of Thought in LLM Judge Prompt

1. *Query Analysis Phase:* First, carefully analyze the query to understand what legal question is being asked.
2. *Ideal Answer Examination Phase:* Next, examine the ideal answer to understand what a comprehensive response should include. Pay special attention to what information is labeled as "Required Information" versus "Helpful Information." Only penalize for missing "Required Information."
3. *Model Answer Review Phase:* Carefully review the model answer to ensure you fully understand its meaning and how it uses citations. Note any ambiguous statements that might affect your evaluation.
4. *Comparative Evaluation Phase:* Next, review the provided reference materials (search results) and compare the model answer against the ideal answer. Evaluate:
   (a) Whether all key legal principles from the Required Information in the ideal answer are correctly identified
   (b) If citations are properly used to support claims (citations in square brackets [#] should correspond to relevant paragraph IDs in the search results)
   (c) Whether the information provided is accurate and relevant to the question
   (d) Any gaps or errors in legal reasoning
   (e) If ALL Required Information from the ideal answer is present (a response can only receive an "Excellent" grade if it includes all Required Information)

5. *Issue Identification Phase:* Based on your analysis, identify specific issues using these labels:

   Incorrect: Contains factually or legally inaccurate statements
   Misattribution: Cited sources do not support the statements they're meant to support
   Missing information: Lacks essential information in "Required Information" of the gold answer
   Citation needed: Contains statements requiring citation but none is provided
   Irrelevant: Includes information unresponsive to the legal question
   Wrong jurisdiction: Based on laws from a different jurisdiction than specified
   Repetitive: Unnecessarily repeats the same legal points multiple times

6. *Final Grading Phase:* Determine the final grade by weighing the strengths and weaknesses identified.

C.3 Pointwise Evaluation Prompt

The prompt template below is domain-agnostic and can be adapted across various specialized fields, including legal, medical, and financial domains.

Pointwise Evaluation Prompt Template

You are a [legal/medical/financial] expert evaluating AI-generated responses to specialized queries. Your task is to assess response quality against gold standard answers using established evaluation criteria.

Evaluation Framework:
[The domain-specific rubric. A legal domain rubric is specified in Appendix C.1]

Evaluation Process:
Follow the structured chain-of-thought methodology below:
[A domain-specific chain-of-thought process; a legal example is provided in Appendix C.2]

Response Format:
Provide your complete reasoning process followed by structured output in JSON format:

```
{
  "reasoning": "Detailed chain-of-thought analysis...",
  "score": [numerical_grade],
  "justification": "Concise explanation for assigned score"
}
```

Reference Examples:
Consult the graded response examples below to calibrate your evaluation standards across different quality levels:
[Insert example answers of varying quality for the same query]

Evaluation Input:
Query: {query}
Search Results: {search_results}
Gold Standard Answer: {gold_answer}
Model Response: {model_response}

C.4 DeCE Prompt Templates

DeCE employs four distinct prompt templates corresponding to each evaluation step. All templates are presented below:

C.4.1 DeCE Criterion Extraction Prompt

The prompt template below is domain-agnostic and can be adapted across various specialized fields to extract evaluation criteria from gold standard answers.

Gold Criterion Extraction Prompt Template

You are a [legal/medical/financial] analysis expert tasked with converting comprehensive domain-specific answers into structured assessment criteria.
Input:

1. A specialized domain query

2. A gold standard answer containing structured information sections (e.g., "Required Information" and "Helpful Information")

Task Specification:
Create a comprehensive checklist of evaluation criteria by:

1. **Extracting mandatory elements** from the Required Information section, including:

   [Required elements vary by domain; for the legal domain, the elements are as follows:]

   - Key domain principles, requirements, and concepts
   - Specific conditions, exceptions, or qualifications
   - Procedural steps or chronological requirements
   - Evidentiary standards or verification requirements
   - Domain-specific regulatory or jurisdictional elements
   - Authoritative sources (cases, regulations, guidelines) and their significance
   - Analytical frameworks or interpretive approaches

2. **Incorporating referenced examples** when the primary section explicitly references supplementary information:

   - Create criteria that verify inclusion of appropriate illustrative examples
   - Specify that mentioning any relevant example is sufficient unless otherwise indicated

Response Format:
Return your response in JSON format as follows:

```
{
  "gold_criterion": [
    "1: [Description of first required element]",
    "2: [Description of second required element]",
    ...
  ]
}
```

Reference Examples:
[Insert domain-specific examples demonstrating criterion extraction]

Evaluation Input:
Query: {query}
Gold Standard Answer: {answer}

C.4.2 Model Answer Evaluation Against Gold Criteria

The prompt template below evaluates AI-generated responses against gold criteria across various specialized domains.

Criterion Evaluation Prompt Template

You are a [legal/medical/financial] expert evaluating an AI-generated response to a specialized domain question. You are provided with ideal answer criteria and an AI-generated response to assess.
Input Data:

```
[BEGIN DATA]
***
[Ideal answer criteria]: {gold_criteria}
***
[AI-generated response]: {model_response}
***
[END DATA]
```

Evaluation Task:
Grade the AI response against each numbered criterion using binary scoring:

- **Score 1:** Criterion is satisfied
- **Score 0:** Criterion is not satisfied

Evaluation Guidelines:
For each criterion, assess satisfaction based on the following rules:
[Evaluation rules vary by domain; the evaluation rules for the legal domain are as follows:]

1. **Content matching:** The response addresses the criterion's requirements, regardless of exact wording
2. **Implicit coverage:** The response implicitly captures the essential elements of the criterion
3. **Authoritative sources:** When the response cites relevant authorities (cases, regulations, guidelines) with appropriate context, score as satisfied
4. **Logical equivalence:** Responses stated in negative versus positive form (or vice versa) that encompass criterion elements are acceptable
5. **Conservative scoring:** When in doubt, assign a score of 0 (not satisfied)

For each criterion, identify and quote specific statements from the response that support your scoring decision and provide clear reasoning.
Response Format:
Provide evaluation in JSON format only. Ensure proper escaping of quotation marks within string fields:

```
{
    "scores": [
        [score of 0 or 1 for first criterion],
        [score of 0 or 1 for second criterion],
        ...
    ],
    "reasoning": [
        "[explanation for first criterion scoring]",
        "[explanation for second criterion scoring]",
        ...
    ]
}
```

Example Output:

```
{
    "scores": [0, 1, 0],
    "reasoning": [
        "The response does not mention the required principle.",
        "The response clearly states that ...",
        "The response lacks specific details about ..."
    ]
}
```

C.4.3 Model Response Element Extraction Prompt

The prompt template below extracts key elements from AI-generated responses across various specialized domains for evaluation purposes.

Element Extraction Prompt Template

You are a [legal/medical/financial] analysis expert tasked with extracting all key elements from a model-generated response for evaluation purposes.
Input:

1. A specialized domain query
2. A model-generated answer to that query

Extraction Task:
Extract and list ONLY the elements that are explicitly present in the model answer. Your job is purely extractive, not evaluative.
Critical Guidelines:

- Only include information that is EXPLICITLY stated in the model answer
- Do NOT mention what is missing from the answer
- Do NOT use phrases like "no specific sources were cited" or "no conditions were provided"
- Do NOT evaluate the quality, completeness, or accuracy of the answer
- If an element category has nothing to extract, simply omit it from your response

Extraction Categories:
Extract only what IS present in the text, organized by these categories when applicable:
[Categories vary by domain; for the legal domain, the categories are as follows:]

1. Key domain principles, requirements, and concepts
2. Specific conditions, exceptions, or qualifications
3. Procedural steps or chronological requirements
4. Evidentiary standards or verification requirements
5. Domain-specific regulatory or jurisdictional elements
6. Authoritative sources cited (cases, regulations, guidelines) – For each source, include:
    - The exact name/citation as mentioned in the text
    - The specific context in which it was cited
    - The claimed proposition or principle the source supposedly supports
    - Any direct quotes attributed to the source
7. Interpretive frameworks or analytical approaches

Response Format:
Return your response in JSON format as follows:

```
{
  "model_elements": [
    "1: [Description of first element extracted from model answer]",
    "2: [Description of second element extracted from model answer]",
    ...
  ]
}
```

Evaluation Input:
Query: {query}
Model Answer: {answer}

C.4.4 Element Verification Against Gold Standard Prompt

The prompt template below verifies whether elements from AI-generated responses are supported by gold standard answers across various specialized domains.

Element Verification Prompt Template

You are a [legal/medical/financial] expert tasked with verifying if elements from an AI-generated response are supported by a gold standard answer.
Input Data:

```
[BEGIN DATA]
***
[Gold standard answer]: {gold_answer}
***
[Elements from an AI-generated response]: {elements}
***
[END DATA]
```

Verification Task:
Grade the elements from the AI-generated response based on the gold standard answer using binary scoring:

- **Score 1:** Element is supported by the gold standard answer
- **Score 0:** Element is not supported by the gold standard answer

Evaluation Guidelines:
For each numbered element, assess support based on the following rules:
[Evaluation rules vary by domain; for the legal domain, the rules are as follows:]

1. **Content alignment:** The gold standard answer addresses the element's content, regardless of exact wording
2. **Implicit support:** The gold standard answer implicitly captures the essential aspects of the element

3. **Authoritative validation:** When the gold standard answer cites relevant authorities (cases, regulations, guidelines) that support the element, score as supported

4. **Logical equivalence:** Gold standard answers stated in negative versus positive form (or vice versa) that encompass the element are acceptable

5. **Conservative scoring:** When in doubt, assign a score of 0 (not supported)

For each element, identify and quote specific statements from the gold standard answer that support your scoring decision and provide clear reasoning.
Response Format:
Provide evaluation in JSON format only. Ensure proper escaping of quotation marks within string fields:

```
{
    "scores": [
        [score of 0 or 1 for first element],
        [score of 0 or 1 for second element],
        ...
    ],
    "reasoning": [
        "[explanation for first element scoring]",
        "[explanation for second element scoring]",
        ...
    ]
}
```

Example Output:

```
{
    "scores": [0, 1, 0],
    "reasoning": [
        "The gold standard answer does not mention this principle.",
        "The gold standard answer clearly states that ...",
        "The gold standard answer lacks support for this element..."
    ]
}
```

C.5 Hyperparameter Settings

All evaluations employ consistent inference parameters to ensure reproducibility and fair comparison across models. We configure Claude 3.5 Sonnet with task-specific hyperparameters optimized for each evaluation component.

C.5.1 Pointwise Evaluation

For holistic quality assessment, we configure the LLM judge to generate detailed chain-of-thought reasoning:

- **Temperature:** 0.0 (deterministic output for evaluation consistency)
- **Max tokens:** Maximum input token count + 2,000 (accommodating CoT reasoning, issue identification, and final scoring)
- **Top-p:** 1.0

C.5.2 DeCE Evaluation Components

Generation Steps (Steps 1 & 3): For the extraction of the gold criteria and the answer elements of the model that require various outputs:

- **Temperature:** 0.3 (introducing controlled randomness for comprehensive element identification)
- **Max tokens:** Maximum input token count + 1,000
- **Top-p:** 1.0

Verification Steps (Steps 2 & 4): For criteria evaluation and model answer element verification requiring deterministic judgments:

- **Temperature:** 0.0 (deterministic output for evaluation consistency)
- **Max tokens:** Maximum input token count + 1,000 (accommodating brief explanations for verification decisions)
- **Top-p:** 1.0

The temperature differentiation reflects the distinct requirements of each evaluation phase: deterministic verification ensures consistent scoring, while moderate randomness in generation steps promotes comprehensive coverage of response elements and gold criteria.

C.6 Correlation Analysis Details

Table 3: Correlation coefficients between automated metrics and human expert judgments. DeCE metrics demonstrate significantly higher correlation with human judgments compared with lexical overlap metrics and pointwise LLM-as-a-judge.

Metric Pair	Pearson	Spearman	P-Value	Instance Count
ROUGE-L vs Human Point.	0.34	0.33	< 0.05	244
BLEU vs Human Point.	0.20	0.17	< 0.05	244
ROUGE-L vs Human F2	0.11	0.15	0.29	100
BLEU vs Human F2	0.12	0.13	0.13	100
Point. Judge vs Human Point.	0.59	0.47	< 0.05	244
Point. Judge vs Human F2	0.35	0.37	< 0.05	100
GPTScore Precision vs Human Precision	0.69	0.66	< 0.05	100
GPTScore Recall vs Human Recall	0.56	0.48	< 0.05	100
GPTScore F2 vs Human F2	0.48	0.39	< 0.05	100
G-Eval Precision vs Human Precision	0.66	0.66	< 0.05	100
G-Eval Recall vs Human Recall	0.53	0.46	< 0.05	100
G-Eval F2 vs Human F2	0.48	0.39	< 0.05	100
RAGCHECKER Precision vs Human Precision	0.62	0.63	< 0.05	100
RAGCHECKER Recall vs Human Recall	0.39	0.34	< 0.05	100
RAGCHECKER F2 vs Human F2	0.38	0.31	< 0.05	100
DeCE F2 vs Human Point.	0.46	0.40	< 0.05	244
DeCE Precision vs Human Precision	0.69	0.67	< 0.05	100
DeCE Recall vs Human Recall	**0.80**	**0.82**	< 0.05	100
DeCE F2 vs Human F2	**0.78**	**0.76**	< 0.05	100

D Detailed Criteria Evaluation Analysis

D.1 Complete Criteria Validation Results

Table 4 provides comprehensive statistics on LLM-generated criteria validation by legal experts.

Validation Outcome	Count	Percentage
No modification required	855 criteria	87.33%
Criteria modified	117 criteria	11.95%
Criteria rejected	7 criteria	0.72%
New criteria added	20 criteria	

Table 4: Expert validation results for LLM-generated evaluation criteria. Among all 979 LLM-generated criteria, only 11.95% needs modification and an extra 20 criteria need to be added.

D.2 Detailed Refinement Patterns

D.2.1 Specificity Calibration Examples

Original: "Does the response identify that Missouri appellate courts apply the substantial evidence standard?" **Refined:** " Does the response identify that Missouri appellate courts apply the substantial evidence standard *when upholding findings of a lower court*?"

D.2.2 Legal Authority Differentiation Examples

Original: " Does the response identify relevant statutory requirements and supporting case law?" **Separated into:** - " Does the response cite relevant statutory or regulatory authority?" - " Does the response include supporting case examples?"

D.2.3 Case Law Flexibility Examples

Original: " Does the response cite Sullivan v. Town of Acton or Canteen Corp. v. City of Pittsfield?" **Refined:** " Does the response cite Sullivan v. Town of Acton, Canteen Corp. v. City of Pittsfield, *or another relevant case establishing municipal liability principles*?"

E Responsible NLP Checklist

Checklist ID	Question	Answer
A1	Did you describe the limitations of your work?	Yes. The limitations are provided in Section 5.

Continued on next page

Checklist ID	Question	Answer
A2	Did you discuss any potential risks of your work?	Yes. Due to the page limit of the main paper, the risks are provided below. **Over-reliance on automated evaluation:** If practitioners become overly dependent on DeCE scores without human oversight, this could lead to deployment of systems that appear to perform well on metrics but fail in real-world scenarios with serious consequences in high-stakes domains like law and medicine. To mitigate this risk, regular validation of evaluation performance by diverse expert panels should be adopted. **Potential for gaming:** Once evaluation criteria become known, there's risk that future LLMs could be optimized specifically to perform well on DeCE metrics without genuine improvement in legal reasoning capabilities. To mitigate this risk, we recommend: continuously updating evaluation criteria, using multiple assessment dimensions beyond precision/recall, and implementing adversarial validation with red team testing to prevent models from overfitting to specific patterns. Additionally, maintaining some evaluation criteria as private, implementing multi-stage validation processes, and establishing continuous monitoring systems with expert oversight can help detect suspicious performance patterns and ensure real-world correlation with DeCE scores.
B1	Did you cite the creators of artifacts you used?	Yes. We cite the creators of artifacts we used throughout the paper, including the models we evaluate, the baseline metrics we compare with, and the judge model: Claude 3.5 Sonnet used as the backbone LLM for DeCE evaluation. Regarding the dataset, while our legal QA dataset is proprietary and cannot be released, we acknowledge this limitation in the paper
B2	Did you discuss the license or terms for use and or distribution of any artifacts?	Yes, terms of use of models are dicussed in LLMs evaluated in Section 3.

Continued on next page

Checklist ID	Question	Answer
B3	Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?	Yes, the discussion of our use of existing models are discussed in LLMs evaluated in Section 3.
B4	Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?	Yes. We mentioned in the paper the data is proprietary and won't be pulicly released. Due to the page limit of the main paper, the discussions are provided below: Our legal QA dataset consists of professionally curated attorney-authored questions and expert gold standard answers that are designed for internal evaluation usage. The dataset contains some personally identifiable information (PII) such as judge names, attorney names, plaintiff and defendant names, which are needed for LLM legal use cases. We do not plan to release the data due to these privacy considerations.
B5	Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?	Yes. While our legal QA dataset is proprietary and cannot be publicly released due to confidentiality constraints, we provide dataset documetation in Appendix B.
B6	Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created?	Yes. Data description is provided in Dataset in Section 3 and details are provided in Appendix B.
C1	Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?	Yes, model parameter size is discussed in LLMs evaluated in Section 3. We access all third-party models through APIs, as mentioned in the footnote of the same section. The paper aims to propose a novel LLM-based, domain-agnostic evaluation framework and to share empirical findings on the limitations of state-of-the-art models for legal QA use cases. Proprietary model training strategies are not within the scope of this paper.

Continued on next page

Checklist ID	Question	Answer
C2	Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?	Yes. Hyperparameter setting for LLM judges are provdided in Appendix C.5. The paper aims to propose a novel LLM-based, domain-agnostic evaluation framework and to share empirical findings on the limitations of state-of-the-art models for legal QA use cases. Proprietary model training strategies are not within the scope of this paper.
C3	Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?	Yes. We report descriptive statistics for our results with full transparency about our experimental setup: **Statistical Reporting:** Correlation coefficients: We report both Pearson and Spearman correlation coefficients between automated metrics and human expert judgments (Table 2). Performance distributions: We present detailed score distributions across models using box plots (Figure 3) and percentage breakdowns (Figure 2). Aggregate statistics: We report mean scores and performance ranges across different jurisdictions (Figure 4) and query types (Figure 5). **Experimental Design:** All reported results are based on single run experiments considering we use temperature=0.0 for all verification steps in DeCE to ensure consistent scoring.
C4	If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, Spacy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?	Yes. Due to the page limit of the main paper, the package versions are provided below: nltk: 3.9.1 rouge: 1.0.1 transformers: 4.52.4

Continued on next page

Checklist ID	Question	Answer
D1	Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?	Partially. We provide a detailed methodology and evaluation criteria, but cannot release the complete annotation instructions due to proprietary constraints. **What we do provide:** Detailed evaluation rubrics: Complete pointwise evaluation criteria (0–4 scale) with specific definitions for each score level (Appendix C.1). Chain-of-thought methodology: A full six-phase evaluation process, including Query Analysis, Ideal Answer Examination, Model Answer Review, Comparative Evaluation, Issue Identification, and Final Grading (Appendix C.2). Comprehensive prompt templates: All four DeCE prompt templates for criterion extraction, model evaluation, element extraction, and verification (Appendix C.4). **What we cannot provide:** Complete annotation instructions, which contain proprietary legal evaluation frameworks developed for internal use. Our provided methodology, rubrics, and prompt templates contain sufficient detail to reproduce the evaluation approach and serve as a guideline for other high-stakes domains, as one purpose of the paper is to propose our novel, domain-agnostic evaluation framework. Domain experts can adapt our framework using the comprehensive templates and evaluation criteria we provide. **Justification:** This approach balances transparency with proprietary constraints while providing sufficient methodological detail for reproducibility in similar expert domains.
D2	Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?	Yes. Due to the page limit of the main paper, the discusssion is provided below: The legal experts who participated in our evaluation are employees of our institution rather than external recruited participants. As institutional employees, their participation in this research evaluation was part of their professional duties rather than separate recruitment with additional compensation. All experts are based in the United States, appropriate for our U.S. legal QA dataset spanning diverse jurisdictions.

Continued on next page

Checklist ID	Question	Answer
D3	Did you discuss whether and how consent was obtained from people whose data you're using/curating?	Yes. Due to the page limit of the main paper, the discussion is provide below: The human evaluation annotations were conducted by legal experts who are employees of our institution. As institutional employees, their participation in annotation tasks was part of their professional duties rather than external data collection requiring separate consent procedures.
D4	Was the data collection protocol approved (or determined exempt) by an ethics review board?	Yes, our legal QA dataset use was reviewed and approved by our institution's internal legal department. The legal professionals who participated in annotation and evaluation tasks are employees of our institution, operating under established professional responsibility frameworks. All data handling and research activities were conducted in accordance with our institution's internal policies for proprietary data use.
D5	Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?	Yes, as discussed in Section 3.1.
E1	If you used any AI assistants, did you include information about your use?	Yes. Due to the page limit of the main paper, the discussion is provide below: We acknowledge the use of large language models for manuscript refinement and code development assistance during the preparation of this work. All conceptual contributions, experimental design, analysis, and conclusions remain the authors' original work.