Are LLMs Court-Ready? Evaluating Frontier Models on Indian Legal Reasoning

Kush Juvekar kush@adalat.ai Adalat AI, India

Arghya Bhattacharya arghya@adalat.ai Adalat AI, India

Sai Khadloya sai@adalat.ai Adalat AI, India

Utkarsh Saxena utkarsh@adalat.ai Adalat AI, India

Abstract

Large Language Models are entering legal workflows, yet we lack a jurisdiction-specific framework to assess their baseline competence therein. We use India's public legal examinations as a transparent proxy. Our multi-year benchmark assembles objective screens from top national and state exams and evaluates open and frontier LLMs under *real world exam conditions*. To probe beyond MCQs, we also include a lawyer-graded, paired-blinded study of long-form answers from the Supreme Court's Advocate-on-Record exam. This is, to our knowledge, the first exam-grounded, India-specific yardstick for LLM court-readiness released with datasets and protocols. Our work shows that while frontier systems consistently clear historical cutoffs and often match or exceed recent top-scorer bands on objective exams, none surpasses the human topper on long-form reasoning. Grader notes converge on three reliability failure modes—procedural/format compliance, authority/citation discipline, and forum-appropriate voice/structure. These findings delineate where LLMs can assist (checks, cross-statute consistency, statute and precedent lookups) and where human leadership remains essential: forum-specific drafting and filing, procedural and relief strategy, reconciling authorities and exceptions, and ethical, accountable judgment.

1 Introduction

LLMs have cleared multiple-choice bar-style screens in several jurisdictions, renewing interest in legal automation, but a jurisdiction-first question remains: *are these systems court ready*? Other fields probe such capability with exam-style settings: broad knowledge suites such as MMLU, Olympiad-level problems in mathematics and science, and clinically oriented reasoning in health (Hendrycks et al., 2020; He et al., 2024; Singhal et al., 2023). By contrast, many AI-and-law studies focus on short-context recall (for example, bail or recidivism prediction and legal-judgment prediction). These are metric-friendly but only indirectly tied to how courts expect lawyers to write and file (Kleinberg et al., 2018; Dressel and Farid, 2018; Cui et al., 2022). India offers a jurisdiction where court-legible benchmarks already exist. We adopt *public exams* already used to gate human entry **Common Law Admission Test (CLAT) –UG/PG** (admissions), **Delhi Judicial Services/ Delhi Higher Judicial Services (DJS/DHJS)** prelims (judiciary), and the Supreme Court's **Advocate-on-Record (AoR)** exam (rights of audience) as court-ready yardsticks (Consortium of National Law Universities, 2025b; Consortium of NLUs, 2025; High Court of Delhi, 2023, 2024; Supreme Court of India, 2025).

Our primary contributions in this paper are:

- **Exam-grounded dataset (objective + subjective):** We curate a multi-year corpus of *objective* questions **6,218 MCQs** plus *subjective* AoR materials (2023). Provenance, year coverage, and marking rules are documented in the Appendix.

  We release the dataset here. 🤗
- **Benchmark under official rules:** We evaluate open and closed models including frontier and strong open baselines under exam-native interfaces and identical scoring policies, enabling comparisons across model families and scales.
- **Blinded AoR study with certified graders.** For each AoR paper, we create paired sets comparing the human-written version with the model-generated version. We anonymize them and have certified AoRs evaluate them using a rubric.

By anchoring evaluation in public exams that every law student, judge, and policymaker recognizes, we present results that legal practitioners can interpret and ML researchers can reproduce. We hope this shared yardstick helps both communities see where

Proceedings of the Natural Legal Language Processing Workshop 2025, pages 359–369
November 8, 2025 ©2025 Association for Computational Linguistics

current models stand today and guides evidence-based adoption and research.

2 Related Work

Legal NLP suites such as LexGLUE and LegalBench cover broad tasks, and IL–TUR targets Indian legal texts, but none align with the public exams that govern entry and practice in India (Chalkidis et al., 2022; Guha et al., 2023; Joshi et al., 2024). Outside law, exam-style evaluations (e.g., MMLU and Olympiad-level benchmarks) stress reasoning but do not test forum-specific procedure or authority discipline (Hendrycks et al., 2020; He et al., 2024; Sun et al., 2025). In AI and law, much of the literature concentrates on economically salient prediction tasks (for example, bail, recidivism, and legal-judgment prediction), where metrics are tractable but only weakly aligned with the linguistic and rhetorical demands of courtroom writing (Kleinberg et al., 2018; Dressel and Farid, 2018; Cui et al., 2022; Shui et al., 2023). Bar-exam studies are informative but jurisdictionally distinct from India's public exams (Katz et al., 2023). To our knowledge, this is the first study to evaluate frontier LLMs on India's public legal examinations, pairing multi-year objective screens with a lawyer-graded subjective study under exam constraints.

3 Exams and Scope

Why these exams? India's legal profession is structured around a publicly administered system of entry and advancement that is mediated through a series of high-stakes examinations. These include: (i) the CLAT–UG/PG for entry into undergraduate and postgraduate law programs, (ii) judicial service examinations such as the DJS and DHJS for recruitment to the judiciary, and (iii) the Advocate-on-Record (AoR) examination, which confers exclusive rights of audience before the Supreme Court of India. Reliance on these standardized assessments provides institutional legibility for stakeholders, who use them to regulate access, allocate professional privileges, and validate competence across the legal system for humans. (Consortium of National Law Universities, 2025b; Consortium of NLUs, 2025; High Court of Delhi, 2023, 2024; Supreme Court of India, 2025).

Why AoR for the subjective study? The Supreme Court's *Advocate-on-Record (AoR)* certification is uniquely consequential: under the *Supreme Court Rules, 2013*, only an AoR may file an appearance or act for a party in the Supreme Court (the AoR is the filing advocate of record) (Supreme Court of India, 2013). Eligibility itself signals seniority and training (four years' practice plus one year of training under an AoR, followed by a Court-conducted examination) (Supreme Court of India, 2025, 2024). The exam is administered by the Supreme Court and consists of four descriptive papers *Practice & Procedure*, *Drafting*, *Advocacy/Professional Ethics*, and *Leading Cases* making it the most premium, publicly administered subjective legal reasoning assessment tied directly to Supreme Court practice (Supreme Court of India, 2025; Careers360 Law, 2023). These features make AoR the clearest lens for open-ended legal reasoning and writing.

On excluding AoR *Drafting*. We exclude Paper II *(Drafting)* from quantitative scoring because drafting in the AoR exam is format critical: cause-title, parties, prayers, affidavits, layout, and citations must follow strict Supreme Court templates. Without document-template tooling, LLMs generate legally plausible text that routinely violates these formal requirements. In a pilot using Gemini 2.5 Pro, the certified examiner deemed the draft "not evaluable" due to pervasive structural non-compliance (see Appendix §E). As Gemini 2.5 Pro is the only model with consistent cross-exam performance across years, this single pilot suffices to indicate drafting limitations. Accordingly, we evaluate AoR on the three papers testing legal reasoning: *Practice & Procedure*, *Advocacy & Professional Ethics*, and *Leading Cases*.

Year selection and provenance. For *objective* exams, we include only years where the governing body released both the official paper and the *official answer key*; years without a key are excluded. For *subjective*, we evaluate AoR 2023 with paired, blinded grading by certified practitioners (three papers as above). All artifacts are sourced from official portals.[1] A compact dataset summary is in Table 1; full syllabi and exact year coverage appear in Tables 8 and 9.

4 Model Selection and Inference Configuration

Objective–exam cohort. We evaluate a broad panel spanning frontier proprietary, large open, and

[1]Representative sources: Consortium of National Law Universities (2025b); Consortium of NLUs (2025); High Court of Delhi (2023, 2024); Supreme Court of India (2025).

Exam	Mod.	Qs/exam	Marking	Total Questions
CLAT UG	Obj.	200/150/120	$+1/-0.25/0$	**3,154**
CLAT PG	Obj.	120	$+1/-0.25/0$	**814**
DJS (Prelim.)	Obj.	200	$+1/-0.25$	**1,400**
DHJS (Prelim.)	Obj.	150	$+1/-0.25$	**850**
AoR 2023 (SC)	Subj.	35	none (per-question marks)	**35**

Table 1: Objective exam sizes and marking (official).

strong small/open baselines. The cohort (i) covers **families** across vendors (Google, OpenAI, Anthropic, Mistral, DeepSeek, Alibaba/Qwen, Meta) to reduce recipe bias; (ii) spans **scales** from ~7B to frontier sizes to observe size trends under identical constraints; (iii) includes both instruction-tuned and **reasoning-tuned** models (e.g., R1) to test whether explicit reasoning helps under exam conditions; and (iv) uses widely available endpoints, increasing reproducibility.

Subjective–exam cohort (AoR). Human grading constraints preclude evaluating all models on long-form papers. We therefore select a principled triad: *Gemini 2.5 Pro* (the objective leader, testing transfer from capacity to drafting), *Gemma 3 27B* (the strongest small/open baseline in our objective runs, offering cost-efficient human evaluation), and *DeepSeek R1* (a large, reasoning-tuned open model, probing whether reasoning training enhances forum-specific drafting). Together, these models represent a **frontier ceiling**, a **competitive small/open baseline**, and a **reasoning-tuned large open** reference point, sufficient to expose transfer gaps without exhausting grader bandwidth.

Inference setup. We replicate *official exam conditions* by incorporating constraints (such as negative marking for MCQs and forum-specific AoR instructions) into our prompts, positioning each model as an exam candidate. Full prompt text is in Appendix §B and §D. Decoding is deterministic (temperature = 0; default top-p); no tools, no retrieval, single-pass inference. All endpoints are invoked via a single gateway *OpenRouter* (OpenRouter, 2025) with date-pinned model identifiers; total evaluation cost was $<$ \$500. For objective papers we enforce a structured output schema and score strictly under official rules; per-model breakdowns appear in the Appendix§G.

5 Evaluation

5.1 Objective exams

We score per question using the official schema. Models return strict JSON with a single `answer_label` using structured outputs; non-conforming outputs are marked wrong in the conservative variant.

5.2 Subjective exams

Paired, blinded design. For AoR 2023 we obtain human answer script (***AoR 2023 Exam Topper***) and generate one LLM script per model in our subjective cohort (Gemini 2.5 Pro; Gemma 3 27B; DeepSeek R1). For each paper, we create three *paired sets*, each containing two anonymized scripts of the same question(s): *Script A* (human or LLM) and *Script B* (the other), with order randomized. **Rubric and aggregation.** Each certified *Advocate-on-Record* (AoR) grader receives all three paired sets and is not told which script is human/AI. Five certified AoR assessors grade all pairs using an official-style rubric covering: (i) Accuracy and application of law to facts, (ii) Authority discipline (presence, correctness, and fit of case/statute citations; penalties for fabrication/miscitation; rewards for pin-point cites), (iii) Forum-specific structure and procedure (orders/rules, cause-titles/party arrays/prayers where relevant), (iv) Depth/nuance and handling of counter-arguments, (v) Language and expression (clarity, concision, tone). We compute per-paper totals, deltas (LLM vs. human), and summarize qualitative failures.

6 Results

6.1 Objective exams

We evaluate the models against human topper scores released by the official exam committees(Consortium of National Law Universities, 2025a; High Court of Delhi, 2025). Across all four objective exams, frontier systems lead consistently. Gemini 2.5 Pro exceeds the historical topper anchors on every exam (e.g., CLAT PG +14.3, DJS +19.8, DHJS +25.3 on average across years), while GPT 5 Chat is near parity on DJS and strongly positive on CLAT PG. Open

Model	Paper	Grader 1 (AoR)		Grader 2 (AoR)		Grader 3 (AoR)		Grader 4 (AoR)		Grader 5 (AoR)	
		AI	Human	AI	Human	AI	Human	AI	Human	AI	Human
Gemini 2.5 Pro	Practice & Proc.	63.5	73.0	68.5	69	68	98	72	85	70	83
	Adv. & Ethics	79.0	78.5	72	70.5	68	70	68	74	73	69
	Leading Cases	73.0	73.0	63	75	58	76	58	73	60	75
	Non-draft. total	**215.5**	**224.5**	**203.5**	**214.5**	**194**	**244**	**198**	**232**	**203**	**227**
DeepSeek R1	Practice & Proc.	67.0	73.0	59.5	69	58	98	50	85	63	85
	Adv. & Ethics	68.0	78.5	64.5	70.5	63	70	60	74	66	69
	Leading Cases	59.0	73.0	45.5	75	42	76	38	73	43	75
	Non-draft. total	**194.0**	**224.5**	**169.5**	**214.5**	**163**	**244**	**148**	**232**	**172**	**227**
Gemma 3 27B	Practice & Proc.	32.0	73.0	57.5	69.0	37	98	41	85	39	85
	Adv. & Ethics	58.0	78.5	41.5	70.5	66	70	66	74	64	69
	Leading Cases	57.0	73.0	55.0	75.0	47	76	47	73	50	75
	Non-draft. total	**147.0**	**224.5**	**154.0**	**214.5**	**150**	**244**	**148**	**232**	**153**	**227**

-

Table 2: AoR 2023 (paired, blinded): combined per-paper scores by evaluator. Drafting excluded from aggregates; pilot note in Appendix §E.

Model	**CLAT UG** (/200/150/120)		**CLAT PG** (/120)		**DJS** (/200)		**DHJS** (/150/125)	
	Avg	Δ	**Avg**	Δ	**Avg**	Δ	**Avg**	Δ
Topper average (anchor)	148.9		88.6		162.7		114.1	
Gemini 2.5 Pro	156.6	$+7.6$	102.8	$+14.3$	182.5	$+19.8$	139.4	$+25.3$
GPT 5 Chat	134.7	-14.3	100.8	$+12.2$	162.7	$+0.0$	119.4	$+5.3$
DeepSeek R1	141.6	-7.3	91.8	$+3.2$	146.7	-16.0	113.3	-0.9
DeepSeek Chat v3	141.5	-7.4	91.0	$+2.5$	137.3	-25.4	114.1	$+0.0$
Claude Sonnet 4	143.5	-5.4	88.5	-0.1	150.7	-12.0	112.8	-1.3
Mistral Medium 3.1	136.7	-12.2	89.9	$+1.3$	148.1	-14.5	106.0	-8.1
Qwen 3 235B	138.1	-10.9	82.4	-6.2	127.8	-34.9	83.1	-31.0
Llama 3.3 70B	123.0	-25.9	82.3	-6.3	119.0	-43.7	92.5	-21.6
GPT 4.1 Mini	122.6	-26.3	78.3	-10.3	121.9	-40.8	88.4	-25.7
Gemma 3 27B	117.5	-31.4	69.8	-18.8	111.3	-51.4	80.6	-33.5
Qwen 2.5 7B	101.2	-47.7	59.7	-28.9	76.9	-85.8	56.9	-57.3
Gemma 3 12B	106.0	-42.9	62.3	-26.3	94.0	-68.7	75.3	-38.8
GPT 3.5 Turbo	100.6	-48.3	63.7	-24.9	85.9	-76.8	62.5	-51.6
Llama 3.1 8B	90.5	-58.4	58.2	-30.4	82.1	-80.5	54.5	-59.6

Table 3: Cross-exam summary (averages across years). Δ is against each exam's topper average. Positive Δ indicates model means at or above topper average. Per-year $\times$ model matrices appear in the Appendix §G

reasoning-tuned DeepSeek R1 is competitive on CLAT PG/DHJS but trails on DJS; instruction-tuned DeepSeek v3 reaches parity on DHJS and small positive on CLAT PG, yet lags on DJS/CLAT UG. Smaller ($\leq$30B) models fall below topper anchors across the board (e.g., Gemma 3 27B: -18.8 on CLAT PG; -51.4 on DJS).

6.2 Subjective exams

Gemini 2.5 Pro demonstrates performance closest to human-level proficiency, achieving near parity on the *Ethics* paper and a statistical tie on *Leading Cases*. However, a notable performance gap remains in the *Practice & Procedure* paper, suggesting that procedural knowledge and its application present a distinct challenge. In contrast, other models such as DeepSeek R1 and Gemma 3 27B exhibit more significant performance differentials across all examination papers. A qualitative analysis of grader feedback, which proved remarkably consistent across different evaluators, converged on three principal failure modes:

- **Deficiencies in Authority Discipline and Doctrinal Rigor:** A critical shortfall identified across multiple models was the inability to consistently adhere to the conventions of legal citation and authority. This manifested in several forms: the complete omission of controlling precedents; the miscitation of peripheral authorities; and a more subtle failure termed "manufacturing authorities," where a model correctly recalls a relevant case but fails to articulate its specific relevance to the question's legal dilemma. For example, the failure to cite the controlling precedent in a given area, such as omission of *Rupa Ashok Hurra v. Ashok Hurra AIR 2022* in a discussion on curative petitions is not a simple mistake. Furthermore, responses often exhibited a tendency towards "generic" legal assertions - mentioning concepts like "the principles of natural justice" or citing a well-known case like *Maneka Gandhi v. Union of India AIR 1978* without providing the requisite relevance to the question it is answering. This lack of precision, whether through omission, misapplication, or inadequate synthesis, strips the legal argument of its persuasive force and demonstrates a failure to engage with the source material at the requisite doctrinal depth, treating legal principles as abstract concepts rather than grounded, citable authorities.

- **Proclivity for Irrelevance and Inefficient Content Generation:** A second pervasive issue was the generation of content that substantively drifted from the core legal or factual premises of the question. Graders, colloquially yet pointedly, categorized this as "slop" - digressive text that, while potentially grammatically correct and thematically adjacent, fails to advance a direct answer. This includes lengthy paraphrases of basic legal principles already assumed by the question, speculative explorations of tangential legal scenarios, or the inclusion of boilerplate disclaimers that add no analytical value. This inefficiency not only obscures the relevant answer but also reflects a model's difficulty in performing the crucial task of issue-spotting and prioritization, a skill wherein human examinees are trained to allocate their limited time and space exclusively to the most salient points.

- **Inapt Voice, Structure, and Rhetorical Framing:** The third failure mode pertains to the formal and stylistic conventions of professional legal communication. Model responses were frequently characterized by a distinctly "AI-sounding" cadence, often beginning with overly broad, rephrased introductions that lack the incisive tone expected in high-stakes legal writing. A particularly jarring convention noted by graders was the use of meta-framing, such as prefacing an answer with "As an aspiring Advocate on Record..." or "In my capacity as a legal AI...", despite instructions in prompts not to use such commentary. Such framing breaks the professional illusion and reveals the artificial nature of the author. Furthermore, the structural preferences of the models, which often favor long, generalized paragraphs, clash with the exam specific expectations for concise, point wise answers. The models struggled to adopt the succinct, authoritative, and forum-specific voice that human graders associate with a well-trained legal professional, instead defaulting to a more verbose and generically informative prose style.

In practice, these are filing-critical defects that attract direct mark deductions. They explain the model–human gap in subjective papers.

7 Conclusion

This study demonstrates a clear dichotomy in AI capabilities for legal tasks. On objective, multiple-choice examinations, frontier models meet or even surpass historical human pass marks, demonstrating a robust capacity for short-context legal recall and rule application. However, this proficiency does not translate seamlessly to the domain of subjective, long-form writing, where no model could match the performance of a human topper. The critical deficits lie both in knowledge and in execution: a lack of procedural fidelity, imprecise authority handling, and a failure to adopt forum-specific structure.

These findings compel a two-part definition of what it means to be truly "court-ready." First is the *capacity* to recall and apply legal doctrine at scale, a benchmark the strongest systems now meet. The second, and more elusive, aspect is *reliability* under the practical constraints of legal practice, specifically adhering to procedural defaults, maintaining strict authority discipline, and producing work products that align with judicial expectations. Our results indicate that current systems fall short on this second, crucial dimension for apex court practice in India.

In practical terms, this means AI is best deployed as a supportive tool rather than an autonomous practitioner. Systems can efficiently assist with tasks like searching and verifying authorities, checking consistency across drafts, or cross-referencing case details. Tasks that require full drafting, independent citation, or any action carrying legal responsibility remain beyond current capabilities and should be handled with human oversight.

Limitations

While our work provides stakeholders with greater insights into today's frontier models it has the following limitations:

- **Exam representativeness.** Publicly available exams serve as proxies for live courtroom practice and may not fully capture real-world complexity.

- **Year coverage & answer keys.** Our results depend on the official answer keys and the subset of years with high-quality scans, which may limit generalizability.
- **Model selection.** We evaluate a triad of models for AoR 2023 constrained by bandwidth and resources; other models may yield different outcomes.
- **Endpoint variance.** Closed-model updates and run-to-run variability can shift results; we snapshot models at evaluation time.
- **Blind grading sample size.** Five assessors of record provide robust signals, but cannot exhaust all stylistic or interpretive variance.

References

Careers360 Law. 2023. Aor exam syllabus — overview. Papers: Practice & Procedure, Drafting, Professional Ethics, Leading Cases.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the Annual Meeting of the ACL*.

Consortium of National Law Universities. 2025a. Clat 2026 (ug and pg) overview. Accessed Sep 1, 2025.

Consortium of National Law Universities. 2025b. Ugclat 2026 syllabus. Accessed Sep 1, 2025.

Consortium of NLUs. 2025. Clat–pg: Pattern and syllabus (consortium of nlus).

Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. 2022. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *arXiv preprint arXiv:2204.04859*.

Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv:2308.11462*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv:2009.03300*.

High Court of Delhi. 2023. Delhi judicial service examination: Instructions and negative marking (official notices).

High Court of Delhi. 2024. Delhi higher judicial service examination: Important instructions (objective with negative marking).

High Court of Delhi. 2025. Delhi higher judicial service: Rules/notifications (preliminary, mains, viva). Accessed Sep 1, 2025.

Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. IL-TUR: Benchmark for indian legal text understanding and reasoning. In *Proceedings of the Annual Meeting of the ACL*.

Daniel Martin Katz, Michael J. Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam. SSRN.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

OpenRouter. 2025. Openrouter: The unified interface for llms. `https://openrouter.ai/`. Accessed: 2025-09-01.

Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. *arXiv preprint arXiv:2310.11761*.

K. Singhal and 1 others. 2023. Large language models encode clinical knowledge and reasoning. *Nature*.

H. Sun and 1 others. 2025. An olympiad-level math benchmark for large language models. *arXiv preprint arXiv:2503.21380*.

Supreme Court of India. 2013. Supreme court rules, 2013. Order IV: Advocates-on-Record; only an AoR can file an appearance or act for a party.

Supreme Court of India. 2024. Advocates-on-record: Training and eligibility notices (specimen form/notice). Eligibility certificate for one-year AoR training referencing Supreme Court Rules, 2013.

Supreme Court of India. 2025. Advocate–on–record examination (official information).

A Prompt Templates and Interfaces

Chat template (all models). We use a simple two–message chat format across models; only exam-specific prompts differ.

```
[
  { "role": "system", "content": "<
    system_prompt>" },
  { "role": "user",   "content": "<
    user_message>" }
]
```

Objective exams require strict JSON outputs; subjective exams are long-form. (Full exam-specific prompts below.)

B Objective Exam Prompts

CLAT PG (objective)

```
system_prompt:
{
  "You are an aspiring law student
    taking the Common Law Admission Test
     (CLAT)
   for Post Graduate programs. You will
    output ONLY strict JSON objects with
     your
   answer. Analyze each question
    carefully and choose the best answer
     from the
   given options. MARKING SCHEME: Each
    question carries 1 mark with
    negative
   marking of 0.25 for wrong answers.
    You can choose to SKIP a question if
     you're
   unsure to avoid negative marking. Do
    not include code fences or
    commentary."
}

user_message:
{
  "You are taking the Common Law
    Admission Test (CLAT) for Post
    Graduate programs.
   Read the question and options and
    choose one answer. Always return
    ONLY a JSON
   object with keys: 'answer_label' and
    'explanation'. The 'answer_label'
    MUST be
   one of: 'A','B','C','D' for selecting
     an option, or 'SKIP' to avoid
    negative
   marking if unsure. Keep the
    explanation concise (1-3 sentences).
     Do not include
   any other keys or commentary.\n
   Question: {question}\n
   Options: {options}\n
   Return JSON now."
}
```

CLAT UG (objective)

Same interface as CLAT PG; the allowed labels are `A,B,C,D,SKIP` and the exam name is changed to CLAT UG.

DJS / DHJS Preliminary (objective)

```
system_prompt:
{
  "You are a knowledgeable legal expert
    taking the Delhi Judicial Service
    Examination.
   You will output ONLY strict JSON
    objects with your answer. Do not
    include code
   fences or commentary."
}

user_message:
{
  "You are taking the Delhi Judicial
    Service (DJS) Examination. This exam
     tests
   knowledge of Indian law, judicial
    aptitude, general knowledge and
    current
   affairs. You are answering questions
    from: {paper_name}\n
   Instructions:\n
   - Each correct answer carries 1 mark
    .\n
   - Each incorrect answer carries
    negative 0.25 mark.\n
   - Each skipped answer carries 0 mark
    .\n
   - Choose the most appropriate answer
    based on Indian law and legal
    principles.\n
   - Return ONLY a JSON object with '
    answer_label' (1,2,3,4 or SKIP) and
      'explanation' (brief legal
    reasoning).\n
   - Do not automatically skip GK/
    current-affairs questions.\n
   - Skip if unsure to avoid negative
    marking.\n
   Question: {question}\n
   Options: {options}"
}
```

C Objective JSON schema (enforced at parse time).

```
{
  "answer_label": "A|B|C|D|SKIP"
  // or "1|2|3|4|SKIP"
  "explanation": "1-3 sentences"
}
```

D AoR (Subjective) Prompts

Practice & Procedure (AoR Paper I)

```
"You are taking the Advocate on Record (AOR)
    Examination.
```

```
You are answering questions for the section
    Practice and Procedure of the Supreme
Court of India.

Important Instructions:
- Be verbose but keep the marks for the
    question in mind.
- Write like a candidate would; do NOT reveal
    that you are an LLM.
- Do not include code fences or meta commentary.

- Provide comprehensive answers; include
    relevant case law and statutory provisions.
- Structure your answer logically with clear
    headings.
- Be precise and accurate in legal terminology
    ."
```

Drafting (AoR Paper II)

```
"You are taking the Advocate on Record (AOR)
    Examination.
 You are answering questions for the section
    Drafting.

 Important Instructions:
 - Write like a candidate would; do NOT reveal
    that you are an LLM.
 - The question carries 20 marks.
 - You will be given context and appendices.
 - Draft the required legal document as
    specified in the question.
 - Follow proper legal drafting format and
    structure.
 - Include all necessary components mentioned in
     'INSTRUCTIONS'.
 - Use appropriate legal language and
    terminology."
```

Advocacy & Professional Ethics (AoR Paper III)

```
"You are taking the Advocate on Record (AOR)
    Examination.
 You are answering questions for the section
    Advocacy and Ethics.

 Important Instructions:
 - Be verbose but respect the marks allotted.
 - Write like a candidate would; do NOT reveal
    that you are an LLM.
 - Do not include code fences or meta commentary
    ."
```

Leading Cases (AoR Paper IV)

```
"You are taking the Advocate on Record (AOR)
    Examination.
 You are answering questions for the section
    Leading Cases of India.

 Important Instructions:
 - Be verbose but respect the marks allotted.
 - Write like a candidate would; do NOT reveal
    that you are an LLM.
 - Do not include code fences or meta commentary
    ."
```

E AoR Drafting (Paper II): Exclusion Rationale and Pilot Grader Note

Rationale. AoR Drafting is *format-critical*: cause title, party array, prayer, affidavits, signatures/verification, pagination/lineation, margining, and citation form must match Supreme Court templates (per *Supreme Court Rules, 2013*, Order IV and allied provisions) (Supreme Court of India, 2013). Autoregressive, text-only LLMs without document-template tooling frequently violate these formal requirements even when the narrative is legally plausible. To avoid scoring noise dominated by page-layout compliance, we exclude Drafting from quantitative comparisons and focus on the three reasoning-centric papers (Practice & Procedure; Advocacy/Professional Ethics; Leading Cases).

Pilot grading (one draft). A certified AoR graded a single LLM Drafting response and marked it "*not evaluable*" due to pervasive formal defects. Representative issues (verbatim categories from the grader):

- Missing or malformed *cause title* and party array; prayer block not in prescribed order.
- Incorrect or absent references to relevant *Supreme Court Rules*; wrong order numbers.
- Affidavit/verification, Vakalatnama, and signature blocks omitted or mispositioned.
- Pagination/line numbers and margining absent; citations inconsistently formatted.
- One court-fee statement incorrect for SLP (Crl).

The full note is archived with the anonymized script (available to reviewers on request). The other three papers were graded under the blinded protocol described in the main text.

F AoR Grader Packet & Instructions

Materials provided. (1) AoR 2023 question paper (with official marks per question); (2) *Answer Script A* (AI-generated, anonymized); (3) *Answer Script B* (human-written topper, anonymized). Graders are not told which script is human or AI.

How to evaluate (high-level). Use the official question paper to guide marking and apply the same standards used in real AoR evaluation. Award marks *per question* out of the official maximum (e.g., a 20-mark question must receive 0–20). Provide short notes where relevant and an overall comment per script.

Rubric dimensions.

- **Accuracy of law & reasoning:** Are principles stated correctly and applied to facts?
- **Case law & statutes:** Verify that cited authorities exist and are relevant; deduct for fabricated or incorrect citations.
- **Structure & coherence:** Clear issue → rule → application → conclusion flow; forum-appropriate organization.
- **Depth of analysis:** Beyond surface points; counter-arguments/nuances addressed where pertinent.
- **Language & expression:** Clear, professional, and appropriate for a Supreme Court exam answer.

Partial credit. Award partial credit wherever reasoning is substantively sound even if incomplete or imperfectly expressed.

Output expected from graders.

- Question-wise marks (out of the official marks allotted).
- Brief evaluator notes (e.g., "case not found," "well-structured," "analysis shallow").
- A 2–3 sentence overall comment on the paper's quality.

Note on Drafting (Paper II). Drafting is not part of the quantitative comparison in this study due to strict layout/form requirements. One pilot draft was graded and deemed "not evaluable" owing to pervasive formal defects; see Appendix §E for the summary note.

G Comprehensive Result Matrix Across LLMs and Years

We present the results of LLM Evaluations for CLAT UG, CLAT PG, DJS and DHJS in Table 4, Table 5, Table 6 and Table 7 respectively.

H Ethical Considerations

This research involved human evaluation of high-stakes professional materials, guided by the following ethical protocols:

- **Voluntary Expert Participation:** Certified Advocate-on-Record (AoR) evaluators participated on a voluntary basis. Their involvement was motivated by a professional interest in advancing understanding of technology within the legal field, and their contribution is gratefully acknowledged.
- **Managed Workload and Anonymization:** To respect the time of our volunteer experts, the evaluation workload was carefully limited to a manageable number of anonymized scripts. This prevented fatigue and ensured the integrity of the subjective assessment process.
- **Blinded Evaluation for Objectivity:** A paired, blind methodology was employed, making certain that evaluators were unable to differentiate between scripts created by humans and those by AI. This was essential in reducing bias and achieving unbiased and equitable comparisons. However, the stylistic disparities between human work and that of LLMs often resulted in comments suggesting suspicion that the responses might have been AI-generated.
- **Integrity and Transparency:** The study uses only officially released public materials to avoid compromising exam integrity. We transparently report both model capabilities and their significant limitations in procedural and drafting fidelity, emphasizing the continued necessity of human oversight in legal practice.

Year	Gemini 2.5 Pro	GPT 5 Chat	DeepSeek R1	DeepSeek Chat v3	Claude Sonnet 4	Mistral Medium 3.1	Qwen 3 235B	Llama 3.3 70B	GPT 4.1 Mini	Gemma 3 27B	Qwen 2.5 7B	Gemma 3 12B	GPT 3.5 Turbo	Llama 3.1 8B	Exam Topper
2008	145.75	134.25	136.5	136	136.75	**147**	136.75	132.75	125	119.25	103.5	114	109	99.75	–
2009	**179.25**	166.75	171.25	169.25	175	165.75	165.25	155.5	149.25	144.5	116	117.5	114.5	97	175
2010	**167**	157.75	162.75	159.25	158.25	154	159	151.25	142.75	139	129	137	128	112	165
2011	**173.5**	146.75	153.25	151.75	153.75	138	141	125.75	129.75	115.5	89.25	108.75	97.75	96.75	173
2012	154.75	138.75	146.25	144.75	152.75	149.75	147.75	120.5	128.5	109.75	96.5	110.5	104	80	**159**
2013	**168.5**	151.25	164.25	166	146.25	149	150	135.25	133.75	132.5	111.5	114.75	107.75	97	160.75
2014	**185.25**	147.75	166.5	172.5	175	169	164.25	151.5	135.25	132.5	114	119	115.25	103.25	171.75
2015	**156.25**	117	133	131	130	114.75	119.5	98	91	100	83.25	86	74	30.75	143.75
2016	**187.75**	164	178.25	165.5	174	166.5	169.75	151.75	145.75	140.5	113.75	116	118	91.5	174.5
2018	**180.25**	144	156.75	149	150.5	147.5	143.5	130.5	131.25	138.75	116.5	112.75	106.75	102.25	159
2019	**184**	145.5	163.25	162.5	165	146.75	154.75	130.5	132.25	127.75	104.75	115.75	94.75	98.5	177.25
2021	**136.25**	126.25	123.75	123.75	121.5	115	115	110.75	103.75	100.75	92.75	83.25	101.25	92.25	125.5
2022	**137.5**	116.5	125.75	116.5	127.25	121.5	122	114.5	106.5	107.75	88	101	87.75	95.5	121
2023	**130.5**	111.75	84	114.25	123	120.5	112	95	114.25	96.5	95.25	98.5	92	94.5	116.75
2024	99.75	93.5	98.5	97.25	100	96	100	87.5	92.25	84	84.75	83.75	87.25	77.75	**108**
2025	**107.75**	92.5	96.25	99	100.25	96.25	94	93.75	96.25	93	82.75	85.5	80	88.25	103.5

Table 4: CLAT UG Results.

Year	Gemini 2.5 Pro	GPT 5 Chat	DeepSeek R1	DeepSeek Chat v3	Claude Sonnet 4	Mistral Medium 3.1	Qwen 3 235B	Llama 3.3 70B	GPT 4.1 Mini	Gemma 3 27B	Qwen 2.5 7B	Gemma 3 12B	GPT 3.5 Turbo	Llama 3.1 8B	Exam Topper
2019	**85.27**	81.75	76.75	75.25	75.75	76.5	76.75	72.75	71.5	66.5	48.75	60.75	56.25	34.5	–
2020	**95.5**	99	79.5	83	76.25	79.5	58.75	76.25	65.25	52.5	40.5	44.25	51.25	47.75	72
2021	**105**	98.75	96.5	90.25	92	91.75	87.25	83.25	89.25	74.5	60.25	61.5	61.75	54	85.75
2022	**106.5**	102	94	93.25	87.25	92.75	90.5	84	77.75	73	61.75	61.75	66	51.75	94
2023	103.75	**105**	88	87.5	92.75	91.25	87.75	81.75	70.25	70.25	67	65.5	67.75	64	95.5
2024	**111.25**	110	100	101.75	94.5	98.75	90	90.5	81.75	76.25	66	73.75	71.5	66.25	104.25
2025	**95**	90.25	93	90.5	88.25	85.5	80	78.25	85.5	72	62.5	62.75	68	65.25	80

Table 5: CLAT PG Results.

Year	Gemini 2.5 Pro	GPT 5 Chat	DeepSeek R1	DeepSeek Chat v3	Claude Sonnet 4	Mistral Medium 3.1	Qwen 3 235B	Llama 3.3 70B	GPT 4.1 Mini	Gemma 3 27B	Qwen 2.5 7B	Gemma 3 12B	GPT 3.5 Turbo	Llama 3.1 8B	Exam Topper
2011	**151.75**	133.75	123.75	111.25	121.25	125	92.5	95	122.5	100	82.5	82.5	56.25	68.75	–
2014	**178.75**	156.25	140	127.5	145	145	131.5	115	125	112.5	85	87.5	78.75	73.5	–
2015	**191.25**	178.75	157.5	145	168.75	165	138.75	131.25	145	128.75	78.75	106.25	103.75	86.75	–
2017	**191.25**	178.75	161.25	147.5	160	153.75	123.75	122.5	131.25	120	102.5	111.25	91.25	76.75	–
2018	**178.75**	162.5	143	135	143	132.5	125	123.75	111.25	108.75	66.75	91.75	87	89.5	–
2019	**170**	155	129.75	132.25	136	145	116.25	112.5	98.75	100	56	83.5	66	78.55	–
2022	**185**	145	148.75	136.25	151.25	147.5	131.5	108.75	120	97.5	72.5	83.75	88.75	87.75	–

Table 6: Delhi Judicial Services (DJS) Results. Topper Marks are not released.

Year	Gemini 2.5 Pro	GPT 5 Chat	DeepSeek R1	DeepSeek Chat v3	Claude Sonnet 4	Mistral Medium 3.1	Qwen 3 235B	Llama 3.3 70B	GPT 4.1 Mini	Gemma 3 27B	Qwen 2.5 7B	Gemma 3 12B	GPT 3.5 Turbo	Llama 3.1 8B	Exam Topper
2013	**151.75**	133.75	123.75	111.25	121.25	125	92.5	95	122.5	100	82.5	82.5	56.25	68.75	–
2017	**130**	112.5	102.5	72.5	105	97.5	80	91.25	91.25	86.25	70	80	75	54.5	–
2019	**145**	128.75	120.5	112.5	118.75	117.5	90	90	96.25	85	60	83.75	63.75	58.5	–
2022	**143.75**	127.5	113.75	111.25	115	106.5	77.5	91.25	86.25	73.75	47.5	62.5	56.25	50.5	–
2023	**138.75**	108.75	116.25	111.25	112.5	102.5	85	91.25	86.25	78.75	50	73.75	55	54.5	–

Table 7: Delhi Higher Judicial Services (DHJS) Results. Topper Marks are not released.

Exam	Modality	Contents (official)
CLAT UG	Objective	English Language; Current Affairs / GK; Legal Reasoning; Logical Reasoning; Quantitative Techniques.
CLAT PG	Objective	Core LL.B. subjects: Constitutional Law; Jurisprudence; Administrative Law; Contract; Torts; Family; Criminal; Property; Company; Public International Law; Tax; Environmental; Labour / Industrial Law.
DJS (Prelim.)	Objective	Constitution; CPC; CrPC / BNSS; IPC / BNS; Evidence; Contract; Partnership; Arbitration; Specific Relief; Limitation; POCSO; Commercial Courts Act; English / GK.
DHJS (Prelim.)	Objective	DJS core plus commercial/statutory expansion: TPA; Sale of Goods; Negotiable Instruments; Succession / Hindu laws; Prevention of Corruption; POCSO; SARFAESI / DRT; Labour laws; Commercial Courts; IT; IPRs; English / GK.
AoR (SC)	Subjective	Four descriptive papers: Practice & Procedure of the Supreme Court; Drafting; Advocacy & Professional Ethics; Leading Cases (official case list).

Table 8: Appendix syllabus/contents (official sources: Consortium of NLUs for CLAT; Delhi High Court for DJS/DHJS; Supreme Court of India for AoR) (Consortium of National Law Universities, 2025b; Consortium of NLUs, 2025; High Court of Delhi, 2023, 2024; Supreme Court of India, 2025).

Exam	Mod.	Years included (official paper + official key)	Qs/exam	Total MCQs
CLAT UG	Obj.	2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2018, 2019, 2021, 2022, 2023, 2024, 2025 (2017 and 2020 excluded due to key issues).	200/150/120	**3,154**
CLAT PG	Obj.	2019, 2020, 2021, 2022, 2023, 2024, 2025	120	**814**
DJS (Prelim.)	Obj.	2011, 2014, 2015, 2017, 2018, 2019, 2022	200	**1,400**
DHJS (Prelim.)	Obj.	2013, 2017, 2019, 2022, 2023	150	**850**
AoR (SC)	Subj.	2023 (blinded grader study)	-	-

Table 9: Appendix year coverage used in this study (objective papers require both official paper and official answer key). AoR is fully descriptive (no MCQs).