SG-LegalClarify: a Large-scale Benchmark for Legal Reasoning and Ambiguity Recognition in Singapore Law

Ziwei Gu[1*], Michael S. Lee[2*], Drew Rein[2], Yash Maurya[2], Vikneswaran Kumaran[3], Darshini Ramiah[3], Chenguang Wang[4], Udari Madhushani Sehwag[2]

[1]Harvard University, [2]Scale AI, [3]Infocomm Media Development Authority (IMDA), [4]University of California, Santa Cruz

udari.sehwag@scale.com

Abstract

Singapore offers a distinctive testbed for legal reasoning by large language models (LLMs). Its legal system combines inherited English common law with jurisdiction-specific statutes, requiring models to distinguish local doctrine from generic common-law intuition. This setting also highlights the challenge of recognizing underspecification in lay legal queries, because the facts material to a legal outcome depend on the governing doctrine. Yet legal benchmarks rarely cover Singapore law or test this form of ambiguity recognition. To address these gaps, we introduce SG-LEGALCLARIFY, a large-scale benchmark for open-ended legal reasoning and ambiguity recognition in Singapore law. SG-LEGALCLARIFY contains 1,100 lawyer-authored source questions spanning 10 legal domains, including 340 fully specified and 760 underspecified questions. Each underspecified question has a fixed hidden-fact assignment and a fully specified counterpart, yielding 1,860 evaluation instances. Responses are graded with lawyer-authored Issue, Rule, Application, and Conclusion (IRAC) rubrics. We evaluate 19 LLMs, closed-book and with web search, using a multi-turn pipeline in which models decide whether to answer or clarify, and a user simulator reveals hidden facts only when asked. We find the benchmark far from saturated: the best models satisfy only 69.8% of mandatory rubric items closed-book (average 53.8%) and 77.8% with web search (average 63.0%). Models also struggle to recognize when clarification is necessary. We will release the public split of our benchmark, the model outputs on it, and the evaluation pipeline to support future research.

1 Introduction

Singapore offers a distinctive setting for evaluating legal AI. Its legal system is among the most highly regarded in the world, ranking in the global top ten on civil and criminal justice [51]. It is also a common-law jurisdiction in which inherited English law applies but subject to local modification and

[*]Equal contribution.

Singapore-specific statutes [25]. A model therefore cannot rely on generic common-law intuition only and must identify when Singapore-specific doctrine governs.

The difficulty is compounded when large language models (LLMs) give legal guidance to lay users, who often do not know which facts the law makes material [11, 29], resulting in underspecified queries (Figure 1). To answer such queries, a model must clarify selectively, asking when a material fact is missing even if the query is linguistically clear, and answering directly when the information is sufficient, rather than burdening users with unnecessary questions. The clarification decision is itself part of legal reasoning, since recognizing that a fact is missing requires identifying the legal issue and recalling the rule that makes the fact material.

Existing legal benchmarks leave gaps in assessing selective clarification and do not pose clarification as a decision. Most evaluate models on fixed inputs treated as answerable as given, emphasizing legal knowledge and reasoning but not whether a model recognizes that legally material information is missing and seeks clarification [20, 21, 26, 36]. Recent consultation benchmarks do require models to elicit facts through multi-turn dialogue [10, 49, 52, 54], but they focus on queries requiring clarification only, so they measure whether a model clarifies, not whether it clarifies only when it should. Additionally, failures can also be harder to trace, as a flawed conclusion may stem from a key fact the model fails to elicit, or a misidentified issue or rule [15]. Rubric-based benchmarks are beginning to localize which criteria a response misses [12, 34, 49, 54], but selective clarification requires evaluation framework beyond fine-grained rubrics. For Singapore, existing work focuses on compliance detection, embedding retrieval, or citation retrieval [9, 19, 45], rather than open-ended legal reasoning and interaction with users.

To address these gaps, we introduce SG-LEGALCLARIFY, a benchmark that contains 1,100 lawyer-authored source questions across 10 domains of Singapore law: 340 fully specified (FS) and 760 underspecified (US). For each US question, the contributor specifies a fixed assignment of values to the omitted facts and writes a paired FS counterpart that states those facts upfront, yielding 1,860 evaluation instances: 1,100 FS and 760 US. Each question carries a lawyer-authored rubric organized by Issue, Rule Recall, Application, and Conclusion (IRAC),[1] and each US question shares its rubric with its FS counterpart. These pairs allow us to compare legal analysis when the same material facts must be elicited versus provided upfront. We evaluate 19 LLMs in closed-book and web-enabled settings. Under a prompt that does not reveal specification status, the model must decide whether to answer or clarify. Then, a tool-gated user simulator adapted from τ^2-Bench [8] reveals a hidden fact only when asked. This design lets us assess whether models recognize when clarification is needed, ask about relevant hidden facts, and produce correct legal analysis within each IRAC component.

Our contributions can be summarized as follows: (1) We introduce SG-LEGALCLARIFY, a large-scale, lawyer-authored benchmark for open-ended legal reasoning and selective clarification across 10 domains of Singapore law. (2) We develop a multi-turn evaluation framework that measures selective clarification and fact elicitation and compares legal analysis across paired US/FS instances. (3) Our evaluation of 19 LLMs reveals substantial headroom even with web search and difficulty recognizing when clarification is necessary.

2 Dataset

[1] IRAC is a widely taught organizational framework for legal analysis [41]. We use its components as a rubric schema; models are not required to use IRAC headings.

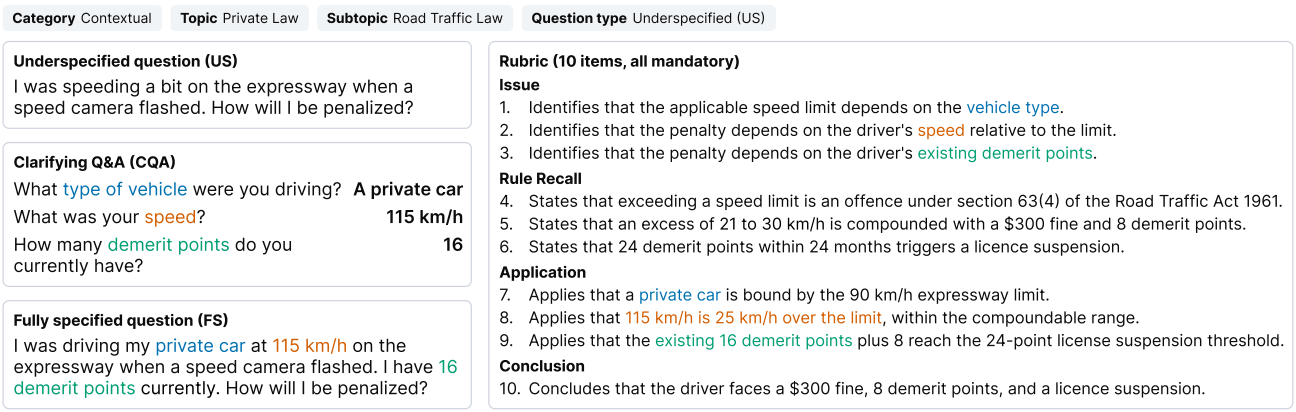


Figure 1. An example source question from SG-LEGALCLARIFY. The underspecified (US) question omits three legally material facts, the clarifying questions and answers (CQA) reveal them, and the fully specified (FS) version states them outright. Responses are graded against a lawyer-written IRAC rubric. Colors mark the same variable across CQA, FS, and rubric.

2.1 Dataset Description

SG-LEGALCLARIFY contains 1,100 English-language source questions about Singapore law, each requiring a free-form response and accompanied by a reference answer and grading rubric (§2.3). The questions span 10 domains, grouped into six practice-oriented legal domains and four legal-system knowledge domains (Figure 2b).

Each source question is labeled by **category** and **specification**. Factual questions (580) ask for a legal rule, definition, process, or institutional fact, whereas Contextual questions (520) require applying the law to a described situation. FS questions (340) include all contributor-identified facts needed to determine the answer, whereas US questions (760) omit at least one legally material variable. By construction, all 520 Contextual source questions are US; the 580 Factual questions comprise 340 FS and 240 US. Each US source question has a manually written paired FS counterpart, yielding 1,860 evaluation instances: 760 matched US–FS pairs and 340 additional FS questions (Figure 2a). Each pair shares the same underlying legal content as well as rubric, enabling paired comparisons under identical grading criteria with facts available upfront versus through interaction.

2.2 Dataset Construction

Contributors. Questions were authored by 23 contributors, all verified lawyers (2 to 37 years of experience, median 6) who held either a Singapore law degree, or an overseas law degree together with active legal practice in Singapore. All completed project-specific quality training. For each item, contributors were assigned a category, domain, subtopic, and specification label, and could decline assignments outside their expertise. Every item was independently reviewed by one of seven contributors. Reviewer edits were final. More details about the contributors are in Appendix A.

Authoring workflow. Contributors were instructed to author questions to resemble requests that a layperson might pose to an LLM assistant. For each question, the contributor supplied the question text, reference answer, supporting legal references, and grading rubric (§2.3). For each US question, the contributor identified the omitted legally material variables and provided, for each variable, a clarifying question and corresponding answer (CQA). We call the joint assignment of these hidden fact values the

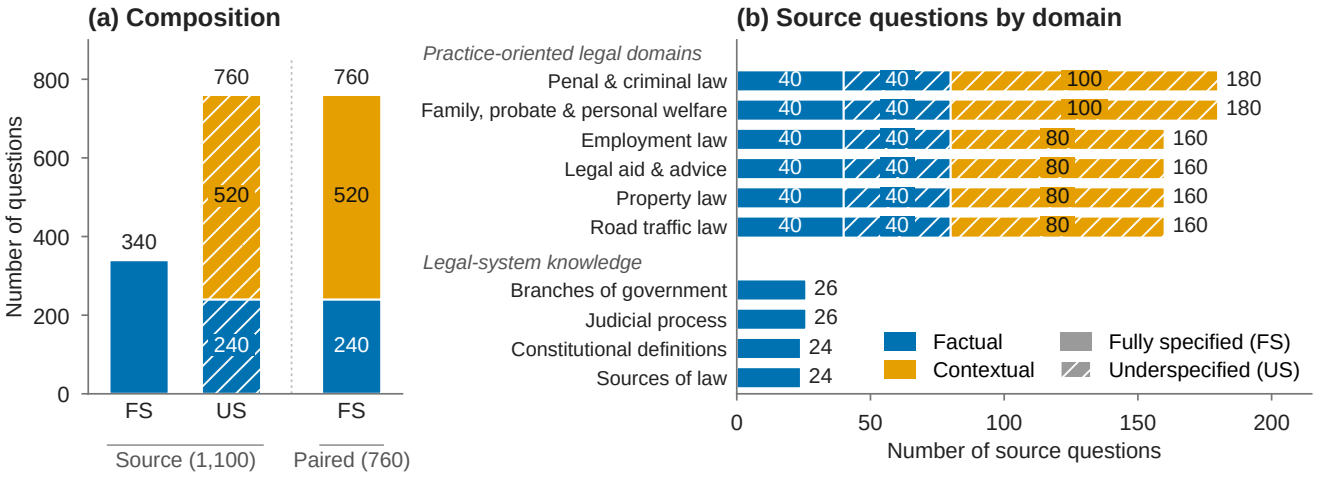


Figure 2. Composition of SG-LEGALCLARIFY. (a) The 1,100 source questions comprise 340 FS and 760 US questions. Each US question has a manually written paired FS counterpart, yielding 1,860 evaluation instances. (b) Source questions per domain, excluding the 760 paired counterparts.

		% of mandatory items			
Dimension	Credits the model for	Factual	Contextual	All	% Mand.
Issue	Identifying a relevant legal issue or variable	13.1	24.3	19.0	98.6
Rule Recall	Stating or correctly describing the governing law	58.5	40.6	49.2	82.5
Application	Applying the rule to the facts of the instance	14.6	21.4	18.2	94.0
Conclusion	Reaching the correct legal outcome	13.8	13.7	13.7	96.0
Mandatory items per source question (mean)		9.2	11.2	10.1	

Table 1. IRAC rubric dimensions and their distribution over mandatory items (Factual n=5,329; Contextual n=5,833; all n=11,162). % Mand. is the share of each dimension tagged Mandatory.

point-instance. The point-instance defines a single completed fact pattern with a determinate answer under the benchmark rubric. Figure 1 illustrates this construction, including an US question, the CQAs for the omitted variables, the paired FS counterpart, and the shared rubric.

US questions omit between 1 and 16 variables (median 3; mean 3.1). Fixing one instance keeps the reference answer and rubric focused on a single fact pattern, while the interactive evaluation still tests which missing variables the model asks about. The contributor then manually created the paired FS counterpart by incorporating the point-instance facts while preserving the underlying question. We will release 50% of source questions in each category and retain the remaining 550 for private evaluation, keeping each US-FS pair in the same split.

2.3 IRAC Rubrics

Each evaluation instance is graded against a rubric of 1–23 binary (satisfied/not satisfied) items. Across the 1,100 source questions, contributors authored 12,546 distinct rubric items. Applying each shared rubric to both members of its US-FS pair yields 21,762 instance-level rubric items in total (mean 11.7 per evaluation instance). Contributors were instructed to make each rubric item atomic, self-contained, and non-overlapping. Each item carries two tags: **dimension** and **importance**.

Dimensions. Following recent rubric-based legal evaluations [12, 20, 34, 49], we score responses using

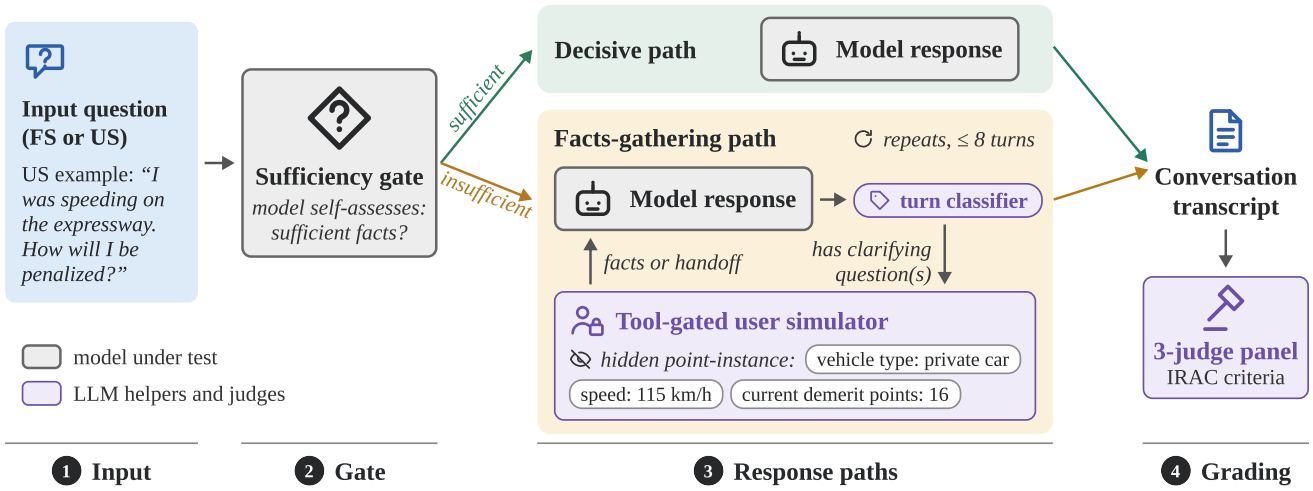


Figure 3. Two-stage sufficiency-gated evaluation. The model under test selects a decisive or facts-gathering path at the gate. A panel of three LLM judges grades the transcript against IRAC rubrics.

itemized criteria. We organize these criteria into four IRAC-inspired dimensions [41]: Issue, Rule Recall, Application, and Conclusion. This scheme evaluates distinct parts of the model's expressed legal analysis, and can localize errors such as correct rule recall followed by flawed application. Not every rubric contains items from all four dimensions. Table 1 defines the dimensions and reports their distribution.

Importance. Of the 12,546 distinct rubric items, 11,162 (89.0%) are Mandatory, meaning they are necessary for a substantively correct response. The remaining 1,384 (11.0%) are Optional and capture useful but nonessential content, such as citing a specific statute when no citation is requested or discussing additional requirements after one condition has already determined the outcome. Headline results in §4 are computed over mandatory items only.

3 Experiments

3.1 Models

We evaluate the same 19 general-purpose LLMs in closed-book and web-search conditions. Open-weight models are grouped by total parameter count. **Proprietary models** (9): Claude Fable 5 [5], Claude Opus 4.8 [6], and Claude Sonnet 5 [7]; GPT-5.6 Sol, Terra, and Luna [46]; Gemini 3.1 Pro [23] and Gemini 3.5 Flash [24]; and Qwen3.7 Plus [4]. **Larger open-weight models** (6): DeepSeek-V4 Pro and Flash [17]; Kimi K2.6 [43] and Kimi K3 [31]; Mistral Large 3 [42]; and Llama 4 Scout [40]. **Smaller open-weight models** (4): Qwen3.6 27B [48] and Qwen-SEA-LION-v4.5-27B-IT [2]; Gemma 3 27B IT [22] and Gemma-SEA-LION-v4-27B-IT [1]. The two SEA-LION models are Southeast Asia-adapted members of AI Singapore's SEA-LION family [44]. Their respective base models, Qwen3.6 27B and Gemma 3 27B IT, are included for comparisons of regional adaptation. Access routes, reasoning and sampling configurations, and output-token ceilings are detailed in Appendix B.

3.2 Multi-Turn Evaluation Pipeline

Figure 3 illustrates our experiment pipeline. Every FS and US input question uses the same two-stage sufficiency-gated prompt scheme (Appendix D) without disclosing specification status. We run closed-

book and web-search conditions, with search available only during the evaluated model's dialogue turns. The gate and helpers remain closed-book.

Sufficiency gate. Given a question, a separate call asks the evaluated model to output `SUFFICIENT_FACTS` or `INSUFFICIENT_FACTS`. The harness then supplies only the chosen path's system prompt. This decoupling makes the sufficiency judgment an explicit, directly measurable decision, separate from the model's subsequent behavior of either clarifying or answering. We keep this gate closed-book so that routing reflects the model's judgment from the question itself and remains comparable across retrieval conditions, rather than depending on retrieved evidence. Appendix C reports how often models actually request facts after being routed to facts-gathering.

Response paths. If the model selects sufficient facts, it enters the Decisive path and is instructed to provide direct legal analysis and a decisive conclusion without clarifying questions. The conversation ends after its first response, without invoking the user simulator. Otherwise, the model enters the Facts-gathering path and interacts with the user simulator for up to eight model turns. An LLM turn classifier (Opus 4.6) detects whether there are clarifying questions about concrete facts in the model response. If not, the classifier ends the conversation. Conversations always end on a model response. The model is instructed to output `INSUFFICIENT_FACTS_GATHERED` rather than guess when facts remain insufficient.

Tool-gated user simulator. Following tool-constrained user simulation [8, 18], we build a simulator (Opus 4.8) that maps fact requests in the latest model response to ground-truth CQAs, selects the variables that are explicitly asked, and rephrases the answers into natural sentences without adding new information. When two consecutive model responses match no fact variable, a decisive handoff is triggered that permits one final model turn within the turn limit. FS questions trigger this handoff immediately. Simulator-selected variables are recorded for each question.

Validated in §3.5, the LLM turn classifier and user simulator run without reasoning under fixed configurations (Appendix B). All reported rubric scores use criterion-level majority votes from the selected three-model panel (§3.4). The planned grid is $3 \times 19 \times 1{,}860 = 106{,}020$ generation trials per condition, or 212,040 total, excluding validation.

3.3 Metrics

Our primary metric, **mandatory score**, is the percentage of mandatory rubric criteria passed in a trial. It is macro-averaged across trials for each model and condition, overall and separately for FS and US questions. Optional rubric criteria are excluded.

Multiple metrics characterize different aspects of a model's clarification behavior. **Gate selectivity** compares the recorded route at the sufficiency gate with the specification label. **Clarification rate** records at least one concrete fact request. On US trials, the user simulator records which hidden variables were requested by the model. For each model and run, let m_i be the number of variables requested, v_i the total hidden variables, and N_{US} the number of US trials. Task-averaged **variable coverage** is $\frac{100}{N_{\text{US}}} \sum_{i=1}^{N_{\text{US}}} m_i / v_i$, including trials with no requests ($m_i = 0$), weighting trials equally rather than pooling variables.

3.4 LLM-as-Judge Meta-Evaluation

Human reference. We stratify 110 source tasks across the taxonomy: 34 FS-only tasks and 76 US–FS pairs, yielding 186 evaluation instances. Opus 4.8, DeepSeek-V4 Flash, and Qwen-SEA-LION-v4.5-27B-IT generate 558 closed-book conversations with reasoning enabled. Two reviewers grade each response

against binary IRAC criteria, without adjudication or label merging. Across all 110 tasks (6,522 paired criterion items), raw agreement is 80.4% and Fleiss' $\kappa = 0.607$ (95% CI $[0.569, 0.643]$; question-cluster bootstrap).

Judge-panel selection. Nineteen judges use a shared frozen grading prompt derived from the human guidelines, with reasoning and search disabled. We enumerate all 969 three-judge panels, take criterion-level majority votes, and rank panels by mean Cohen's κ against the two annotation copies. The top judge panel—**DeepSeek-V4 Flash, GPT-5.6 Sol, and Kimi K3**—scores $\kappa = 0.697$ and 84.8% agreement on this reference. These are in-sample estimates. This selected panel grades both production conditions.

Grouped five-fold cross-validation on the same reference yields pooled held-out $\kappa = 0.690$ and 84.5% agreement, compared with $\kappa = 0.678$ for single-judge selection. Panels and individual judges are selected within training folds, keeping each source's variants, responses, and criteria together. This evaluates panel selection, not the fixed production panel. Appendix E details the reference and protocol.

3.5 Additional Human Validation of LLM Helpers

We audit fact-request recognition by the LLM turn classifier and faithful disclosure of requested facts by the LLM user simulator. Two independent annotations, from three contributors, cover each of 55 tasks: 17 FS-only tasks and 38 US–FS pairs. Annotators agree perfectly on which facts were disclosed (Fleiss' $\kappa = 1.00$) and substantially on which were requested ($\kappa = 0.77$); agreement on whether a turn warrants a simulator reply is $\kappa = 0.82$ for 64 FS turns and 0.71 for 38 US conversations. Across 64 bank facts in 24 US conversations with simulator replies, the simulator achieves **89.1% fact-level disclosure accuracy** (93.8% and 84.4% by annotation copy), where a requested fact should be disclosed and an unrequested fact withheld. Among these 24 conversations, both annotators identify over-disclosure in four (16.7%), with five additional disputed cases; neither identifies under-disclosure. Both reviewers label as fact requests all nine FS turns whose positive classifier decisions can be reconstructed and matched to turn-specific human labels. Appendix F details the scoring, alignment, and annotation reliability.

4 Results

4.1 Web search improves performance but leaves substantial headroom

Table 2 reports mandatory scores for all 19 models, combining FS and US questions. Fable 5 leads closed book at **69.8%**; Kimi K3 leads the web-enabled condition at **77.8%**. All 19 models improve with web search, but gains vary substantially: DeepSeek V4 Flash gains the most, rising from 50.1% to 74.0% (+23.9 percentage points). The mean across models rises from 53.8% to 63.0% (+9.2 percentage points), **leaving substantial headroom even with web search enabled**. Rule Recall remains the weakest IRAC component (43.2% closed book; 56.5% web-enabled). Appendix G disaggregates performance by task type, specification status, IRAC component, and legal domain, and reports model-level search use and per-model clarification behavior.

4.2 Sufficiency recognition limits selective clarification

The closed-book gate routes 37.8% of US trials to the decisive path and 31.3% of FS trials to fact gathering. Figure 4 shows that models cluster within a narrow diagonal band, $12.6 \leq \text{TPR} - \text{FPR} \leq 38.9$ percentage points. Detection and false-positive rates tend to rise together across models ($r = 0.90$): Luna detects 89.3% of US trials but flags 64.1% of FS trials, whereas Mistral detects only 16.2% with 3.5% false

	Model	Closed book	Web enabled
Proprietary	Fable 5	**69.8**±0.5	73.3±0.1
	Opus 4.8	66.6±0.2	72.9±0.2
	GPT-5.6 Sol	64.6±0.3	70.5±0.3
	Gemini 3.5 Flash	62.5±0.4	71.6±0.4
	Sonnet 5	61.2±0.2	70.1±0.3
	GPT-5.6 Terra	59.8±0.4	67.1±0.1
	GPT-5.6 Luna	58.4±0.3	66.7±0.1
	Gemini 3.1 Pro	57.2±0.1	59.9±0.7
	Qwen3.7 Plus	49.9±0.5	67.9*±0.6
Larger open-weight	Kimi K3	65.9±0.3	**77.8**±1.2
	Kimi K2.6	54.2±0.4	68.9±0.7
	DeepSeek V4 Pro	52.4±0.1	73.1±1.1
	Mistral 3 Large	51.8±0.2	57.2±0.2
	DeepSeek V4 Flash	50.1±0.3	74.0±0.4
	Llama 4 Scout	36.0±0.4	37.5†
Smaller open-weight	Qwen3.6 27B	47.0±0.4	58.9±0.3
	Qwen-SEA-LION v4.5 27B	46.1±0.3	54.7±1.1
	Gemma 3 27B	35.0±0.3	37.7±0.4
	Gemma-SEA-LION v4 27B	34.5±0.1	37.3±0.4

Table 2. Mandatory scores (%), FS+US combined. Model rows show mean ± sample SD across three seeds; SD is estimated from rounded seed scores. Rows are sorted by Closed book score within each group.

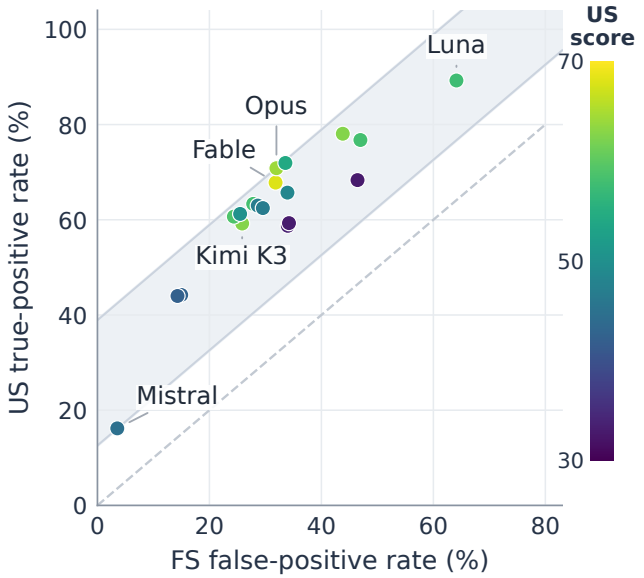


Figure 4. Closed-book gate selectivity. Each point represents one model. Color denotes mandatory score across all US trials. TPR and FPR are the proportions of US and FS trials routed to fact gathering. Shading marks the observed TPR-FPR band.

positives. Even the best balanced gate accuracy, weighting FS and US equally, is only 69.4% (Opus). **Greater sensitivity to missing facts thus accompanies more unnecessary routing to fact gathering**: models struggle to distinguish queries that need clarification from those already answerable. Evaluating clarification as an active decision exposes a failure mode that evaluations presupposing a need to clarify would miss.

4.3 Targeted clarification is associated with smaller missing-fact penalties

Figure 5(a) compares 760 matched US-FS pairs, holding the source questions constant. Every model scores higher on FS on average in both conditions. With web search, the gap appears to widen for 13 of 19 models, though we treat this as a descriptive trend. Thus, higher web-enabled scores do not imply that models have overcome underspecification. Models that clarify more frequently tend to have smaller paired gaps ($r = -0.87$ in both conditions; Figure 5(b)), as do those requesting a larger share of target facts ($r = -0.82$ closed book; $r = -0.90$ web-enabled; Figure 5(c)).

For absolute US performance, **what models ask is more informative than asking frequency alone** (Figure 6(c-d)): in closed book, target-request coverage correlates more strongly with US score than clarification rate does ($r = 0.73$ versus 0.40), and the web-enabled condition sees a similar pattern ($r = 0.67$ versus 0.35). These model-level associations do not establish causal effects: coverage measures facts requested, not verified receipt or successful use, and complete coverage need not be necessary when one fact determines the outcome.

*Two-seed mean. †One seed; no SD. Both models were deprovisioned mid-run, so only partial data are available and reported.

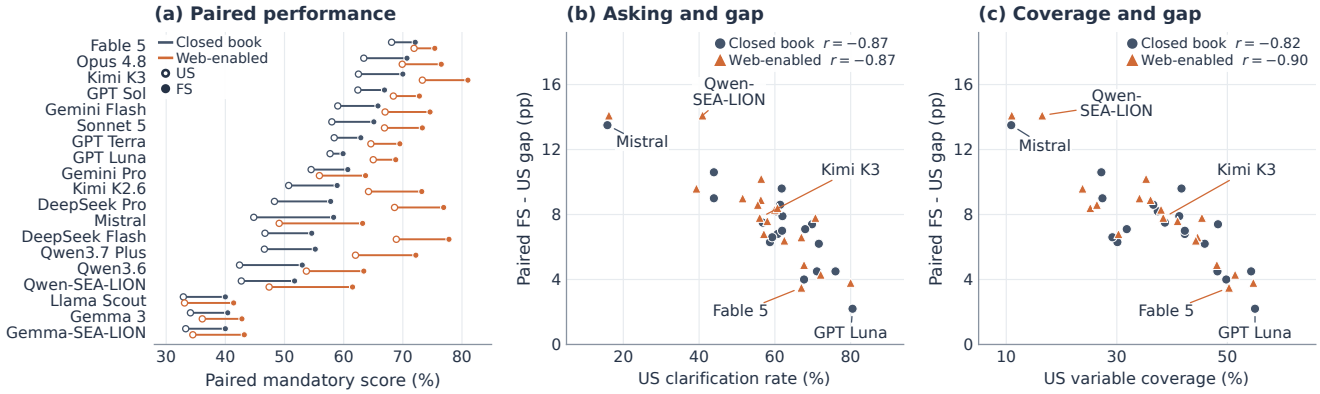


Figure 5. Paired performance and clarification. (a) Mean mandatory scores on 760 matched US-FS pairs. Each segment joins US and FS for one model and condition. (b-c) Paired FS–US gap versus US clarification rate and target-request coverage, respectively. Each point represents one model and condition. Pearson r is computed across the 19 models within each condition.

4.4 Conclusion-rubric and non-conclusions reveal distinct failure modes

Figure 6(a–b) separates rubric-passing and rubric-failing responses treated as committed from detected non-conclusions. Among closed-book US facts-gathering trials, the rubric-failing share decreases with greater target-request coverage: 45.8%, 36.6%, and 34.1% for zero, partial, and complete coverage. Complete-coverage trials also have a smaller rubric-failing share than direct-answer trials (34.1% versus 54.5%), though 23.2% are detected as non-conclusions, leaving rubric-passing shares similar (42.8% versus 45.5%). Among FS trials routed to fact gathering, 49.3% are detected as non-conclusions, despite the required facts already being supplied.

With web search, detected non-conclusion rates are lower across all facts-gathering groups, but rubric-failing shares do not consistently decline. For complete-coverage US trials, the rubric-passing share rises to 55.6%: detected non-conclusions fall from 23.2% to 12.3%, whereas the rubric-failing share declines only from 34.1% to 32.1%. This failing share instead rises slightly in the zero-target US group (45.8% to 48.6%) and the FS facts-gathering group (29.2% to 31.6%). The more consistent change is therefore in detected non-conclusions, with rubric-failing responses remaining common. Group membership varies across routes and conditions, so these comparisons are descriptive.

5 Related Work

5.1 Legal Benchmarks for LLMs

A growing family of benchmarks evaluates LLMs on jurisdiction-specific law. LegalBench [26] collaboratively assembles tasks spanning six types of legal reasoning, primarily over U.S. law. For Chinese law, LawBench [21] tests legal tasks across three cognitive levels, LexEval [36] covers a broader range of legal tasks, LAiW [16] organizes its tasks around legal syllogism, and LegalAgentBench [37] evaluates tool-using agents. Exam-derived benchmarks include LEXam [20], built from Swiss, European, and international law exams, and GreekBarBench [12], which evaluates free-text answers grounded in statutes and case facts from Greek Bar examinations. Further work covers Korea [30, 33], Thailand [3], and India [28], as well as Arabic-language legal evaluation [27]. For Singapore law, AutoLaw studies compliance and violation detection [45], MLEB evaluates legal embeddings for retrieval [9], and SG-

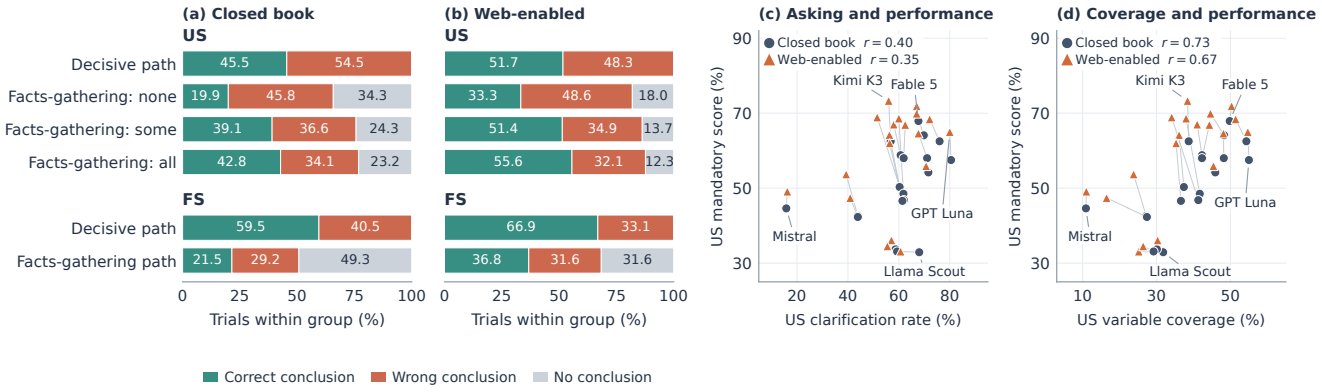


Figure 6. Clarification and performance. (a-b) Recorded conclusion outcomes by route and target-request coverage. Percentages normalize within rows. "Correct conclusion" and "Wrong conclusion" denote Conclusion-rubric pass/fail among responses treated as committed. "Decisive path" and "Facts-gathering" identify recorded routes; for US facts-gathering rows, "none/some/all" refers to target variables detected as requested. (c-d) US mandatory score versus US clarification rate and target-request coverage, respectively. Color and shape denote condition; segments connect the same model across conditions. Pearson r is computed across the 19 models within each condition.

LegalCite evaluates citation retrieval from Singapore judgments [19]. SG-LEGALCLARIFY complements these benchmarks by evaluating open-ended legal analysis and selective clarification for Singapore law. A parallel line of work develops legal-domain-adapted LLMs, including SaulLM [13], ChatLaw [14], and DISC-LawLLM [53]. Our experiments instead focus on contemporary general-purpose models and regional models for Southeast Asia, evaluating their ability to handle Singapore-specific reasoning without legal-domain adaptation.

Recent legal benchmarks begin to evaluate multi-turn consultation and fact elicitation. LexRAG [38] evaluates answers over pre-authored consultation dialogues, while PLawBench's public legal consultation task [49] mandates a one-shot list of follow-up questions without simulating responses or grading subsequent legal analysis. LeCoDe [52] lets a lawyer agent choose between questions and advice, evaluating key-fact recall and advice quality. FSDBench [10] evaluates multi-turn fact elicitation in Polish tax law without assessing the resulting legal answer. Closest to our setting, DLawBench [54] lets a model choose between further client questions and submitting a legal memorandum, assessing inquiry and legal analysis. SG-LEGALCLARIFY complements this literature by testing whether to clarify across mixed fully specified and underspecified inputs. Paired questions share a fixed fact assignment and rubric, allowing comparison of legal analysis when material facts must be elicited versus provided upfront.

5.2 Underspecification and Clarification in LLM Evaluation

Outside the legal domain, CLAMBER [56] finds that LLMs struggle to detect and clarify ambiguity, while AbstentionBench [32] reports abstention failures on unanswerable and underspecified questions. QuestBench studies which clarification questions resolve underspecified reasoning tasks [35], while Clarify When Necessary [55] and AskBench [57] study when and what models should ask. ClarifyMT-Bench [39] and RegretBench [50] evaluate multi-turn clarification policies. LHAW [47] and HiL-Bench [18] address underspecified reasoning in long-horizon agentic tasks. For user simulation, we adapt τ^2-Bench's use of tools and observable state to constrain our LLM-based user simulator [8]. In legal consultation, underspecification is conditioned by the governing doctrine. SG-LEGALCLARIFY evaluates this connection between jurisdiction-specific legal knowledge and selective clarification.

6 Limitations

Our study is subject to a set of limitations. First, human validation of the LLM helpers is limited in scope, and turn-classification errors or simulator over-disclosure may affect dialogue trajectories and downstream scores. Second, provider differences prevent identical search-tool access, so web-enabled results reflect both model capabilities and tool configurations. While the decoupled, closed-book gate makes sufficiency judgments explicit and supports controlled comparisons across retrieval conditions, its up-front routing may alter clarification behavior, reducing ecological validity. Finally, not every annotated variable is necessary to resolve a point-instance: a subset of facts may be dispositive, so incomplete target-request coverage need not indicate inadequate clarification.

7 Conclusion

We introduced SG-LegalClarify, a benchmark for selective clarification in Singapore legal reasoning that combines fully specified and underspecified questions, a fact-grounded user simulator, and rubric-based evaluation. Our evaluation of 19 models reveals persistent weaknesses in recognizing when clarification is needed and conducting correct legal analysis. Web search improves overall scores, but fact-gathering groups show more consistent reductions in non-conclusions than in rubric-failing conclusions. Together, these findings motivate evaluating legal assistants beyond answer quality or clarification frequency alone. SG-LegalClarify provides a testbed for jointly improving when models ask, what they ask, and how correctly they answer.

AI use statement

In this work, we used generative AI tools for coding and editing paper drafts more efficiently. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics Statement

SG-LEGALCLARIFY is intended to support research on reliable legal reasoning and clarification. Questions and rubrics were authored and reviewed by verified lawyers, who could decline assignments outside their expertise. Sensitive-topic assignments were restricted to contributors who opted in. Incorrect legal guidance can have serious consequences, and benchmark scores should not be interpreted as evidence that a model is safe for unsupervised legal advice.

Reproducibility Statement

We describe dataset construction, rubric design, contributor workflow, and review procedures. We further document the evaluation protocol, model and tool configurations, seed handling, metrics, and all system prompts for easier reproduction of our work. We will release the public benchmark split, full evaluation pipeline code, and model outputs on the public split.

References

[1] AI Singapore. Model card for Gemma-SEA-LION-v4-27B-IT, 2025. URL https://huggingface.co/aisingapore/Gemma-SEA-LION-v4-27B-IT.

[2] AI Singapore. Qwen-SEA-LION-v4.5-27B-IT, 2026. URL https://huggingface.co/aisingapore/Qwen-SEA-LION-v4.5-27B-IT.

[3] P. Akarajaradwong, P. Pothavorn, C. Chaksangchaichot, P. Tasawong, T. Nopparatbundit, K. Pratai, and S. Nutanong. NitiBench: Benchmarking LLM frameworks on Thai legal question answering capabilities. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34304–34327, Suzhou, China, Nov. 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1739. URL https://aclanthology.org/2025.emnlp-main.1739/.

[4] Alibaba Cloud Community. Qwen3.7-Plus: Multimodal agent intelligence, 2026. URL https://www.alibabacloud.com/blog/qwen3-7-plus-multimodal-agent-intelligence_603206.

[5] Anthropic. Claude Fable 5 and Claude Mythos 5, 2026. URL https://www.anthropic.com/news/claude-fable-5-mythos-5.

[6] Anthropic. Introducing Claude Opus 4.8, 2026. URL https://www.anthropic.com/news/claude-opus-4-8.

[7] Anthropic. Introducing Claude Sonnet 5, 2026. URL https://www.anthropic.com/news/claude-sonnet-5.

[8] V. Barrès, H. Dong, S. Ray, X. Si, and K. Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment. *CoRR*, abs/2506.07982, 2025. doi: 10.48550/ARXIV.2506.07982. URL https://doi.org/10.48550/arXiv.2506.07982.

[9] U. Butler, A. Butler, and A. L. Malec. The massive legal embedding benchmark (MLEB). *CoRR*, abs/2510.19365, 2025. doi: 10.48550/ARXIV.2510.19365. URL https://doi.org/10.48550/arXiv.2510.19365.

[10] M. Bystroński, K. Tagowski, D. Janiak, J. Farganus, L. Augustyniak, M. Kajdanowicz, and T. J. Kajdanowicz. Factual state discovery benchmark: Evaluating fact elicitation in Polish tax law. In S. T.Y.S.S., J. D. Rodriguez, and O. de Gibert, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1262–1279, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-393-7. doi: 10.18653/v1/2026.acl-srw.114. URL https://aclanthology.org/2026.acl-srw.114/.

[11] I. Cheong, K. Xia, K. J. K. Feng, Q. Z. Chen, and A. X. Zhang. (A)I am not a lawyer, but...: Engaging legal experts towards responsible LLM policies for legal advice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3-6, 2024*, pages 2454–2469. ACM, 2024. doi: 10.1145/3630106.3659048. URL https://doi.org/10.1145/3630106.3659048.

[12] O. S. Chlapanis, D. Galanis, N. Aletras, and I. Androutsopoulos. GreekBarBench: A challenging benchmark for free-text legal reasoning and citations. In C. Christodoulopoulos, T. Chakraborty,

C. Rose, and V. Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25099–25119, Suzhou, China, Nov. 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1368. URL https://aclanthology.org/2025.findings-emnlp.1368/.

[13] P. Colombo, T. P. Pires, M. Boudiaf, R. Melo, G. Hautreux, E. Malaboeuf, J. Charpentier, D. Culver, and M. Desa. Saullm-54b & saullm-141b: Scaling up domain adaptation for the legal domain. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ea3f85a33f9ba072058e3df233cf6cca-Abstract-Conference.html.

[14] J. Cui, Z. Li, Y. Yan, B. Chen, and L. Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*, abs/2306.16092, 2023. doi: 10.48550/ARXIV.2306.16092. URL https://doi.org/10.48550/arXiv.2306.16092.

[15] M. Dahl, V. Magesh, M. Suzgun, and D. E. Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, 2024.

[16] Y. Dai, D. Feng, J. Huang, H. Jia, Q. Xie, Y. Zhang, W. Han, W. Tian, and H. Wang. LAiW: A Chinese legal large language models benchmark. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10738–10766, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.716/.

[17] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *CoRR*, abs/2606.19348, 2026. doi: 10.48550/ARXIV.2606.19348. URL https://doi.org/10.48550/arXiv.2606.19348.

[18] M. Elfeki, T. Trinh, K. Luu, G. Luo, N. Hunt, E. Montoya, N. Marwaha, Y. He, C. Wang, F. Crabedo, A. Castilo, and B. Liu. Hil-bench (human-in-loop benchmark): Do agents know when to ask for help? *CoRR*, abs/2604.09408, 2026. doi: 10.48550/ARXIV.2604.09408. URL https://doi.org/10.48550/arXiv.2604.09408.

[19] S. L. Y. Ern, K. Feng, Y. Du, C. L. E. Jia, and Z. Sun. Sg-legalcite: A principle-augmented benchmark for legal citation retrieval in singapore law. *CoRR*, abs/2605.21057, 2026. doi: 10.48550/ARXIV.2605.21057. URL https://doi.org/10.48550/arXiv.2605.21057.

[20] Y. Fan, J. Ni, J. Merane, Y. Tian, Y. Hermstrüwer, Y. Huang, M. Akhtar, E. Salimbeni, F. Geering, O. Dreyer, D. Brunner, M. Leippold, M. Sachan, A. Stremitzer, C. Engel, E. Ash, and J. Niklaus. LEXam: Benchmarking legal reasoning on 340 law exams. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xNhbMyXsJn.

[21] Z. Fei, X. Shen, D. Zhu, F. Zhou, Z. Han, A. Huang, S. Zhang, K. Chen, Z. Yin, Z. Shen, J. Ge, and V. Ng. LawBench: Benchmarking legal knowledge of large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.452. URL https://aclanthology.org/2024.emnlp-main.452/.

[22] Gemma Team. Gemma 3 technical report. *CoRR*, abs/2503.19786, 2025. doi: 10.48550/ARXIV.2503.19786. URL https://doi.org/10.48550/arXiv.2503.19786.

[23] Google DeepMind. Gemini 3.1 Pro. Model card, Google DeepMind, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

[24] Google DeepMind. Gemini 3.5 Flash. Model card, Google DeepMind, 2026. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/.

[25] Government of Singapore. Application of English Law Act 1993 (2020 revised edition). Singapore Statutes Online, 1993. URL https://sso.agc.gov.sg/Act/AELA1993.

[26] N. Guha, J. Nyarko, D. E. Ho, C. Ré, A. Chilton, K. Aditya, A. Chohlas-Wood, A. Peters, B. Waldon, D. N. Rockmore, D. Zambrano, D. Talisman, E. Hoque, F. Surani, F. Fagan, G. Sarfaty, G. M. Dickinson, H. Porat, J. Hegland, J. Wu, J. Nudell, J. Niklaus, J. J. Nay, J. H. Choi, K. Tobia, M. Hagan, M. Ma, M. A. Livermore, N. Rasumov-Rahe, N. Holzenberger, N. Kolt, P. Henderson, S. Rehaag, S. Goel, S. Gao, S. Williams, S. Gandhi, T. Zur, V. Iyer, and Z. Li. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/89e44582fd28ddfea1ea4dcb0ebbf4b0-Abstract-Datasets_and_Benchmarks.html.

[27] F. Hijazi, S. AlHarbi, A. AlHussein, H. Abu Shairah, R. AlZahrani, H. AlShamlan, O. Knio, and G. Turkiyyah. ArabLegalEval: A multitask benchmark for assessing Arabic legal knowledge in large language models. In N. Habash, H. Bouamor, R. Eskander, N. Tomeh, I. Abu Farha, A. Abdelali, S. Touileb, I. Hamed, Y. Onaizan, B. Alhafni, W. Antoun, S. Khalifa, H. Haddad, I. Zitouni, B. AlKhamissi, R. Almatham, and K. Mrini, editors, *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 225–249, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.arabicnlp-1.20. URL https://aclanthology.org/2024.arabicnlp-1.20/.

[28] A. Joshi, S. Paul, A. Sharma, P. Goyal, S. Ghosh, and A. Modi. IL-TUR: Benchmark for Indian legal text understanding and reasoning. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11460–11499, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.618. URL https://aclanthology.org/2024.acl-long.618/.

[29] S. Kapoor, P. Henderson, and A. Narayanan. Promises and pitfalls of artificial intelligence for legal applications. *CoRR*, abs/2402.01656, 2024. doi: 10.48550/ARXIV.2402.01656. URL https://doi.org/10.48550/arXiv.2402.01656.

[30] Y. Kim, Y. Choi, E. Choi, J. Choi, H. J. Park, and W. Hwang. Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5573–5595, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.319. URL https://aclanthology.org/2024.findings-emnlp.319/.

[31] Kimi Team. Kimi K3: open frontier intelligence. *CoRR*, abs/2607.24653, 2026. doi: 10.48550/ARXIV.2607.24653. URL https://doi.org/10.48550/arXiv.2607.24653.

[32] P. Kirichenko, M. Ibrahim, K. Chaudhuri, and S. J. Bell. Abstentionbench: Reasoning llms fail on unanswerable questions. In D. Belgrave, C. Zhang, L. N. Montoya, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, N. Chen, I. V. M. Ruíz, and A. Loaiza-Bonilla, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/fb122bfc3f0127a94ded048b5b03496f-Abstract-Datasets_and_Benchmarks_Track.html.

[33] J. Lee, D. Kim, S. Hwang, H. Kim, and G. Lee. KoBLEX: Open legal question answering with multi-hop reasoning. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4019–4053, Suzhou, China, Nov. 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.200. URL https://aclanthology.org/2025.emnlp-main.200/.

[34] J. Lee, K.-W. On, S. S. Han, A. Cohan, and J. Hockenmaier. Evaluating legal reasoning traces with legal issue tree rubrics. In M. Liakata, V. P. Moreira, J. Zhang, and D. Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3299–3322, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.150. URL https://aclanthology.org/2026.acl-long.150/.

[35] B. Z. Li, B. Kim, and Z. Wang. Questbench: Can llms ask the right question to acquire information in reasoning tasks? In D. Belgrave, C. Zhang, L. N. Montoya, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, N. Chen, I. V. M. Ruíz, and A. Loaiza-Bonilla, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/c42c8d51556fabb4b57fc86d3d3d0d09-Abstract-Datasets_and_Benchmarks_Track.html.

[36] H. Li, Y. Chen, Q. Ai, Y. Wu, R. Zhang, and Y. Liu. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/2cb40fc022ca7bdc1a9a78b793661284-Abstract-Datasets_and_Benchmarks_Track.html.

[37] H. Li, J. Chen, J. Yang, Q. Ai, W. Jia, Y. Liu, K. Lin, Y. Wu, G. Yuan, Y. Hu, W. Wang, Y. Liu, and M. Huang. LegalAgentBench: Evaluating LLM agents in legal domain. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2322–2344, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.116. URL https://aclanthology.org/2025.acl-long.116/.

[38] H. Li, Y. Chen, Y. Hu, Q. Ai, J. Chen, X. Yang, J. Yang, Y. Wu, Z. Liu, and Y. Liu. Lexrag: Benchmarking retrieval-augmented generation in multi-turn legal consultation conversation. In N. Ferro,

M. Maistro, G. Pasi, O. Alonso, A. Trotman, and S. Verberne, editors, *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 3606–3615. ACM, 2025. doi: 10.1145/3726302.3730340. URL https://doi.org/10.1145/3726302.3730340.

[39] S. Luo, H. Yi, M. Li, S. Meng, F. Liu, Z. Hu, J. Feng, and Q. Liu. Clarifymt-bench: Benchmarking and improving multi-turn clarification for conversational large language models. *CoRR*, abs/2512.21120, 2025. doi: 10.48550/ARXIV.2512.21120. URL https://doi.org/10.48550/arXiv.2512.21120.

[40] Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

[41] J. Metzler. The importance of irac and legal writing. *U. Det. Mercy L. Rev.*, 80:501, 2003.

[42] Mistral AI. Introducing Mistral 3, 2025. URL https://mistral.ai/news/mistral-3/.

[43] Moonshot AI. Kimi K2.6: Advancing open-source coding, 2026. URL https://www.kimi.ai/blog/kimi-k2-6.

[44] R. Ng, T. N. Nguyen, H. Yuli, T. N. Chia, L. W. Yi, W. Q. Leong, X. Yong, J. G. Ngui, Y. Susanto, N. Cheng, H. Rengarajan, P. Limkonchotiwat, A. V. Hulagadri, K. W. Teng, Y. Y. Tong, B. Siow, W. Y. Teo, T. C. Meng, B. Ong, Z. H. Ong, J. R. Montalan, A. Chan, S. Antonyrex, R. Lee, E. Choa, D. O. Tat-Wee, B. J. D. Liu, W. C. Tjhi, E. Cambria, and L. Teo. SEA-LION: Southeast Asian languages in one network. In K. Inui, S. Sakti, H. Wang, D. F. Wong, P. Bhattacharyya, B. Banerjee, A. Ekbal, T. Chakraborty, and D. P. Singh, editors, *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 512–526, Mumbai, India, Dec. 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. doi: 10.18653/v1/2025.ijcnlp-long.30. URL https://aclanthology.org/2025.ijcnlp-long.30/.

[45] T. D. Nguyen, L. H. Pham, and J. Sun. AUTOLAW: enhancing legal compliance in large language models via case law generation and jury-inspired deliberation. *CoRR*, abs/2505.14015, 2025. doi: 10.48550/ARXIV.2505.14015. URL https://doi.org/10.48550/arXiv.2505.14015.

[46] OpenAI. GPT-5.6 system card. System card, OpenAI, 2026. URL https://deploymentsafety.openai.com/gpt-5-6.

[47] G. Pu, M. S. Lee, U. M. Sehwag, D. J. Lee, B. Zhu, Y. Maurya, M. Raghavendra, Y. Xue, and S. M. Denton. LHAW: controllable underspecification for long-horizon tasks. *CoRR*, abs/2602.10525, 2026. doi: 10.48550/ARXIV.2602.10525. URL https://doi.org/10.48550/arXiv.2602.10525.

[48] Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

[49] Y. Shi, H. Liu, Y. HU, S. Gaojie, X. Xinran, Y. Ma, T. Tang, L. Zhang, Q. Chen, F. Di, W. Lv, W. Wu, K. Yang, S. Yang, W. Wang, R. Shi, Q. Yuanyang, Y. Qi, Z. Jingwen, S. Xiaoyu, Y. Chen, Z. Yi, A. Yang, B. Yu, D. Liu, J. Lin, W. Shen, B. Zhao, C. L. A. Clarke, and H. Wei. PLAWBENCH: A rubric-based benchmark for evaluating LLMs in real-world legal practice. In M. Liakata, V. P. Moreira, J. Zhang, and D. Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10067–10116, San Diego, California,

United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.458. URL https://aclanthology.org/2026.acl-long.458/.

[50] M. N. Ta, M. A. T. Nguyen, D. D. Nguyen, Y. Wang, and P. Nakov. One more turn, less regret: A regret-based multi-turn benchmark for llms' clarification policies. *arXiv preprint arXiv:2607.21143*, 2026.

[51] World Justice Project. WJP rule of law index 2025, 2025. URL https://worldjusticeproject.org/rule-of-law-index/downloads/WJPIndex2025.pdf.

[52] W. Yuan, K. Song, Z. Jiang, J. Cao, Y. Zhang, J. Lin, K. Kuang, J. Zhang, and X. Liu. LeCoDe: A benchmark dataset for interactive legal consultation dialogue evaluation. In M. Liakata, V. P. Moreira, J. Zhang, and D. Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36019–36048, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1667. URL https://aclanthology.org/2026.acl-long.1667/.

[53] S. Yue, W. Chen, S. Wang, B. Li, C. Shen, S. Liu, Y. Zhou, Y. Xiao, S. Yun, X. Huang, and Z. Wei. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *CoRR*, abs/2309.11325, 2023. doi: 10.48550/ARXIV.2309.11325. URL https://doi.org/10.48550/arXiv.2309.11325.

[54] L. Zhang, Y. Shi, Y. Hu, J. Zhang, W. Lv, Y. Ma, W. Wang, R. Shi, Y. Qiu, X. Xu, Y. Qi, L. Miao, J. Savelka, Y. Liu, K. Ashley, B. Zhao, H. Wei, and L. Qu. Dlawbench: Evaluating llms through multi-turn legal consultation. *CoRR*, abs/2606.13931, 2026. doi: 10.48550/ARXIV.2606.13931. URL https://doi.org/10.48550/arXiv.2606.13931.

[55] M. J. Zhang and E. Choi. Clarify when necessary: Resolving ambiguity through interaction with LMs. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5541–5558, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.306. URL https://aclanthology.org/2025.findings-naacl.306/.

[56] T. Zhang, P. Qin, Y. Deng, C. Huang, W. Lei, J. Liu, D. Jin, H. Liang, and T.-S. Chua. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10746–10766, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.578. URL https://aclanthology.org/2024.acl-long.578/.

[57] J. Zhao, K. Fang, and L. Cheng. When and what to ask: AskBench and rubric-guided RLVR for LLM clarification. In M. Liakata, V. P. Moreira, J. Zhang, and D. Jurgens, editors, *Findings of the Association for Computational Linguistics: ACL 2026*, pages 17120–17140, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.845. URL https://aclanthology.org/2026.findings-acl.845/.

A Contributor Details

Recruitment and qualifications. All contributors were verified lawyers with demonstrated expertise in Singapore law. Two groups took part (Table 3): 23 contributors authored and reviewed benchmark items, and nine provided the human labels used to validate the LLM judges (Section 3.4). Items involving sensitive subject matter were assigned only to contributors who had opted in to such content.

Pool	*N*	Mean	Median	Min	Max
Question authors and reviewers	23	9.2	6	2	37
Meta-evaluation annotators	9	5.6	4	2	13

Table 3. Contributor pools and years of legal experience.

Review. Seven of the 23 authors also served as reviewers. The review covered (i) the specification label (fully specified or underspecified), (ii) the accuracy of the clarifying question and answer pairs, (iii) the legal accuracy of the expected answer, (iv) the accuracy of the listed legal references, (v) the rubrics, and (vi) the alignment of the category and domain labels with the item. Reviewers could edit any field or relabel the item. The reviewed version superseded the author's and was final, with no further adjudication.

Meta-evaluation annotators. Nine lawyers (2 to 13 years of experience, median 4) produced the human labels for candidate-judge evaluation (Section 3.4). Each response and rubric item received two independent binary judgments. Panel selection compared predictions separately with both annotation copies and averaged the two Cohen's κ values; no majority human reference or adjudication was used. Annotator identities can vary across questions.

B Model and Inference Configuration

Tables 4 and 5 list the 19 models' closed-book configurations and web-search routes and controls, respectively. For closed-book calls, supported models are requested to reason at high effort; Qwen-SEA-LION uses template-based thinking. The closed-book output ceiling is 16,384 tokens per call, except for Llama 4 Scout (8,192). Available reasoning outputs are logged alongside conversations and run metadata.

Routing. Gemini 3.1 Pro uses `gemini/gemini-3.1-pro-preview`; Mistral 3 Large uses `mistral/mistral-large-latest`; and Qwen3.6 27B uses `fireworks_ai/deployed/t5cvm6h4`. Llama 4 Scout uses `bedrock/us.meta.llama4-scout-17b-instruct-v1:0` closed-book and `groq/meta-llama/llama-4-scout-17b-16e-instruct` for its retained web-search run. Gemma 3 27B IT uses `bedrock/google.gemma-3-27b-it` closed-book, but switches to local `gemma-3-27b-it` for web search because the Bedrock route rejects the search-tool schema.

Web-search configuration. The eight OpenAI, Anthropic, and Gemini models use native provider search through the OpenAI Responses API, Anthropic citations, and Gemini grounding, respectively; the remaining 11 use a custom Serper–Jina search-and-visit loop. Native search uses a medium context

Model	Closed-book route	Reasoning	Temp.
Claude Fable 5	Anthropic	adaptive, high	1.0
Claude Opus 4.8	Anthropic	adaptive, high	1.0
Claude Sonnet 5	Anthropic	adaptive, high	1.0
GPT-5.6 Sol	OpenAI	high	1.0
GPT-5.6 Terra	OpenAI	high	1.0
GPT-5.6 Luna	OpenAI	high	1.0
Gemini 3.1 Pro	Google	high	default
Gemini 3.5 Flash	Google	high	default
Qwen3.7 Plus	Fireworks	high	0.7
Kimi K2.6	Fireworks	high	0.6
Kimi K3	Fireworks	high	0.6
DeepSeek-V4 Pro	Fireworks	high	0.6
DeepSeek-V4 Flash	Fireworks	high	0.6
Mistral 3 Large	Mistral	—	0.0
Qwen3.6 27B	Fireworks (private)	high	0.7
Llama 4 Scout	AWS Bedrock	—	0.0
Qwen-SEA-LION-v4.5-27B-IT	local vLLM	Qwen3 thinking	1.0
Gemma-SEA-LION-v4-27B-IT	local vLLM	—	1.0
Gemma 3 27B IT	AWS Bedrock	—	0.0

Table 4. Closed-book inference settings. High effort is requested through the provider-specific reasoning interface; — denotes no reasoning mode. Gemini temperature is unset. The client retries without temperature when a provider rejects it. SEA-LION sampling uses top-p 0.95, with top-k 20 for Qwen and 64 for Gemma; Qwen additionally uses min-p 0 and repetition penalty 1.

setting, with Anthropic limited to 100 search-tool uses per API call; OpenAI and Gemini use provider-managed limits. The custom loop permits up to 999 tool-call rounds, 9,999 searches, and 9,999 page visits per assistant turn. The models otherwise retain closed-book sampling settings shown in Table 4.

Seeds and run aggregation. The planned design comprises three runs per condition, with $19 \times 1{,}860 = 35{,}340$ generation trials per run: 106,020 per condition and 212,040 overall, including 43,320 planned US trials per condition. Metrics are computed within each model/run and then averaged over eligible runs. Following deprovisioning, web-search results retain run 1 for Llama 4 Scout and runs 1–2 for Qwen3.7 Plus. All other models retain three runs per condition. Validation studies and ablations are excluded from these totals.

Run directories 1–3 correspond by repository convention to seeds 42, 43, and 44; individual trial records do not stamp the seed/run ID. The closed-book client forwards seeds only to OpenAI and Mistral routes; the custom web-search client forwards seeds to local vLLM. Helpers are unseeded and use the same configuration across conditions; this does not imply identical generated replies.

Helper configuration. The configured defaults are Claude Opus 4.6 for the turn classifier, and Claude Opus 4.8 for simulator selection/rendering. Helpers run without reasoning; the sufficiency gate uses the evaluated model. Reported rubric scores use the selected three-model panel (§3.4). Helper model IDs are not stamped in individual trial records; the defaults above describe the inspected runner configuration.

Model	Web-search route	Search	Tool request
Claude Fable 5	Anthropic	Native	—
Claude Opus 4.8	Anthropic	Native	—
Claude Sonnet 5	Anthropic	Native	—
GPT-5.6 Sol	OpenAI	Native	—
GPT-5.6 Terra	OpenAI	Native	—
GPT-5.6 Luna	OpenAI	Native	—
Gemini 3.1 Pro	Google	Native	—
Gemini 3.5 Flash	Google	Native	—
Qwen3.7 Plus	Fireworks	Custom	API
Kimi K2.6	Fireworks	Custom	API
Kimi K3	Fireworks	Custom	API
DeepSeek-V4 Pro	Fireworks	Custom	API
DeepSeek-V4 Flash	Fireworks	Custom	API
Mistral 3 Large	Mistral	Custom	API
Qwen3.6 27B	Fireworks (private)	Custom	Passthrough
Llama 4 Scout	Groq	Custom	Wording
Qwen-SEA-LION-v4.5-27B-IT	local vLLM	Custom	Wording
Gemma-SEA-LION-v4-27B-IT	local vLLM	Custom	Wording
Gemma 3 27B IT	local vLLM	Custom	Wording

Table 5. Web-search routes and controls for the retained evaluation runs. Native denotes provider-hosted search; Custom denotes the Serper–Jina loop. Tool requests apply to the first custom-loop round on the facts-gathering path: API forces a tool call via `tool_choice`; Passthrough forwards this parameter through LiteLLM; Wording relies on a prompt instruction. — denotes not applicable. Local vLLM endpoints accept `tool_choice`, but the harness uses wording to avoid tool-format leaks in Gemma variants.

Condition	US (%)	FS (%)
Closed-book	97.7	90.9
Web-search	94.6	80.1

Table 6. Clarification conditional on routing to facts gathering, including inseparable parse-failure fallbacks. Rates average eligible runs within each model, then weight all 19 models equally. They describe behavior after routing and do not remove the gate's constraint on the opportunity to clarify.

C Routing and conditional clarification.

Gate precision and recall compare the recorded route with FS/US labels, treating US and facts gathering as positive. FS includes 340 original FS questions and 760 paired FS counterparts; US includes the 760 original US questions. We compute the conditional fraction separately for each model, eligible run, and specification status, then average over runs within each model. Table 6 averages these model means equally across the 19 models.

D System Prompts

D.1 Sufficiency gate.

The evaluated model receives the same gate prompt in both conditions, closed-book with reasoning requested where supported.

> You are an expert legal AI assistant advising a user on Singapore law. Your objective is to provide accurate, definitive legal conclusions based on the facts provided by the user.
>
> Decision protocol:
> 1) First determine whether the facts given are legally sufficient to reach a decisive conclusion on the question asked. Facts are legally sufficient if they allow for a decisive legal conclusion, and any missing facts would not be outcome-determinative.
> 2) If sufficient, output the exact phrase SUFFICIENT_FACTS. Or if insufficient, output the exact phrase INSUFFICIENT_FACTS.

The parser checks `INSUFFICIENT_FACTS` before `SUFFICIENT_FACTS`, retries an unrecognized output once, and then defaults to facts gathering. The selected path prompt replaces the gate prompt for dialogue; search is available only during web-condition dialogue turns.

D.2 Decisive path.

> You are an expert legal AI assistant advising a user on Singapore law. Your objective is to provide accurate, definitive legal conclusions based on the facts provided by the user.
>
> General rules:
> - Provide a direct legal analysis and a decisive conclusion based only on provided facts. Do not ask any clarifying questions.
> - Do not hedge solely for caution or risk-avoidance.
> - Identify and keep reasoning focused on legally relevant variables and applicable Singapore law.
> - Verbalize the legally relevant variables in your response.

The harness ends the conversation after one assistant response.

D.3 Facts-gathering path.

> You are an expert legal AI assistant advising a user on Singapore law. Your objective is to provide accurate, definitive legal conclusions based on the facts provided by the user.
>
> You are engaging in a multi-turn conversation.
> PHASE 1: FACTS GATHERING
> Act responsibly and ask clarifying questions for outcome-determinative variables that are missing from the user's initial query. Do not ask redundant or low-impact questions.
> PHASE 2: LEGAL ASSESSMENT
> If you have gathered sufficient facts from the user, you must provide the most decisive conclusion possible.
> OR
> If the facts are still legally insufficient, DO NOT guess or hallucinate hidden assumptions. You must output the exact phrase INSUFFICIENT_FACTS_GATHERED and state what is still missing.
>
> General rules:
> - Do not hedge solely for caution or risk-avoidance.
> - Identify and keep reasoning focused on legally relevant variables and applicable Singapore law.

- Verbalize the legally relevant variables in your response.

A turn containing a concrete question invokes the simulator even if it also contains a conclusion. A conclusion without a question, or neither flag, ends the conversation. No simulator reply is generated after the last allowed assistant turn.

D.4 Turn classifier.

You are a precise classifier. Respond only with valid JSON.
Classify the following assistant message from a legal consultation.

Message:
"""
{assistant_message}
"""

Respond with a JSON object containing:
- "has_clarifying_question": true ONLY if the assistant message includes at least one CONCRETE question requesting particular facts about the user's situation. Set to false for general invitations like 'feel free to share more details' or 'if you have more information, let me know' — these are NOT clarifying questions.
- Treat numbered or bulleted fact requests as clarifying questions even if there is no literal '?' character, as long as they are CONCRETE questions requesting particular facts about the user's situation.
- If the assistant asks the user to answer specific factual items (e.g., dates, participants, events, amounts, prior actions), set "has_clarifying_question" to true.
- "has_definitive_answer": true if the assistant provides a case-specific conclusion, including probabilistic but committed conclusions (e.g., 'likely valid/invalid', 'likely liable/not liable', 'the claim likely succeeds/fails'). Set to false if the assistant only lists generic possibilities without connecting to the user's facts.

Respond ONLY with the JSON object, no other text.

Output is a JSON object with `has_clarifying_question` and `has_definitive_answer`. Parsing removes code fences and applies `json.loads`. On failure, the heuristic marks a question when it finds a question mark or a numbered/bulleted list with request cues; it sets the answer flag to false. Cues include "please answer," "please provide," "to assess," "to determine," "I need," and "clarify."

D.5 Single-pass fact selector.

You are a strict JSON key selector. Respond only with valid JSON.

You are selecting variable keys for a user simulator.

You receive:
1) The assistant's latest message
2) Available variable keys, each with a key question and key fact answer

IMPORTANT MATCHING RULES:
- Evaluate EACH key independently and decide whether it should be selected.
- Select a key only when the assistant explicitly asks about that key's fact topic, or a clear semantic equivalent of that exact fact.
- Use BOTH the key question and key answer meaning when deciding.
- If the assistant asks compound questions spanning multiple topics, select all and only the directly asked topics.
- Do NOT select keys that are merely related downstream implications of a broader question.
- Do NOT select keys for generic discussion, process-only suggestions, or legal consequence summaries.
- If uncertain whether a key was directly asked, set select=false.

Assistant message:
"""
{assistant_message}
"""

Available keys (question + answer descriptions):
{tools_block}

Respond ONLY with valid JSON in this format:
{
"decisions": [
{"key": "key_1", "select": true, "reason": "short reason"},
{"key": "key_2", "select": false, "reason": "short reason"}
],
"selected_keys": ["key_1"],
"reason": "brief explanation"
}

The selector receives the latest assistant message and per-key descriptions, not the dialogue history. After stripping fences and parsing JSON, the implementation reads selected `decisions` and the legacy `selected_keys/tool_calls` forms, then maps keys to tools. Exceptions return an empty selection. The optional second recall pass is disabled.

D.6 Answer rendering.

Binary-answer system message and template:

You rewrite yes/no facts into concise sentences while preserving the question subject.
Rewrite the binary answer into one short, natural sentence.
Use the same subject/perspective as the question.
Do not add facts. Preserve only what is implied by the question and yes/no answer.
Return plain text only.

Question: {question}
Answer: {answer}

Non-binary-answer system message and template:

You rewrite short fact answers into concise, self-contained sentences.
Rewrite the answer into one short, natural, complete sentence.
Use the question text for context so the sentence is self-contained.
Do not add new facts beyond the question and answer.
Keep uncertainty if the answer is uncertain.
Return plain text only.

Question: {question}
Answer: {cleaned_answer}

Both return plain text. Exceptions use deterministic question-grounded sentences. The harness joins tool outputs; an empty selection yields "I'm not sure about that."

D.7 Rubric judge.

You are an expert legal evaluator assessing the quality of an LLM's response to a Singapore law question, applying the same grading guidelines as human contributors. The goal of this project is to ensure that chatbots provide accurate, Singapore-specific answers when users in Singapore ask questions in English that have legal ramifications.

You will be given:
1. The original question
2. The full conversation (including any clarifying questions and answers)
3. A set of rubric criteria to evaluate, each tagged with its dimension (Issue, Rule Recall, Application, or Conclusion)

For each rubric criterion, decide whether the assistant's responses in the conversation meet it: Met ("passed": true) or Not Met ("passed": false). When a criterion is ambiguous, apply the rules below rather than personal judgment.

Core "Met" bar (applies throughout):
- Grade substance, not phrasing. A criterion is Met if the response conveys the required point in any wording; it need not mirror the criterion's words.
- Met requires the point to be present and correct, not merely gestured at. A vague allusion a layperson couldn't recognize as the point = Not Met.
- Partial = Not Met (binary; no half credit). If a criterion bundles two required elements and only one is present, mark Not Met.
- Judge each criterion independently, only against what it asks. Correct content elsewhere doesn't help it; errors elsewhere don't fail it.
- Match the criterion's specificity. Specifically named points must be specifically met; generally named points must be generally met.
- In general, these guidelines support a strict reading and application of the rubric criteria.

Dimension-specific rules:
- Issue criteria ("Identifies whether ..."): Met only if the response explicitly surfaces the specific variable or scope condition as something the outcome depends on. A response can do this by asking about the variable, conditioning its analysis on it, or stating its legal significance outright. The test:

would a layperson reading the response learn that this specific fact matters to the outcome? A passing mention inside general doctrine coverage that does not treat the variable as a fact the outcome turns on is Not Met.
- Rule Recall criteria ("State that under section X of the Y Act, ..."): Met if the legal rule is stated correctly — the exact section number is NOT required (a correct citation is a bonus). A wrong statute/section or misstated rule = Not Met (incorrect law is worse than uncited law). Require the citation only when the provision itself is the point (e.g., "Identify Section 320(g) as the provision defining grievous hurt"). Leniency on the section number does not extend to omitting the rule itself.
- Application criteria ("Applies fact X ..."): Met only if the response uses the fact in its legal reasoning (fact -> rule/outcome). A fact merely restated, or not referenced at all, is Not Met. Any linkage between fact and rule/outcome suffices; the response does not need to use the word "apply".
- Conclusion criteria ("Conclude that ..."): Met if the response commits to a direct answer/position on the user's question (caveats are fine). A hedged answer (e.g., "you are likely to face ...") still counts as a conclusion if it identifies a specific outcome, and earlier clarifying questions do not negate a final committed answer. Not Met if the response only lists considerations, asks for more facts, or never commits. An explicit "I can't answer yet" / INSUFFICIENT_FACTS_GATHERED is not a conclusion = Not Met. Conditional forks ("if dangerous then X; if negligent then Y") with the deciding fact left unresolved do not count as a conclusion.

Cross-cutting rules:
- Criteria with multiple elements ("State A and B"): all required elements must be present to be Met. Conversely, if all are present, mark Met — do not over-require. Grade each element in the context the criterion specifies (a figure stated correctly but attached to a different case than the one the criterion names does not help).
- Criteria with qualifiers ("Outcome is X if Y is met"): grade the qualifier at the criterion's level of specificity. Conditioning only on unnamed, generic qualifications when the criterion names a specific gate is Not Met; stating the outcome as unconditional, with the named qualifier missing entirely, is Not Met.

Respond with a JSON array where each element has:
- "id": the rubric criterion id
- "passed": true (Met) or false (Not Met)
- "reasoning": a brief explanation (1-2 sentences)

Include exactly one element for every rubric criterion, in the order given.
Respond ONLY with the JSON array, no other text.

The corresponding evaluation-message template is:

Original Question: {scenario['user_question']}

Full Conversation:
{conv_text}

Rubric Items to Evaluate:
{rubric_text}

Evaluate each rubric item. Respond with a JSON array only.

Each rubric line has the form `- ID {id} [dimension: {dimension}]: {item}`; importance tags are hidden. The parser strips fences, requires a JSON array, and coerces supported boolean, integer, and textual verdicts to true/false; unrecognized verdicts become null. A parsing exception returns null for every criterion with an evaluation-error note. Verdicts are merged by criterion ID; an omitted ID becomes null/"Not evaluated." The candidate-judge sweep retries incomplete outputs as described in Appendix E.

D.8 Deterministic handoff.

The simulator injects one of these user messages without an LLM call:

> I've provided all the information I have, please make a decisive conclusion based on the information you have so far.

> I'm not sure, please make a decisive conclusion based on the information you have so far.

No-variable questions receive the first message. Repetition is detected when two "I'm not sure" replies repeat or selected keys are already covered; the handoff uses the first message if all keys are covered and the second otherwise. The harness permits one final assistant response within the turn limit.

D.9 Conclusion-presence classifier.

Only final responses containing the insufficiency marker receive this analysis-time check. The default is Opus 4.8, with verdicts cached by a hash of the final text.

> You are a strict legal-output classifier. Determine whether the assistant's final message contains a clear legal conclusion/position for the user's case. A clear conclusion means the assistant ultimately takes a legal position (for example liable/not liable, entitled/not entitled, can/cannot, likely/unlikely), even if caveated. Mere fact-listing, process advice, or saying facts are insufficient without any legal position is not a clear conclusion.
> Respond with exactly one token: YES or NO.

E Judge Meta-Evaluation Details

Response generation. The three response models use their closed-book inference settings (Appendix B), with reasoning enabled and no web search. We generate one unseeded response per evaluation instance and model.

Judge inference. All 19 candidate judges use the same prompt, SHA-256-stamped per record, and grade closed-book with reasoning disabled. Temperature is 1 for Claude and GPT-5.6 models and 0 for the remaining candidates. The runner requests 16,384 output tokens, uses no seed, and retries up to three times while any verdict is null.

Annotation guidelines. Reviewers grade substance rather than wording, assess each criterion independently, and award no partial credit. Issue criteria require explicitly surfacing a fact or condition that

Panel	Copy 1 κ	Copy 2 κ	Mean κ
DeepSeek-V4 Flash + GPT-5.6 Sol + Kimi K3	0.7008	0.6939	0.6973
DeepSeek-V4 Flash + GPT-5.6 Terra + Kimi K3	0.6963	0.6965	0.6964
GPT-5.6 Terra + Gemma 3 27B + Kimi K3	0.7023	0.6900	0.6961

Table 7. Top three panels among 969 candidates on the common 6,522-item reference. The first row is the retained production panel. Mean values are computed before rounding; all values are in-sample estimates.

affects the outcome. Rule Recall criteria require the correct substantive rule; exact citations are optional unless identifying the provision is itself required, whereas an incorrect statute or section fails the criterion. Application requires connecting a fact to a rule or outcome, not merely restating it. Conclusion requires commitment to a specific outcome: caveats are permitted, but unresolved conditional alternatives are insufficient. Required elements and qualifiers must be satisfied at the criterion's stated specificity.

Human reference and agreement. The common reference contains 6,522 criterion items with paired human judgments and matching predictions from all 19 judges. It covers all 110 source questions, 186 evaluation instances, and 558 conversations; auxiliary conclusion-presence checks, deleted entries, and unmatched criteria are excluded. Raw agreement is 80.4%, with Fleiss' $\kappa = 0.607$ (95% CI $[0.569, 0.643]$). The percentile interval uses 20,000 bootstrap resamples of source questions, retaining all variants, responses, criteria, and both judgments within each sampled question. Low-confidence ratings are retained. Contributor identities can differ by question: Fleiss' κ summarizes human agreement without treating the annotation copies as two fixed individuals and is invariant to swapping their order within an item. Panel–human comparisons instead use Cohen's κ against each complete annotation copy.

Human disagreement. Reviewers disagree on 1,280 of 6,522 criterion instances (19.6%). Using the evaluation rubric's dimension assignments, Fleiss' κ is 0.505 for Issue, 0.555 for Rule Recall, 0.582 for Application, and 0.682 for Conclusion. Qualitative inspection identifies recurring differences over citation optionality, whether a fact is merely mentioned or applied, and whether a qualified answer commits to an outcome. Overlapping rubric criteria can also produce correlated disagreements.

Panel objective and comparison. For panel P, let $\hat{y}_i^P$ be its majority-vote label on rubric-item instance i, and let $y^{(r)}$ denote annotation copy r. We rank panels by

$$\bar{\kappa}(P) = \frac{1}{2}\left[\kappa(\hat{y}^P, y^{(1)}) + \kappa(\hat{y}^P, y^{(2)})\right].$$

Human labels are never averaged into a synthetic reference. Table 7 reports the top three panels on the common reference. The strongest individual judge is Kimi K3 ($\kappa = 0.678$). These are in-sample comparisons and do not establish a uniquely superior panel.

Missing predictions. The loader lets later records supersede earlier records for the same judge and item. After this resolution, all 19 judges have non-null predictions for every item in the common reference. The panel rule ignores null votes and maps ties, including all-missing cases, to Not Met. These branches are not exercised here. Individual judges would otherwise be scored on their own non-null subsets.

Analysis	Paired items	Mean κ	Agreement (%)
Panel selection, held-out	6,522	0.6903	84.51
Single-judge selection, held-out	6,522	0.6781	83.90

Table 8. Pooled out-of-fold predictions on the common reference. Panels and single judges are selected separately within each training fold. These are point estimates from one fixed partition; no significance claim is made.

Fold	Test items	Training-selected panel	Held-out κ
1	1,281	GPT-5.6 Sol + Mistral 3 Large + Qwen3.7 Plus	0.7205
2	1,083	DeepSeek V4 Flash + GPT-5.6 Terra + Kimi K3	0.6517
3	1,536	DeepSeek V4 Flash + GPT-5.6 Sol + Kimi K3	0.6893
4	1,251	DeepSeek V4 Flash + GPT-5.6 Terra + Kimi K3	0.7113
5	1,371	GPT-5.6 Terra + Gemma 3 27B + Kimi K3	0.6663

Table 9. Source-grouped fold results on 6,522 paired items. Each fold holds out 22 source questions. Four panels are selected across five folds; the production panel is selected once.

Grouped cross-validation. We permute sorted source-question IDs with NumPy and divide them into five folds of 22 sources. Each source's FS/US variants, response models, and criteria remain together. In every fold, we select among all 969 panels on the other 88 sources, using the mean-κ objective above and lexicographic tie-breaking. We then evaluate only on the held-out sources. We pool each item's one held-out prediction, compute κ against each annotation copy, and average the two values, rather than averaging fold κs. A single-judge comparator is selected independently within each training fold; Kimi K3 wins all five. Table 8 reports pooled results, and Table 9 reports the selected panels.

Scope of the CV result. On the common reference, full-data selection favors DeepSeek-V4 Flash, GPT-5.6 Sol, and Kimi K3 ($\kappa = 0.6973$). Grouped CV evaluates the panel-selection procedure conditional on fixed prompts and judge outputs, not the fixed production panel or earlier prompt tuning. We retain the top panel on the common reference; these results do not establish a uniquely best panel or a statistically significant advantage over individual judges.

Own-response diagnostic. On the common 6,522-item reference, own-versus-other response pass rates are 0.630 versus 0.454 for Opus 4.8, 0.440 versus 0.502 for DeepSeek-V4 Flash, and 0.431 versus 0.547 for Qwen-SEA-LION. These contrasts are confounded by response quality and do not independently establish self-preference. GPT-5.6 Sol and Kimi K3 have no own responses in this reference set, so their self-preference cannot be assessed here. Grouped CV does not resolve self- or same-family preference; no production judge-replacement sensitivity is established by these diagnostics.

F Human Validation of LLM Helpers

Components and protocol. The turn classifier should recognize explicit fact requests warranting another simulator reply. The user simulator should supply requested facts without unrequested disclosures or unsupported additions. Three contributors provided two independent annotations for each of 55 tasks (17 FS-only and 38 US–FS pairs). One annotated all 55; the others supplied second copies for 40 and 15 tasks, yielding 110 annotations. All results are pre-adjudication.

Annotation units. *Engagement* asks whether the simulator should provide another factual response. FS engagement is annotated per assistant turn; US engagement is exported once per conversation, without a defined referent turn in multi-round dialogues. Per-fact labels indicate whether each bank item was explicitly requested and disclosed anywhere in the conversation. Engagement can be Yes when every bank fact-request label is No, because requests may concern facts outside the bank. Annotators separately flag unsupported factual information outside the bank (*hallucination*) and FS replies exceeding the fixed handoff template.

Simulator performance. We evaluate disclosure accuracy on the 64 bank facts in the 24 US conversations containing simulator replies, excluding 14 conversations without a reply. For each fact and reviewer, handling is correct when a requested fact is disclosed or an unrequested fact is withheld. The simulator achieves **89.1% accuracy**: 114 of 128 fact–reviewer comparisons are correct (98 requested-and-disclosed; 16 unrequested-and-withheld). This equals the mean of the two annotation-copy scores, 60/64 (93.8%) and 54/64 (84.4%); the denominator represents two ratings of 64 facts. All 14 errors are unrequested disclosures. Disclosure precision is 98/112 (87.5%), and requested-fact recall is 98/98 (100%) in this reply-containing subset. These conversation-wide labels do not establish that requests preceded disclosures; outside-bank hallucination is assessed separately.

Turn-classifier performance. All 93 annotated FS/US transcripts match saved calibration conversations after Unicode and whitespace normalization. Run metadata and handoff messages distinguish classifier invocations from bypasses: 46 FS and 14 US conversations take the decisive path, and nine FS second turns follow a forced handoff. These turns bypass classification, so absence of a reply cannot be scored as a classifier prediction. In the remaining nine FS first turns, continuation establishes a positive fact-request decision, and both reviewers independently label every turn Yes: **9/9 turns agree with both reviewers**. This positive-only subset leaves performance on non-requests unmeasured. US engagement labels cannot be matched to individual classifier decisions because their referent turn is undefined. We therefore do not report an overall turn-classifier accuracy.

Human agreement. Table 10 reports human–human agreement, Fleiss' κ for the varying reviewer pairs, and Gwet's AC1. Reported 95% percentile confidence intervals use 2,000 task-level bootstrap resamples. N/A is a separate category where present. Fact-level agreement includes all 96 bank facts across the 38 US conversations.

Disagreements and scope. All 11 fact-request disagreements across the 96 bank facts reflect the common contributor's broader interpretation of a request; six concern the 64 facts used for simulator performance. Reviewers agree on which bank facts were disclosed. In the 24 reply-containing conversations, neither annotator flags under-disclosure, and both hallucination-positive annotations are disputed. All ten FS beyond-template disputes concern No versus N/A conventions. The small sample, common-rater design, visible continuation, and absence of adjudication limit generalization. The simulator score measures annotated bank-fact handling; the classifier result covers only the nine aligned positive decisions.

Judgment	n	Agree. (%)	Fleiss' κ [95% CI]	AC1
Engagement (FS turns)	64	92.2	0.82 [0.61, 0.96]	0.86
Engagement (US conversations)	38	86.8	0.71 [0.47, 0.92]	0.83
Fact requested	96	88.5	0.77 [0.58, 0.92]	0.77
Fact disclosed	96	100.0	1.00	1.00
FS beyond-template	64	84.4	0.52 [0.19, 0.74]	0.77
Under-disclosure	24	100.0	undefined	1.00
Hallucination	24	91.7	−0.04	0.91

Table 10. Agreement between human annotators, distinct from helper performance. Each unit has two ratings. The final two rows include only US conversations with a simulator reply. Confidence intervals are shown where reported.

Note. Under-disclosure κ is undefined because all ratings in this subset are negative: expected agreement is $p_e = 1$, making $1 - p_e = 0$.

G Additional Results

Figures 7–9 break down performance by IRAC component, task type, and legal domain. These summaries pool available trials across models and runs. Table 11 reports search behavior in the web-search condition.



Figure 7. IRAC component pass rates. Closed-book and web-enabled results, pooling FS and US trials. Each component uses its own eligible trials, excluding trials without applicable mandatory criteria for that component. Rates are separate component summaries.

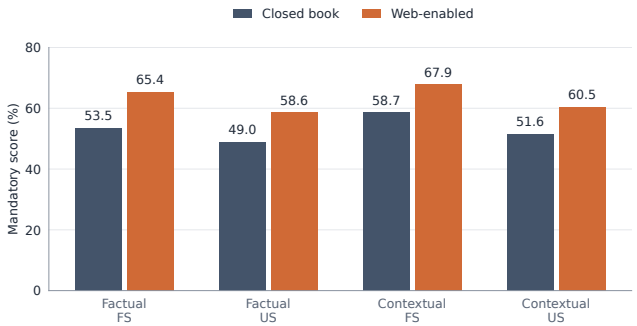


Figure 8. Performance by task type and specification. Mandatory scores for factual and contextual questions, separately for FS and US. Factual FS includes original unpaired questions and paired counterparts; all contextual questions have FS–US pairs. The factual FS–US contrast therefore compares unmatched sets.



Figure 9. Performance across legal domains. Mandatory scores pooled across FS and US trials in each domain. Segments connect closed-book and web-enabled scores; labels show web-enabled minus closed-book differences in percentage points. Domains are ordered by this difference.

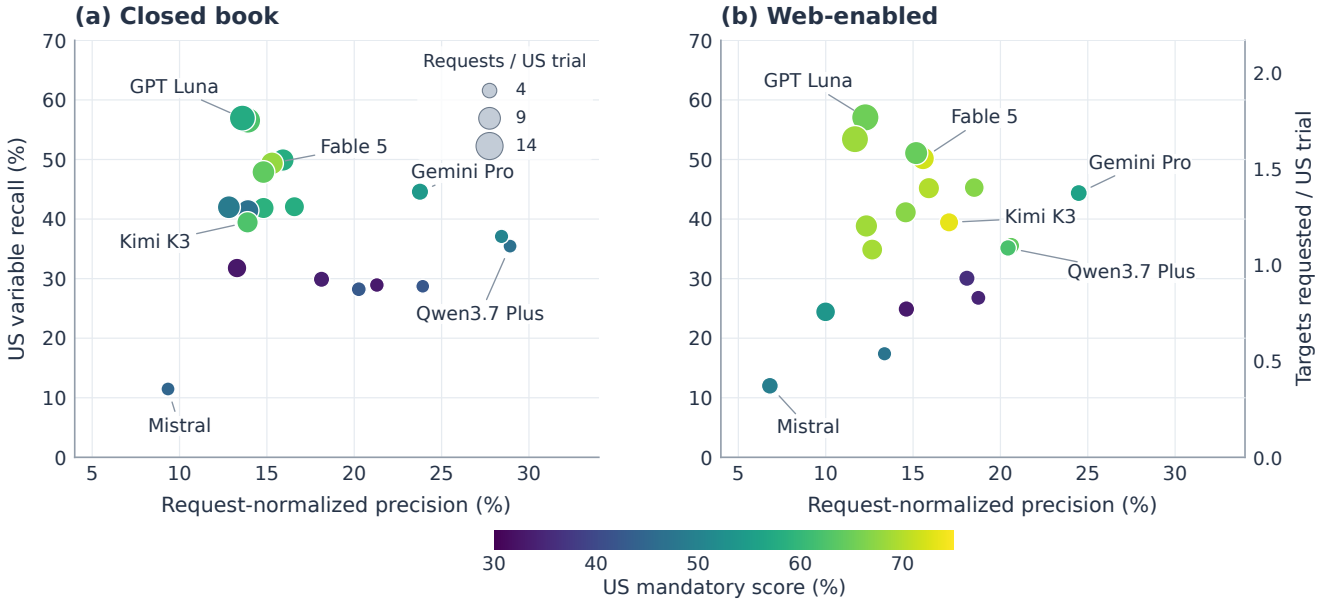


Figure 10. Per-model clarification behavior. Each point represents one model in (a) closed-book or (b) web-enabled evaluation. All metrics cover all US trials, including direct-answer trials. The horizontal axis is request-normalized precision, which is target-variable detections divided by the number of distinct fact requests. The vertical axis is pooled variable recall. Bubble area shows fact requests per US trial and color shows US mandatory score.

Rank	Model	Path	Tool use (%)	Searches (mean)	Visits (mean)	Distinct URLs (mean)	Duration (median, s)
1	Kimi K3	Custom	92.1	5.8	4.4	37.5	377.8
2	DeepSeek V4 Flash	Custom	99.0	5.2	4.7	35.4	171.5
3	Fable 5	Native	88.2	—	—	—	58.8
4	DeepSeek V4 Pro	Custom	96.5	5.0	4.7	33.3	187.0
5	Opus 4.8	Native	82.9	—	—	—	59.8
6	Gemini 3.5 Flash	Native	99.7	—	—	—	68.3
7	GPT-5.6 Sol	Native	100.0	—	—	—	73.2
8	Sonnet 5	Native	88.1	—	—	—	48.0
9	Kimi K2.6	Custom	97.1	8.6	6.6	47.0	300.5
10	Qwen 3.7 Plus*	Custom	87.0	9.3	5.7	43.0	85.7
11	GPT-5.6 Terra	Native	99.9	—	—	—	47.5
12	GPT-5.6 Luna	Native	99.9	—	—	—	50.6
13	Gemini 3.1 Pro	Native	46.6	—	—	—	50.9
14	Qwen 3.6 27B	Custom	89.6	5.3	1.2	31.7	59.1
15	Mistral 3 Large	Custom	48.1	0.9	1.1	7.7	47.3
16	Qwen-SEA-LION 27B	Custom	71.2	4.5	0.7	24.9	649.1
17	Gemma 3 27B	Custom	37.8	0.4	0.0	3.5	22.3
18	Llama 4 Scout†	Custom	86.4	1.1	0.1	9.1	10.6
19	Gemma-SEA-LION 27B	Custom	45.1	2.1	0.0	4.2	19.7

Table 11. Search behavior in the web-search condition. Rows follow descending web-enabled mandatory score, with FS and US combined. Tool use is the percentage of trials with at least one search or page-visit call. Searches, visits, and distinct URLs are means per trial; URLs include those encountered in search results or direct visits. Duration is median wall-clock time from prompt to final answer, including generation and tool calls. Native denotes provider-hosted search; custom denotes our tool-call loop. Dashes indicate unavailable native-provider telemetry. Values summarize available runs, with three seeds per model unless marked.

*Two seeds. †One seed.