Beyond the Single-Expert Ceiling: Localizing LLM-as-Judge Stability on Free-Form Legal Essays

Anonymous ACL submission

Abstract

LLM-as-a-judge systems are typically validated by agreement with human experts, but expert grades can vary, making "human-level" agreement an unstable target. We study Thai bar-exam grading, where three Bar Council-trained examiners and multiple LLM judges score the same long-form legal answers under identical inputs (question, official rubric, gold answer, and student response). Aggregate stability is higher for the LLM panel, but the difference is entirely driven by one rubric-boundary case: the guideline under-specifies how to score an otherwise correct answer that omits a decisive statutory citation. On this item, LLM judges overwhelmingly follow the majority human interpretation and do not reproduce a plausible stricter minority reading. Across 26 models, 22 cluster in or near the majority band, 3 fall in the rubric-silent middle gap, and only 1 (GPT-5.4 Nano) approaches the minority band; *zero LLMs reproduce the minority reading*. We call this the *LLM majority-amplifier effect*: selecting judges to maximize agreement with a human reference panel systematically suppresses legitimate minority readings on rubric-silent items.

1 Introduction

LLM-as-a-judge systems are often evaluated by their agreement with human experts (). In practice, this is operationalized with inter-annotator agreement or correlation between the model's scores and an expert reference on a shared set of responses. This framing implicitly treats the expert reference as a stable target. LLM-as-a-judge has become practical because modern LLMs can follow detailed grading instructions where surface metrics fall short (Zheng et al., 2023; Liu et al., 2023; Gu et al., 2024; Akarajaradwong et al., 2025; Laban et al., 2024; Es et al., 2024). Standard practice selects a judge (or prompt) by maximizing agreement with expert scores, often reporting Cohen's kappa or agreement scores (Thakur et al., 2025; Zheng et al., 2023; Liu et al., 2023; Es et al., 2024). However, when experts disagree, human–LLM agreement can reflect limitations of the target as much as limitations of the judge (Plank, 2022; Pavlick and Kwiatkowski, 2019).

Prior work often treats human-LLM agreement as the ceiling for automated judging, especially when a single expert or panel defines the target (Zheng et al., 2023; Liu et al., 2023; Thakur et al., 2025; Es et al., 2024). Yet this ceiling is unstable when experts disagree about applying the grading guideline to a specific case. This problem is exacerbated in an interpretation-intensive domain like law, where both answering the question and grading the answer require reasoning rather than surface-level matching.

Central thesis. In interpretation-intensive domains such as law, human–LLM-judge alignment must be interpreted relative to human–human alignment, which determines the stability of the expert grading target before any automated judge is evaluated. *Accordingly, we complement aggregate agreement (e.g., panel α) with targeted analyses that make explicit which plausible expert readings an LLM judge reproduces (and which it misses).*

We study this in a Thai bar exam grading setting, where graders apply an official regulation and gold answer, and both answering and grading require legal reasoning rather than surface-form matching. Three Bar Council-trained human examiners and three anchor LLM judges independently scored the same 15 answers using identical inputs: the question, the official 2014 Bar Council grading regulation, the official gold answer curated by the judge consortium, and the participant answer.

Results indicate that at the aggregate level, the LLM panel shows higher inter-rater stability (Artstein and Poesio, 2008) than the human panel. However, this difference is not uniform across questions. For two of the three questions, the human and LLM

panels achieve comparable pooled stability. The gap is driven by a single question in which the rubric does not specify how to grade an otherwise correct answer that omits a decisive statutory citation. In this case, the LLM judges do not replace human judgment; they reproduce a single human reading of the rubric, aligning with the majority rather than a stricter minority interpretation. This prevents the aggregate stability gap from being read as a general claim that LLM judges are more stable than human experts.

Extending the LLM panel to 26 models reveals an asymmetric structural finding (Section 4): classifying each judge by its contested-cell mean, 22 of 26 LLMs land in or near B/C's reading band, 3 sit in the regulation-silent middle gap, and only 1 (GPT-5.4 Nano) approaches A's band without consistently scoring within it. *Zero LLMs in the 26-judge panel reproduce A's reading.* The asymmetry holds across model size, vendor, and price.

Our contributions are: (i) **Finding.** We identify a population-level bias, the *LLM majority-amplifier effect*: on rubric-silent (contested) items, agreement-maximizing judge selection concentrates on the majority human reading; across 26 LLM judges, *zero* reproduce the stricter minority reading. (ii) **Method.** We propose augmenting LLM-judge validation with contested-cell analysis: compute each judge's mean score on contested rubric cells, cluster judges by direction, and report these clusters alongside panel α to audit rubric-reading coverage. (iii) **Reproducibility support.** We release an identical-inputs six-rater protocol with a dual-role design (LLMs as examinees and judges), plus a replicable Thai bar-exam corpus with official questions, gold answers, and multi-rater human/LLM scores.

2 Study Design

We compare human and LLM judgments under an identical-inputs protocol: Bar Council-trained examiners and anchor LLM judges independently score the same Thai bar-exam-style questions using the official grading regulation and gold answers, with all raters blinded to each other's scores.

2.1 Stage 1: Grading instrumentation

We use three Thai bar-style Civil and Commercial Code essay questions that differ in the kind of legal distinction they require. (See Appendix A.)

Q1 asks whether an ordinary partnership and its partners are externally bound when one partner enters a purchase contract despite internal objection; the decisive distinction is between internal limits on partner authority and external liability to a third party.

Q2 asks whether a company is bound by contracts made around a registered but unpublished change of directors; the decisive issue is the effect of CCC Section 1023/1 on good-faith third parties, making omission of that citation especially consequential.

Q3 asks whether a shareholder or creditor may annul procedurally defective shareholder-meeting resolutions or fabricated meeting minutes; the answer turns on quorum, standing, and the one-month limitation period.

This makes Q2 contestable for grading, because an answer can reach the correct conclusion while omitting the statutory citation that explains why the conclusion follows.

Raters apply the official Bar Council regulation and gold answer (Appendix B) and score with an 8-tier ordinal rubric (Table 1): choose a tier, then a half-integer within its range. Each question and its corresponding gold answer were drafted by each examiner based on their expertise before collecting any responses to minimize question leakage from the internet.

	Reasoning quality				
Final answer	Unusable	Some	Acceptable	Good	Very good
matches gold	T4: 0–1	T5: 2–5	T6: 5–6	T7: 7–8	T8: 9–10
differs	T1: 0	T2: 1–2	–	T3: 2–4	–

Table 1: Official Thai Bar Council 8-tier rubric: grader picks a tier, then a half-integer score in its range. Ranges overlap and break monotonicity (T4 caps at 1, below T3 at 2–4); the regulation-silent T4-vs-T7 boundary drives the disagreement in Section 3.

2.2 Stage 2: Exam sitting

We conducted a bar-qualification-style mock examination in early 2026 with 50 examinees (**42 humans**; **8 LLMs**).

Human examinees are Thai law students and legal practitioners preparing for the bar examination under informed consent; they sat a proctored 1.5-hour session with no electronic aids, writing longhand.

LLM examinees are closed-frontier models inserted into the roster as human-indistinguishable sequence IDs. Each was queried once per question via provider API at temperature 0 with a zero-shot prompt (no retrieval, tools, or in-context examples) to produce Thai legal prose in a handwritten-answer register. All responses were transcribed

into a common e-document (Google Docs) format, pseudonymized, and graded by examiners unaware of the LLM examinees.

2.3 Stage 3: Grading and LLM-judge scoring

To measure judge stability on identical items, we select five examinees[1] and cross-grade their answers on all three questions ($5 \times 3 = 15$ cells[2]; see Table 2 in Section 3 for the full score matrix). For each cell, the two examiners who did not primary-grade that question independently assign a rubric tier and score given the same inputs (question, regulation, gold answer, and response).

We score the same 15 cells with three LLM judges (Claude 4.6 Opus, Gemini 3.1 Pro, GPT-5.4 Pro) at temperature 0, producing JSON outputs (tier, half-integer score, justification) with no tools, retrieval, or few-shot examples (Appendix D). Raters are blinded to one another's scores. Section 3 therefore compares scoring functions under identical information. We avoid pre-calibrating examiners to retain naturally occurring interpretive variance (which any single-expert "ground truth" would also embody).

3 Results

Table 2 reports the full 15-cell cross-graded score matrix. On the 15 cells, panel-level Krippendorff α is 0.36 for the human panel $\{A, B, C\}$ and 0.77 for the anchor LLM panel $\{\text{Cla}, \text{Gem}, \text{GPT}\}$.

To clarify the discrepancy in human agreement, we further conduct error analysis and identify the explanation for the human agreement gap.

The gap is localized to one question. Table 3 decomposes panel stability by question. On Q1 and Q3, the two panels are equivalently stable ($\alpha = 0.75$ each on the 10-cell pool). The full-panel gap of $+0.41$ is driven by the 5 Q2 cells, three of which are the cells where the participant reaches the correct final answer but omits a decisive statutory citation (Section $1023/1$). On those contested cells, the human panel falls to $\alpha = -0.34$ while the LLM panel holds at $\alpha = 0.68$.

The Q2 contested cells. Q2 concerns company liability after a registered but unpublished change of directors. The decisive issue is whether the company may invoke the change of authority against a good-faith third party, for which CCC Section 1023/1 is central. To understand the Q2 contested cells, we inspect the judges' scores and comments on this question. Across these cells, all judges agree on the final-answer correctness rubric, which is checked against the gold answer, but no consensus is reached on the reasoning-quality rubric. Specifically, on pid 21/Q2, Judge A scored the response 1 (*unusable*), while B and C scored it 7 and 8 (*good*); the three LLM judges gave scores of 5.5/6/5.5 (*acceptable*). Judge A's comment indicates that the low score was due to the answer's lack of a decisive citation.

Although the other judges, including the LLMs, also appear to be aware of this missing citation, their interpretations of penalization intensity vary.

pid	Q	A	B	C	Cla	Gem	GPT
4	1	9.0	9.0	9.0	9.0	10.0	9.0
4	2	3.0	4.0	6.0	4.0	4.0	5.0
4	3	6.0	6.0	5.0	4.0	5.5	5.5
6	1	5.0	7.0	5.0	5.5	6.0	4.5
6	2	3.0	4.0	5.0	4.0	4.0	5.0
6	3	8.0	7.0	6.0	7.0	8.0	7.5
10	1	6.0	6.0	7.0	7.0	4.0	6.0
10	2	1.0	7.0	7.0	7.0	5.5	7.0
10	3	8.0	7.0	7.0	7.5	6.0	6.0
21	1	5.0	6.0	6.0	7.5	9.0	8.0
21	2	1.0	7.0	8.0	5.5	6.0	5.5
21	3	8.0	8.0	8.0	7.0	8.0	7.5
27	1	0.0	4.0	3.0	4.0	4.0	5.5
27	2	2.0	6.0	6.5	5.5	6.0	5.5
27	3	5.0	7.0	6.0	4.5	6.0	5.5
mean		4.67	6.33	6.30	5.93	6.13	6.20
SD		2.90	1.45	1.49	1.61	1.88	1.29
α			**0.36**			**0.77**	

Table 2: All 15 cross-graded cells scored by the three human examiners (A, B, C) and the three anchor LLM judges (Claude 4.6 Opus, Gemini 3.1 Pro, GPT-5.4 Pro) under the same four inputs. Contested Q2 cells (pids 10/21/27) drive the human collapse: A scores 1, 1, 2 where B and C score 6–8.

Slice	N	Human α	LLM α	Gap
Q1	5	0.78	0.79	$+0.01$
Q2	5	-0.34	0.68	$\mathbf{+1.02}$
Q3	5	0.55	0.67	$+0.12$
Q1 and Q3 pooled	10	**0.75**	**0.75**	0.00
All 15 cells	15	0.36	0.77	$+0.41$

Table 3: Judge's Krippendorff α decomposed by question. Q1 and Q3 give identical human and LLM reliability. The cross-panel gap concentrates on Q2 cells.

4 Judge Population Analysis

The contested Q2 cells show that expert interpretations can vary when the rubric underspecifies how

[1] Participant IDs 4, 6, 10, 21, 27; one is Claude 4.6 Opus, still unidentified to cross-graders

[2] We coined the term "cell" as a single grade provided by human judges during cross-grader phase.

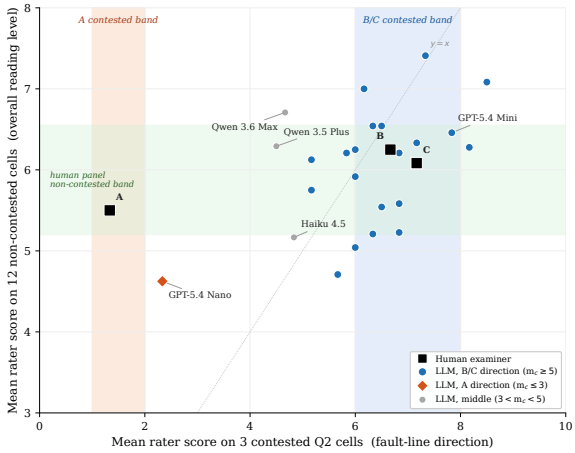

Figure 1: **LLM judge clustering on contested and non-contested cells.** Each marker represents one judge: black squares are human examiners and circles/diamonds are LLM judges. The x-axis shows the mean score on the 3 contested Q2 cells (m_c), while the y-axis shows the mean score on the 12 non-contested cells (m_{nc}).

to penalize a correct answer that omits a decisive citation. This creates an unstable ceiling for LLM-judge calibration, since the target depends on which expert reading is treated as reference. We extend the three-LLM anchor sub-panel to 26 LLMs spanning closed frontier large, closed frontier small, flagship open-weight, and small open-weight models (see Appendix E). Each judge receives the same four inputs as the three anchors (Section 2.3). For each rater, we compute the mean score on the 3 contested Q2 cells (m_c) and the mean score on the 12 non-contested cells (m_{nc}); Figure 1 plots both. Per-judge m_c, m_{nc}, and pairwise Pearson with each human cross-grader on the full 15-cell vector are in Appendix E, Table 6.

No LLM aligns with Judge A on contested Q2 cells. On the non-contested cells, all raters fall within a narrow score range ($m_{nc} \in [4.5, 7.0]$), showing broad agreement between human and LLM judges when the rubric is not ambiguous. The disagreement appears on the contested Q2 cells. There, the human judges split into two readings: Judge A assigns scores in the lower band, while Judges B and C assign scores in the upper band. The LLM population does not reproduce this split as almost all LLM judges align with Judges B and C. Only GPT-5.4 Nano comes closest to Judge A, and even it does not consistently fall within Judge A's band. Thus, *zero LLMs in our 26-judge panel reproduce Judge A's reading.*

The LLM majority-amplifier effect. Most LLM judges align with Judges B/C on the contested cells, indicating convergence on the majority human reading rather than coverage of the full expert-interpretation space. This does not make Judge A's stricter reading invalid; it shows that a plausible minority reading can be suppressed when LLM judges are evaluated by agreement with the majority. Prior work shows that LMs can exhibit strong output homogeneity on open-ended tasks, and that LM judges and reward models become less calibrated when human annotators highly disagree (Jiang et al., 2026). Our results raise a related concern for grading in interpretation-intensive domains such as law: **LLM judges may appear stable not because they cover all experienced human judgments, but because they converge on one dominant reading while missing others.**

5 Conclusion

This paper asks how to interpret human–LLM judge alignment when expert grading is itself unstable. We find that human and LLM panels are similarly stable when the rubric and gold answer constrain scoring, but diverge on an underspecified Q2 boundary. There, humans split between a *stricter minority* and a *more lenient majority* reading, while LLMs overwhelmingly follow the latter.

Synthesis. This localization supports our central thesis: human–LLM agreement cannot be interpreted without first examining human–human agreement. Once contested grading is separated, aggregate stability no longer shows that LLM judges are simply *"more stable."* It shows which expert reading they reproduce. In our setting, LLM stability reflects convergence on one plausible reading, not coverage of the full space of judgments.

Implication. While majority alignment may be acceptable for scalable scoring, it can be undesirable to legal pedagogy and high-stakes evaluation, where a lenient majority reading can flatten distinctions captured by a stricter minority reading. LLM-as-a-judge evaluations should therefore report local agreement profiles alongside aggregate agreement: where experts agree, where they diverge, and which readings LLMs reproduce or miss.

Future work. Defining the quality ceiling requires protocols for adjudicating expert readings while preserving legitimate disagreement. We support this direction by releasing exam questions, gold answers, rater scores, and related resources.

Limitations

Small cross-graded subset. The core stability comparison uses 15 fully cross-graded answers with six certified raters per cell. Scaling cross-grading at this rater density for free-form bar essays is institution-bound: deploying multiple Bar Council-certified examiners across a larger shared subset is feasible only within an official Bar Council session, which requires Council and Ministry of Justice coordination beyond the scope of an academic study. The effect size is substantial, and the mixed-panel sweep corroborates the same pattern across all 20 three-rater panels, but specific α values will move under replication; a Council-supported replication at ≥ 30 cross-graded cells would tighten confidence intervals without overturning the direction.

Panel size and outlier-driven variance. The human panel is three examiners, and the localization depends on how we characterize Judge A's stricter reading of the tier-4-vs-tier-7 boundary. In a different three-examiner panel, that boundary might be resolved by a different rater, and the observed human α might rise toward $r(\mathrm{B}, \mathrm{C}) = 0.75$ or fall further. The structural claim (one regulation-silent boundary drives the cross-panel gap) is panel-independent, but the specific 0.36 value and the identity of the outlier axis should be read as one panel's result.

Doctrinal and linguistic scope. Three essay questions from one doctrinal cluster (Thai partnership and corporate law); a real Thai bar session has ten questions across multiple areas. Whether other legal systems have analogous silent zones where expert panels fracture is an open empirical question; the structural mechanism (panels agree where the rubric is specified, diverge where it is silent) is not Thai-specific, but the numeric gap may differ in English or common-law systems.

Ethics Statement

All 42 human participants consented and were compensated at a fair-market rate. The three examiners consented to the release of their comments and scores under the Thai Bar Council's customary grading-fee structure and were paid the Council's standard examiner rate. Language-model queries were made under each provider's terms of service. Released answer texts and grading comments have been stripped of personal identifiers; only the sequence-number participant identifier remains. The dataset is intended for research on legal-reasoning evaluation methodology and should not be used for individual assessment or certification decisions.

References

Pawitsapak Akarajaradwong, Pirat Pothavorn, Chompakorn Chaksangchaichot, Panuthep Tasawong, Thitiwat Nopparatbundit, Keerakiat Pratai, and Sarana Nutanong. 2025. NitiBench: Benchmarking LLM frameworks on Thai legal question answering capabilities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34304–34327, Suzhou, China. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on LLM-as-a-judge. *arXiv preprint*. ArXiv:2411.15594.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2026. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Philippe Laban, Alexander Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context LLMs and RAG systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9885–9903, Miami, Florida, USA. Association for Computational Linguistics.

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

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. *Proceedings of EMNLP*.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

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

A Full Question Details and Gold Answers

The three essay questions were drawn from the Thai Civil and Commercial Code (CCC); each carried 10 points. Below we provide a faithful English translation of (i) each question's fact pattern and call and (ii) the corresponding gold answer (*thong kham-top*, "answer-flag") issued by the examiners as the official model answer. The original Thai texts are included in the data release.

A.1 Q1: Ordinary Partnership Liability

Decisive provisions: CCC Sections 1012, 1025, 1033, 1050.

Question. The Three Friends Grilled Chicken Partnership (*hang-hun-suan saam-sa-hai kha kai-yang*) was formed with the purpose of trading grilled chicken. Its partners are Mr. Wanmai, Mr. Duenmai, and Mr. Pimai, each contributing 100,000 baht as capital. In the course of the partnership's business, Mr. Wanmai wished to purchase chickens from Charoen Food Co., Ltd. for grilling and sale, but Mr. Duenmai and Mr. Pimai objected. Mr. Wanmai did not heed the other partners' objections and, in violation of them, entered into a contract to purchase fresh chicken from Charoen Food Co., Ltd. for 100,000 baht. When the delivery date arrived, Charoen Food Co., Ltd. delivered the chicken, but Mr. Duenmai and Mr. Pimai refused to accept it on the ground that Mr. Wanmai had acted without authority. From whom, if anyone, may Charoen Food Co., Ltd. demand liability, and on what ground?

Reference answer. The issue is from whom Charoen Food Co., Ltd. may demand liability. The relevant provisions of the Civil and Commercial Code are as follows. Section 1012 provides that a partnership or company contract is a contract whereby two or more persons agree to unite for a common undertaking with a view to sharing the profits that may be derived from it. Section 1025 provides that an ordinary partnership is a partnership in which all partners are jointly and severally liable without limit for all debts of the partnership. Section 1026 provides that every partner must contribute something to the partnership, whether money, other property, or labour. Section 1033 provides that if the partners have not agreed otherwise on the manner of management, every partner may manage the partnership, but a partner may not enter into any contract to which another partner has objected; in such case every partner is deemed a managing partner. Section 1050 provides that any act done by a partner in the ordinary course of the partnership's business binds all partners, who are jointly and severally liable without limit for the debts so created.

Analysis. Because the Three Friends Grilled Chicken Partnership was formed with the purpose of trading grilled chicken, with Mr. Wanmai, Mr. Duenmai, and Mr. Pimai as partners each contributing 100,000 baht, the case is one in which two or more persons have agreed to unite by contributing 100,000 baht each under Section 1026 to conduct the business of trading grilled chicken with a view to sharing profits. The arrangement is the formation of an ordinary partnership under Sections 1012 and 1025, in which all partners bear unlimited joint and several liability for all partnership debts. Because the facts do not show any agreement on the manner of management, all three are deemed managing partners and each may manage the partnership under Section 1033.

As for Mr. Wanmai's entry into the purchase contract with Charoen Food Co., Ltd. over the objections of Mr. Duenmai and Mr. Pimai: although the final clause of Section 1033 paragraph one provides that a partner may not enter into a contract to which another partner has objected, that provision governs the *internal* relation among partners, not the *external* relation between partners and third parties (governed by Section 1050). For the external relation, the sole consideration is whether the transaction was in the ordinary course of the partnership's business.

Since the partnership was formed to trade grilled chicken, the purchase of fresh chicken for 100,000 baht is within the ordinary course of its business. All partners (Mr. Wanmai, Mr. Duenmai, and Mr. Pimai) are therefore bound by this purchase and are jointly and severally liable without limit for the debt Mr. Wanmai created, each for the full 100,000 baht. However, because the Three Friends Grilled Chicken Partnership is not registered as a juristic person separate from its partners, Charoen Food Co., Ltd. cannot demand liability from the partnership itself.

Conclusion. Charoen Food Co., Ltd. may demand that Mr. Wanmai, Mr. Duenmai, and Mr. Pimai be jointly and severally liable for 100,000 baht.

A.2 Q2: Company Director Authority

Decisive provisions: CCC Sections 1022, 1023, 1023/1.

Question. Jipatha Electricals Co., Ltd. was formed with the purpose of dealing in all types of electrical appliances and equipment. Its directors were Mr. Somchai, Mr. Somchoke, and Mr. Somchaiya. On 26 February 2012 a shareholders' meeting was held and the shareholders resolved to remove Mr. Somchai, Mr. Somchoke, and Mr. Somchaiya from office and to appoint Mr. Praneet, Mr. Pranom, and Mr. Pranit as replacement directors. The change of directors was duly registered with the Ministry of Commerce on 28 February 2012, but the Ministry had not yet published the change in the Royal Gazette.

(a) If on 1 March 2012 Mr. Somchai and Mr. Somchoke entered into a contract in the name of Jipatha Electricals Co., Ltd. to purchase 20 refrigerators from Arctic Cooling Co., Ltd., received the refrigerators from Arctic Cooling Co., Ltd., and took them for themselves without delivering them to Jipatha Electricals Co., Ltd., may Arctic Cooling Co., Ltd. demand that Jipatha Electricals Co., Ltd. pay the price of the refrigerators? On what ground?

(b) If on 1 March 2012 Mr. Praneet and Mr. Pranom entered into a contract to purchase 20 refrigerators from AI-Freeze Co., Ltd., and when the delivery date arrived AI-Freeze Co., Ltd. delivered the refrigerators, but Jipatha Electricals Co., Ltd. refused to accept delivery on the ground that at the time of the contract Mr. Praneet and Mr. Pranom were not yet lawful directors because the registration had not been published in the Royal Gazette, may AI-Freeze Co., Ltd. demand damages from Jipatha Electricals Co., Ltd.? On what ground?

Note: the facts in (a) and (b) are unrelated.

Reference answer. The change of directors from Mr. Somchai and Mr. Somchoke to Mr. Praneet and Mr. Pranom by resolution of the shareholders' general meeting, duly registered, is a valid change of the persons authorised to act for Jipatha Electricals Co., Ltd., with the consequence that Mr. Somchai and Mr. Somchoke no longer have authority to act on the company's behalf. Whether the company may assert this change against third parties in (a) and (b) is analysed as follows.

(a) Although in general Jipatha Electricals Co., Ltd. may invoke the registered change of directors against third parties under CCC Section 1023 paragraph one (by arguing that the 1 March 2012 contract signed by the former directors Somchai and Somchoke does not bind the company because they were no longer authorised), Section 1023/1 establishes an exception: a company may not raise Section 1023 as a defence against a third party acting in good faith in order to escape liability on the ground that the directors lacked authority. Here, there is no indication that Arctic Cooling Co., Ltd. knew that Somchai and Somchoke were no longer directors, nor does it appear that the change of directors had been published in the Royal Gazette so as to constitute constructive notice under CCC Section 1022. Arctic Cooling Co., Ltd. is therefore a third party acting in good faith. Jipatha Electricals Co., Ltd. cannot raise the lack of authority of Somchai and Somchoke against Arctic Cooling Co., Ltd. under CCC Section 1023/1. Arctic Cooling Co., Ltd. may demand the refrigerator price from Jipatha Electricals Co., Ltd. (5 points).

(b) The contract entered into by the new directors Praneet and Pranom on 1 March 2012 binds Jipatha Electricals Co., Ltd. because it was made by directors vested with authority by a shareholders' general-meeting resolution that had already been followed by a registered change of directors. Publication in the Royal Gazette under Section 1022 serves only to constitute notice to all third parties; it is not a condition for the effectiveness of the change of directors to Praneet and Pranom. AI-Freeze Co., Ltd. is therefore entitled to demand damages from Jipatha Electricals Co., Ltd. for breach of the contract that Praneet and Pranom concluded on the company's behalf (5 points).

Conclusion. (a) Arctic Cooling Co., Ltd. may demand the refrigerator price from Jipatha Electri-

cals Co., Ltd. (b) AI-Freeze Co., Ltd. may demand damages from Jipatha Electricals Co., Ltd.

A.3 Q3: Shareholder-Meeting Annulment

Decisive provisions: CCC Sections 1178, 1193, 1195.

Question. Sudthawin Tour Co., Ltd. was formed to provide overseas tour services. Its registered capital is 200,000 baht, divided into 100 shares of 2,000 baht each, held by 100 shareholders of 1 share each. Its directors are Mr. Ake, Mr. Toh, Mr. Tri, and Mr. Jattawa. The board resolved to convene the 2/2557 ordinary general meeting of shareholders, which was held on 4 April 2014.

At the meeting, 22 shareholders attended. The chair opened the meeting and the meeting considered the payment of an annual dividend, the company having a reasonable profit. On this agenda item, 12 shareholders abstained, 6 voted to approve the dividend, and 4 voted against. The chair declared that the meeting had approved the annual-dividend payment.

On 2 May 2014, Mr. Khao, a shareholder, considered that the 2/2557 general meeting had been convened and had passed a resolution in violation of law, and wished to sue to annul the dividend-approval resolution. He also wished to sue to annul the minutes of meeting 6/2556 dated 6 December 2013, which had approved the appointment of Mr. Tawan as auditor, on the ground that no meeting had actually been held, since Mr. Toh had fabricated those minutes. The other shareholders disagreed and objected to Mr. Khao's proposed action, but Mr. Khao insisted on proceeding alone. Determine:

(a) Can Mr. Khao petition the court to annul the 2/2557 general-meeting resolution and the 6/2556 meeting minutes? Give reasons.

(b) If Mr. Somchai, a major creditor of the company, wishes to petition the court to annul the 2/2557 general-meeting resolution and the 6/2556 meeting minutes, may he do so, and how?

Reference answer. *(a)* The issue is whether Mr. Khao may petition the court to annul the 2/2557 general-meeting resolution and the 6/2556 meeting minutes.

As to the 2/2557 general-meeting resolution, CCC Section 1178 provides that a general meeting must be attended by shareholders holding in aggregate at least one quarter of the company's capital in order to pass any resolution. On the facts, the company's registered capital is 200,000 baht divided into 100 shares of 2,000 baht, and only 22 shareholders attended, each holding 1 share. Only 22 shares attended, representing 44,000 baht, which is less than one quarter of the company's capital. The meeting therefore lacked a quorum and could not pass a resolution on the dividend; the dividend-approval resolution is unlawful, notwithstanding that a majority of those voting approved it.

As to the right to sue for annulment, CCC Section 1195 provides that, where a general meeting has been convened or has passed a resolution in violation of a provision of the law on companies or of the company's articles, a director or any shareholder may petition the court to annul the resolution, but the petition must be filed within one month from the date of the resolution. Mr. Khao is a shareholder and may exercise the right to petition alone; the disagreement of other shareholders does not deprive him of that right. The 2/2557 general-meeting resolution was passed on 4 April 2014 and his intended filing on 2 May 2014 is within the one-month period.

As to the 6/2556 meeting minutes dated 6 December 2013, which approved Mr. Tawan as auditor: Mr. Khao wishes to sue for annulment on the ground that no meeting actually took place and the minutes were fabricated. This is not a case of annulling a general-meeting resolution that violated the law on companies or the company's articles, to which the one-month limitation in CCC Section 1195 applies. Even after the one-month period from the resolution date has elapsed, Mr. Khao may still sue for annulment.

Conclusion. Mr. Khao may petition the court to annul the 2/2557 general-meeting resolution because the meeting lacked a quorum, Mr. Khao is a shareholder with standing to act alone, and he is within the one-month period. He may also petition for annulment of the 6/2556 meeting minutes, even after the one-month period, because no meeting actually took place; this is not the annulment of an unlawful general-meeting resolution under Section 1195.

(b) The issue is whether Mr. Somchai, a major creditor, may petition the court to annul the 2/2557 general-meeting resolution and the 6/2556 meeting minutes.

CCC Section 1195 provides that a director or any shareholder has the right to petition the court to annul a general-meeting resolution that was passed in violation of the law on companies or of the com-

pany's articles. On the facts, Mr. Somchai is a major creditor of the company but does not appear to be a director or a shareholder. He therefore has no right to petition for annulment of the 2/2557 general-meeting resolution. As to the 6/2556 meeting minutes, even though this is not strictly a case falling under CCC Section 1195, Mr. Somchai is merely a creditor, and the law does not confer on creditors a power to sue directly in respect of a company's resolutions or meeting minutes; he therefore cannot petition the court.

Conclusion. Mr. Somchai has no right to petition the court to annul the 2/2557 general-meeting resolution or the 6/2556 meeting minutes, because he is merely a creditor, not a director or shareholder, and no law confers on a creditor standing to petition for such annulment.

B Thai Bar Council Grading Regulation

The official grading regulation of the Office of Legal Education Training of the Thai Bar Council is the criterion our three examiners applied. It has been used by Thai bar examiners as the standard scoring instrument for free-form essay questions on the 10-point scale. Below we give a faithful English translation of the part of the regulation that establishes the scoring criteria; preamble, authorising-citation boilerplate, and enacting matter are paraphrased.

Background. The Executive Committee of the Office of Legal Education Training of the Thai Bar Council issued the regulation in order to standardise the exercise of discretion by grading-committee members on essay examinations at the barrister-at-law level.

Clause 1: Scoring criteria (faithful translation). The grading shall consider the final answer and the supporting reasons together, on a per-question basis, as follows:

(1) Final answer does not match the gold answer, and reasoning unusable: score 0.

(2) Final answer does not match the gold answer, but reasoning somewhat acceptable: score 1–2.

(3) Final answer does not match the gold answer, but reasoning good: score 2–4.

(4) Final answer matches the gold answer, but reasoning unusable: score 0–1.

(5) Final answer matches the gold answer, and reasoning somewhat acceptable: score 2–5.

(6) Final answer matches the gold answer, and reasoning acceptable: score 5–6.

(7) Final answer matches the gold answer, and reasoning good: score 7–8.

(8) Final answer matches the gold answer, and reasoning very good: score 9–10.

If a question contains multiple issues, the grader shall fix the score for each issue first and then apply the criteria above to each issue.

Clause 2. This regulation takes effect from the date of announcement.

Structure of the scale. The scale is 8-tier ordinal with two correctness classes crossed against four reasoning-quality levels, with the overlapping score ranges shown above. The regulation does *not* enforce a clean correctness split: tier 3 (incorrect final answer, reasoning good) scores 2–4 and tier 4 (correct final answer, reasoning unusable) scores 0–1, so a correct final answer with unusable reasoning scores strictly below an incorrect final answer with good reasoning. The numeric score alone therefore does not recover the tier: a score of 1 could be tier 2 (incorrect, somewhat acceptable reasoning) or tier 4 (correct, unusable reasoning); a score of 2 could be tier 2, 3, or 5; a score of 5 could be tier 5 or 6. Section 3 reports both tier and score for each LLM judge; the human graders' tier assignments are inferred from their Thai-language comments. A conventional pass is 15 out of 30 on a three-question session.

What the regulation does not prescribe. The regulation fixes the tier structure and the tier-to-score map, but does not prescribe per-sub-decision weights, question-specific decomposition trees, or natural-language anchors attached to each tier beyond the short tier descriptors above (e.g. what distinguishes "somewhat acceptable" from "acceptable" reasoning). Examiners fill this additional calibration from their training and prior bar-grading experience. The floor-level stability the three examiners reach under this instrument (Section 3) is therefore a measurement on this shared-experience-plus-regulation calibration, not on the axes in isolation.

C Grader Analysis: Extended

This appendix supplements Section 3 with the by-tribe correlation summary and the complete panel-α sweep. The full 15-cell raw score matrix used by the stability analysis is Table 2 in Section 3. Judges A, B, C are the three human examiners (primary on Q1, Q2, Q3 respectively); Claude 4.6 Opus, Gemini 3.1 Pro, GPT-5.4 Pro are the three anchor LLM judges at temperature 0 given identical inputs.

C.1 By-tribe correlation summary

Judge	$\bar{r}$ cross-tribe	$\bar{r}$ own-tribe
Humans (own = other 2 humans)		
A	0.57	0.52
B	0.70	0.68
C	0.61	0.59
LLMs (own = other 2 LLMs)		
Claude	0.68	0.79
Gemini	0.64	0.79
GPT-5.4 Pro	0.57	0.84

Table 4: Each judge's mean pairwise Pearson with the opposite tribe vs. with their own tribe, on the 15 cross-graded answers. B, C, Claude, and Gemini all sit inside a single cross-tribe cluster with mutual r in the 0.55–0.76 range; Judge A's mean cross-tribe Pearson (0.57) sits below the mean within-tribe value of every other grader, reflecting A's stricter interpretive reading on the tier-4-vs-tier-7 boundary.

C.2 All 20 three-rater panel alphas

Table 5 lists Krippendorff α for every three-rater panel drawn from $\{A, B, C, \text{Cla}, \text{Gem}, \text{GPT}\}$, sorted within each column by α. All ten panels that include A sit at $\alpha \leq 0.49$; all ten that exclude A sit at $\alpha \geq 0.63$.

With A ($n = 10$)		**Without A** ($n = 10$)	
Panel	α	**Panel**	α
A, C, GPT	0.30	C, Gem, GPT	0.63
A, B, GPT	0.35	B, C, GPT	0.65
A, B, C	0.36	C, Cla, Gem	0.65
A, C, Gem	0.38	B, C, Gem	0.67
A, C, Cla	0.39	C, Cla, GPT	0.69
A, Cla, GPT	0.42	B, Cla, GPT	0.71
A, B, Cla	0.44	B, Gem, GPT	0.72
A, Gem, GPT	0.44	B, C, Cla	**0.72**
A, B, Gem	0.46	B, Cla, Gem	0.73
A, Cla, Gem	0.49	Cla, Gem, GPT	**0.77**
max	**0.49**	**min**	**0.63**

Table 5: Panel α for all 20 three-rater combinations. Removing A from every panel it appears in raises α by at least 0.14; no A-including panel reaches $\alpha \geq 0.50$.

D LLM-Judge Prompt

This appendix reproduces the verbatim prompt used for each of the three LLM judges (Claude 4.6 Opus, Gemini 3.1 Pro, GPT-5.4 Pro) in Section 3. The prompt is identical across all three judges and across all 150 primary-graded answers; only the question text, gold answer, and participant answer change per item. Sampling: temperature 0, fixed seed where available, single forward pass, no tool use, no retrieval, no few-shot examples, no chain-of-thought scaffolding.

D.1 System prompt (English translation)

```
You are an expert grader of Thai bar
examination free-form essay answers. You
will apply the official Thai Bar Council
grading criteria (reproduced below) to
the given answer. The regulation is the
same instrument used by trained Thai bar
examiners.

You have three inputs: (1) the question
text, (2) the official gold answer
(thong kham top), and (3) the student
answer to be graded.

You must produce:

• A tier number (1–8) from the regulation
  that best fits the student answer.
• A numeric score within that tier's
  score range (half-integers allowed).
• A one-paragraph justification
  (Thai) citing the decisive Civil
  and Commercial Code sections and
  explaining why the answer falls into
  the chosen tier.

Return a JSON object with exactly
three fields: {"tier": integer 1–8,
"score": number 0–10, "justification":
one-paragraph Thai text}. The numeric
score must fall within the tier's range.
Do not include any prose outside the
JSON.

Thai Bar Council Grading Regulation,
verbatim Thai followed by faithful
English translation (see Appendix B for
the full text).
```

D.2 User message schema

```
QUESTION: {question text}
OFFICIAL GOLD ANSWER: {primary grader's
gold answer}
STUDENT ANSWER: {participant answer}
```

D.3 Expected JSON output

```
{ "tier": 7, "score": 7.5,
"justification": "[one-paragraph
Thai legal justification citing Code
sections]" }
```

The gold answer is the primary grader's own *thong kham top* document (also received by the human cross-graders in the Section 2.3 cross-grading).

Neither the primary grader's numerical score nor any other grader's score is included in the judge's inputs. The tier field disambiguates score values that are feasible under multiple tiers (e.g. a score of 1 is feasible under tier 2 or tier 4); without the tier field the judge's class assignment would not be recoverable.

E LLM Judge Panel and Cluster Numbers

The 26-LLM judge panel used in Section 4 spans four model classes:

Closed frontier, large:

1. Opus 4.6/4.7
2. GPT-5.4 Pro
3. GPT-5.5 Pro
4. Gemini 3.1 Pro

Closed frontier, small-to-medium:

1. Sonnet 4.6
2. Haiku 4.5
3. GPT-5.4 Mini/Nano
4. Gemini 3 Flash
5. Gemini 3.1 Flash Lite
6. Grok 4.1 Fast
7. Qwen 3.6 Max

Flagship, large open-weight:

1. GLM 5/5.1
2. DeepSeek v4 Pro
3. Kimi K2.6

Small open-weight:

1. Qwen 3.6 Plus/35B
2. Gemma 4 26B/31B
3. DeepSeek v4 Flash
4. Qwen 3.6 Flash

Table 6 reports each judge's contested-cell mean (m_c), non-contested-cell mean (m_{nc}), and pairwise Pearson correlation with each of the three human cross-graders on the full 15-cell vector. Rows are grouped by contested-cell direction and sorted within each group by m_c. The three human cross-graders are listed at the bottom for reference.

F Robustness

This appendix supplements Section 3 with the bootstrap CIs, alternative IRR coefficients, leave-one-out sweeps, the determinism probe, and the two LLM-disadvantaged input ablations cited in the body.

F.1 Bootstrap confidence intervals

We bootstrap-resample the 15 cells with replacement 5,000 times and recompute panel α on each resample.

F.2 Pairwise Cohen κ_w and Spearman ρ

Pair-by-pair coefficients on the 15 cross-graded answers. Cohen κ_w uses integer-rounded scores on the $\{0, 1, \ldots, 10\}$ ordinal scale.

F.3 Leave-one-out stability

Two sweeps check that the contrast is not carried by a single item or a single rater.

The A-less pair (B, C) reaches $\alpha = 0.76$, inside the LLM range (0.73–0.80). Every two-LLM pair lies in $[0.73, 0.80]$ regardless of which LLM is dropped, ruling out the inversion being a property of any specific model.

Excluding the three contested Q2 cells (pids 10/Q2, 21/Q2, 27/Q2), the remaining 12 cells give human $\alpha = 0.73$ and LLM $\alpha = 0.79$: the panels are stability-equivalent on the non-contested subset. The whole-panel gap is a property of three specific items at the regulation-silent boundary, not of base-rate disagreement on bar-essay scoring.

F.4 Determinism probe at $T = 0$

Each LLM judge was rescored on the 15 cells with three independent seeds at $T = 0$ (seeds 44, 45, 46), giving 135 additional judgements. Within-judge cross-seed Krippendorff α is 0.998 for Claude, 0.945 for Gemini, and 0.877 for GPT-5.4 Pro; mean within-cell SD across seeds is 0.00, 0.25, and 0.39 respectively (max 1.16 on a single GPT cell). $T = 0$ is therefore not universally deterministic.

The three-LLM anchor sub-panel α across the three independent k-slices stays in $[0.76, 0.82]$, below the within-judge ceiling of ≈ 0.94. Inter-model diversity reduces α rather than inflating it: ensemble agreement is bounded above by single-model determinism, and the headline 0.77 is the diversity-suppressed value rather than an artefact of identical-input-identical-output replay.

F.5 Input ablations

Two LLM-disadvantaged ablations test whether the LLM panel agreement depends on having the regulation or the gold answer in the prompt. Humans retain both inputs throughout.

Per-judge mean score shifts under each ablation are substantial (GPT mean $+0.87$ under regulation-withheld; maximum $|\Delta| = 3.5$ under gold-withheld), so the ablations genuinely perturb the judges. The panel direction (LLM $\alpha >$ human

Judge	m_c	m_{nc}	r_B	r_C	r_A
A-direction ($m_c \leq 3$); $n = 1$, α(A, Nano) $= 0.23$					
GPT-5.4 Nano	2.33	4.62	−0.03	0.10	0.22
Middle ($3 < m_c < 5$); $n = 3$					
Qwen 3.5 Plus	4.50	6.29	0.32	0.44	0.67
Qwen 3.6 Max	4.67	6.71	0.37	0.36	0.64
Claude Haiku 4.5	4.83	5.17	0.29	0.35	0.49
B/C-direction ($m_c \geq 5$); $n = 22$; panel $\alpha = 0.56$					
Kimi K2.6	5.17	6.12	0.57	0.77	0.75
Gemma 4 31B	5.17	5.75	0.67	0.56	0.51
DeepSeek v3.2 Speciale	5.67	4.71	0.35	0.39	0.15
Gemini 3.1 Pro	5.83	6.21	0.76	0.55	0.61
Claude 4.6 Opus	6.00	5.92	0.74	0.72	0.59
GPT-5.4 Pro	6.00	6.25	0.61	0.57	0.52
GLM 5.1	6.00	6.25	0.46	0.59	0.43
Qwen 3.6 Flash	6.00	5.04	0.51	0.54	0.25
Gemini 3 Flash	6.17	7.00	0.77	0.74	0.69
DeepSeek v4 Pro	6.33	5.21	0.37	0.69	0.16
Qwen 3.6 Plus	6.33	6.54	0.51	0.54	0.44
Grok 4.1 Fast	6.50	5.54	0.48	0.75	0.25
Gemma 4 26B-A4B	6.50	6.54	0.73	0.67	0.56
Claude 4.7 Opus	6.83	6.21	0.70	0.68	0.32
Claude Sonnet 4.6	6.83	5.58	0.60	0.77	0.38
Qwen 3.6 35B-A3B	6.83	5.23	0.42	0.47	0.00
DeepSeek v4 Flash	6.83	6.21	0.42	0.38	0.34
GLM 5	7.17	6.33	0.61	0.67	0.38
GPT-5.5 Pro	7.33	7.41	0.83	0.69	0.58
GPT-5.4 Mini	7.83	6.46	−0.21	0.28	−0.28
Minimax M2.5	8.17	6.28	0.43	0.32	0.07
Gemini 3.1 Flash Lite	8.50	7.08	0.67	0.67	0.31
Human cross-graders (for reference)					
A	1.33	5.50	0.61	0.43	1.00
B	6.67	6.25	1.00	0.75	0.61
C	7.17	6.08	0.75	1.00	0.43

Table 6: Each rater's mean score on the 3 contested cells (m_c), mean score on the 12 non-contested cells (m_{nc}), and pairwise Pearson with each of the three human cross-graders on the full 15-cell vector. Rows are grouped by contested-cell direction and sorted within each group by m_c. The A-direction cluster has one member; no LLM scores within A's $[1, 2]$ contested band consistently. The B/C-direction cluster (22 of 26 LLMs) spans every model size, vendor, and price tier in the panel.

Quantity	Point	95% CI
α (A, B, C)	0.36	$[-0.09,\ 0.71]$
α (Cla, Gem, GPT)	0.77	$[0.54,\ 0.91]$
α (B, C, Cla)	0.72	$[0.41,\ 0.86]$
$\Delta\alpha$ (LLM − human)	+0.39	$[0.03,\ 0.76]$
$\Delta\alpha$ (A→Cla swap)	+0.35	$[0.04,\ 0.63]$

Table 7: Item-level bootstrap CIs at 5,000 resamples. Both direction contrasts have positive lower bounds: LLM−human is positive in 98.5% of resamples, swap-A-to-Claude in 98.8%.

α) survives both. Withholding the gold answer is the harsher ablation: it forces each judge to construct the correct legal reasoning from the question alone before grading, and the residual $\alpha = 0.47$ is most informative for legal-NLP pipelines that lack a maintained gold-answer key.

F.6 Cross-examinee replication table

Table 12 reports each Phase 1 LLM examinee reused as a judge against the three human cross-graders. Under the contested-cell direction classification of Section 4, seven of eight examinees take B/C direction on the fault line; one (Qwen 3.5 Plus, $m_c = 4.50$) sits in the regulation-silent middle gap. Zero examinees take A direction. The 8-rater LLM panel α is 0.57 on the 15 cells (vs. human 0.36). Examinee capability does not predict judge direction: Pearson(exam, Δ) $= -0.34$, -0.90 without Qwen 3.5 Plus.

Pair	Cohen κ_w	Spearman ρ
A–B	0.38	0.55
A–C	0.29	0.34
B–C	0.75	0.68
Human mean	0.47	0.52
Claude–Gemini	0.71	0.71
Claude–GPT-5.4 Pro	0.63	0.83
Gemini–GPT-5.4 Pro	0.77	0.66
LLM mean	0.71	0.73

Table 8: Pair-by-pair Cohen κ_w and Spearman ρ among the three humans and among the three LLMs. The within-LLM mean is $1.5\times$ the within-human mean under both coefficients, smaller than the $2.1\times$ panel-α gap but in the same direction. Pairwise averaging dilutes any single-rater outlier effect, which is why panel α widens the ratio relative to pairwise coefficients.

Drop	α_H	Drop	α_L
A	0.76	Claude	0.79
B	0.22	Gemini	0.80
C	0.33	GPT	0.73

Table 9: Two-rater Krippendorff α after dropping each single rater from the three-human panel (α_H) and the three-LLM panel (α_L). The human panel is asymmetric; the LLM panel is symmetric.

Drop	α_H	α_L	Drop	α_H	α_L
4/1	0.23	0.67	21/1	0.36	0.75
4/2	0.36	0.75	21/2	0.49	0.77
4/3	0.36	0.78	21/3	0.31	0.76
6/1	0.37	0.78	27/1	0.25	0.77
6/2	0.34	0.75	27/2	0.40	0.77
6/3	0.35	0.76	27/3	0.37	0.78
10/1	0.36	0.82			
10/2	0.46	0.78		range:	
10/3	0.34	0.78		$[0.23, 0.49]$	$[0.67, 0.82]$

Table 10: Leave-one-item-out α on the 14-item subset after each cell is dropped. The two ranges $[0.23, 0.49]$ (humans) and $[0.67, 0.82]$ (LLMs) are disjoint.

Configuration	LLM panel α
Baseline (full inputs)	0.77
Regulation withheld	0.57
Gold answer withheld	0.47
Human reference (full inputs)	0.36

Table 11: LLM panel α under the two LLM-disadvantaged input ablations. Both reduced configurations stay above the human reference.

Examinee	Exam	r_A	r_B	r_C	Δ
Claude 4.6 Opus	20	0.59	0.74	0.72	$+0.14$
Gemini 3.1 Pro	20	0.61	0.76	0.55	$+0.05$
GPT-5.4 Pro	16	0.52	0.61	0.57	$+0.07$
DeepSeek v3.2 Speciale	12	0.15	0.35	0.39	$+0.22$
Qwen 3.5 Plus	11	0.67	0.32	0.44	$\mathbf{-0.29}$
GLM-5	9	0.38	0.61	0.67	$+0.26$
Minimax M2.5	9	0.07	0.43	0.32	$+0.30$
Grok 4.1 Fast	8	0.25	0.48	0.75	$\mathbf{+0.36}$

Table 12: Each Phase 1 examinee scored as an LLM judge on the 15 cross-graded cells. **Exam** is the examinee's bar-exam total (out of 30). r_A, r_B, r_C are Pearson with the three human cross-graders. $\Delta = (r_B + r_C)/2 - r_A$ is the cluster strength with the B/C majority over A.