CAN LLM EXAMINERS EXPLAIN THEIR OWN JUDGMENTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

LLMs may replace or complement human decision makers in domains requiring professional discretion. In such settings, providing justifications for decisions may be ethically and legally necessary. LLMs can provide convincing justifications and explanations for their decisions, but it remains unclear whether these accounts correspond to decision policies expressed in their behavior. This issue is particularly important in high-stakes discretionary evaluation, where several considerations must be integrated without a fixed rule determining their relative importance. This study explores a pragmatic method to compare justifications and potential decision policies in the context of LLM-based exam grading using 120 synthetic answers to an authentic law exam. Each answer is scored independently by five LLM *sub-judges* on ten dimensions of answer quality derived from the exam grading rubric. The averaged sub-judge scores give an emulated objective quality assessment for each of the 120 answers. The *examiner* LLM, the main protagonist, assigns holistic grades from A to F and provides a justification for the same 120 answers without knowledge of the dimensions or the sub-judges' dimension scores. The examiner grades are then regressed on the sub-judge dimension assessments to examine which dimensions explain the variance in the examiner's grading most strongly. The resulting behavioral policy is then compared with a post-grading self-report and justifications given for individual grades. The study is replicated four times with different models as examiners (Gemma 3 (27B), Ministral 14B, DeepSeek V4 Pro, and GPT-5.6 Sol). The ten quality dimensions accounted for $69\% - 80\%$ variation in individual grades ($R^2 = 0.685 - 0.798$). The four examiner models largely agreed on which answers were stronger or weaker ($\rho = 0.83 - 0.96$), despite differences in grading severity. Behavioral cue-utilization profiles showed varying correspondence with both post-grading self-reports ($\rho = 0.32$–0.70) and local justification profiles ($\rho = 0.31$–0.71). Self-reports and local justifications were more consistently aligned with each other ($\rho = 0.67$–0.86). The local justifications largely fit the predefined answer dimensions but also identified additional evaluative considerations. Overall, the findings suggest that LLM explanations can contain meaningful information about judgments, even without fully revealing how information is systematically weighted across decisions.

1 INTRODUCTION AND BACKGROUND

Many professional decisions require judgment beyond the direct application of explicit rules. Decision makers often consider several relevant cues, determine their relative importance, and integrate them into a single conclusion (Kaufmann et al., 2013). When explicit criteria guide but do not uniquely determine the final decision, judgment necessarily involves discretion. University examination grading is one such setting, requiring examiners to weigh multiple aspects of an answer and combine them into a single overall grade (Sadler, 2009). As no fixed rule fully determines how these considerations should be balanced in every case, examiners must exercise professional discretion.

Large language models (LLMs) are increasingly being used in tasks that require evaluative judgment rather than factual retrieval, including the assessment of student work, peer review, legal analysis, and clinical decision support (Li et al., 2025). In these settings, an LLM can provide both a decision

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

and a natural language account of the considerations it claims to have used. Such explanations have the potential to improve transparency, support accountability, and increase confidence in AI-assisted decision-making (Miller, 2019). However, the value of these explanations depends on whether they reliably represent the basis of the judgment, or, at the very least, whether they reflect attributes that are related to variations in the outcomes of the evaluation. That is, explanations and justifications should include factors that are normatively and empirically related to the judgments. If this is the case, explanations and justifications provide information beyond the evaluation itself. However, explanations and justifications may represent convincing rationalizations that do not correspond to the model's observable or unobservable decision policy (Turpin et al., 2023).

The present study investigates LLMs' justifications and explanations in the context of grading a law exam. The exam included an official grading rubric from which we could derive dimensions of normatively relevant qualities of exam answers. We assessed whether these dimensions were reflected in the LLMs grading behavior, and in their justifications and explanations of their gradings. Comparing the recovered grading behavior and the LLMs self-reports allows one to investigate whether justifications and explanations are truly informative or just post hoc rationalizations unrelated to the judgment. Such results are informative about whether justifications can provide meaningful input to users and persons for whom LLMs are used as evaluators.

1.1 Discretion, Multiple-Cue Judgment, and Examination Grading

Research on professional judgment characterizes discretionary decisions as multiple-cue judgments in which decision-makers weight and integrate relevant information under incomplete specification (Kaufmann et al., 2013). Examination grading is a clear instance. Sadler (2009) argues that the quality of complex academic work cannot always be represented as a mechanical aggregation of independently specified criteria, particularly when the work must be judged as an integrated whole (Wyatt-Smith & Klenowski, 2013). Think-aloud studies similarly show examiners integrating evidence, prior comparisons, and assessment guidance when forming judgments (Crisp, 2010). This makes grading a useful setting for studying LLM in discretionary tasks. The outcome is explicit, the relevant dimensions can be grounded in a published grading guideline, and the same examiner can judge a population of answers that vary across those dimensions. At the same time, grading preserves the feature that matters for the present study: no rubric can fully specify in advance how every combination of strengths and weaknesses should be resolved. The examiner's observable grading policy must therefore be inferred from the pattern of decisions rather than assumed from the wording of the rubric or from the examiner's own account.

1.2 LLM-as-a-Judge and Automated Assessment

The notion of "LLM-as-a-judge" refers broadly to the use of language models to score, rank, classify, or select outputs according to stated criteria (Li et al., 2025). This provides a scalable approach to evaluating open-ended outputs that are difficult to assess with reference-based metrics. Recent surveys place LLM-as-a-judge within a broader evaluation landscape spanning programmatic metrics and human assessment, while emphasizing the importance of consistency, bias, robustness, and human validation (Krishnan, 2025; Gu et al., 2026).

Prior work on LLM-based evaluations has demonstrated different ways in which these models can be used as evaluators. For example, G-Eval prompts a language model with evaluation criteria and a structured response format to assess generated summaries and dialogs (Liu et al., 2023). MT-Bench and Chatbot Arena use capable language models to compare responses to open-ended prompts and have helped establish LLM judging as a scalable evaluation paradigm (Zheng et al., 2023). A related line of work decomposes quality into rubric dimensions. LLM-Rubric elicits rubric-level judgments and calibrates them to human ratings (Hashemi et al., 2024); CheckEval decomposes subjective criteria into traceable binary judgments (Lee et al., 2025); and SAJA maps multidimensional rubric features to human-aligned scores through lightweight calibration (Kola et al., 2026). These approaches demonstrate the value of structured criteria but primarily target reliable or human-aligned scoring. They do not examine whether a holistic LLM judge's justifications and self assessed cue weights correspond to the policy expressed in that judge's own decisions.

Educational applications illustrate both the appeal and current limitations of LLM-based assessment. Chiang et al. (2024) report the use of GPT-4 as an assignment evaluator in a university course with

more than one thousand students and documented failures to follow evaluation instructions and a vulnerability to prompt manipulation. Further, Chamieh et al. (2024) find that zero-shot and few-shot approaches struggle particularly with complex reasoning and domain-specific knowledge. More recent work finds systematic differences between human and LLM grading, including variation in the rubric traits that different models emphasize (Mathew et al., 2026). Rubric-guided prompting can improve agreement with human grades, yet attribution analyses suggest that scores may depend strongly on instruction and formatting tokens, while generated rationales can hallucinate rubric elements (Haque et al., 2026), raising concerns about the interpretability of rubric-guided assessment.

Broader LLM-judge research has identified instability and bias, including position, verbosity, self-enhancement, and response-order effects (Zheng et al., 2023; Wang et al., 2024). Similar limitations arise in automated peer review, particularly for long-document evaluation and critical feedback (Zhou et al., 2024). Evaluation design is therefore consequential: explicit criteria improve alignment with human judgments, whereas chain-of-thought reasoning before scoring adds little when criteria are already well specified (Yamauchi et al., 2026).

1.3 Plausibility and Faithfulness of LLM Explanations

Explanations help users interpret, question, and communicate a decision (Miller, 2019), but plausibility is distinct from faithfulness. A plausible explanation is one that appears coherent and convincing to a reader, whereas a faithful explanation accurately represents the factors that produced the model's output (Jacovi & Goldberg, 2020). Empirical studies reinforce this distinction. Models may omit influential biasing cues and rationalize resulting answers (Turpin et al., 2023), while interventions on generated reasoning show that stated chains of thought might not reflect the basis of final answers (Lanham et al., 2023).

More recent work moves toward behavioral and counterfactual tests of faithfulness. Matton et al. (2025) define faithfulness in terms of whether the concepts identified as influential in an explanation correspond to concepts that produce causal changes in model output. Their method uses realistic counterfactual modifications and estimates concept-level effects across examples and datasets. Faithfulness is thus seen as a testable correspondence between reported reasons and behaviorally consequential features, rather than as a property inferred from linguistic quality alone.

In research on human judgment and decision-making (JDM), a line of research from the 70s has concerned the comparison of statistical models of judgments and judges' own accounts of their behavior. One typical conclusion is that judges have limited self-insight (Slovic & Lichtenstein, 1971), although more recent results are more optimistic (Morris et al., 2025). One method from this research tradition is to regress a judgment outcome (e.g., a score) on cues relevant to the judgment. The pattern of associations between cue values and decisions across cases can be used to estimate a cue-utilization policy (Karelaia & Hogarth, 2007). Comparing this behaviorally recovered policy with the importance the judge reports assigning to the same cues provides a way to assess self-insight into judgment (Reilly & Doherty, 1992). The approach has been used for human examination marking, where examiners' ratings of specific assessment dimensions were used to model their overall marks (Elander & Hardman, 2002)

Related work has similarly compared what LLMs report with what their decisions reveal. In medical decision-making, models' stated beliefs and objectives can diverge from the preferences or policies inferred from their actual choices (Yamin et al., 2026a;b). In rubric-conditioned grading, agreement between LLM and expert judgments declines as rubric granularity increases (Deng et al., 2025). Marioriyad et al. (2026) used controlled cue perturbations to test whether injected metadata influences model verdicts. They report substantial behavioral effects for some cues, while judges rarely acknowledge those cues in their natural-language rationales, providing a clear example of a gap between the factors that affect an evaluation and the factors that the evaluator claims, or appears, to have used. Furthermore, counterfactual work also identifies rubric artifacts and failures to update judgments when responses or criteria are reversed (Bagaria et al., 2026).

1.4 The Present Research

In the present study we are interested in the following question:

To what extent are LLM examiners' justifications and explanations informative about dimensions that are associated with their actual grading behavior?

More specifically, we investigate whether normatively relevant dimensions of answer quality derived from the grading rubric are reflected in the examiners' grading behavior and local justifications. We then compare the emphasis on these dimensions in justifications and retrospective importance reports with their predictive importance for actual grading. We will use the approach developed in the JDM literature, where self-reported information about the judgment is compared with statistically recovered plausible policies in regression analyses.

The contribution is threefold. First, we have connected research on LLM explanation reliability with the established JDM literature on policy self-insight. Second, we offer a black-box method for comparing what an LLM judge says influenced its decisions with what statistically appears to influence them across repeated cases. Third, we evaluate this correspondence in a controlled instantiation of discretion-based judgment: holistic grading of synthetic university law examination answers generated against an authentic exam and grading rubric. The focus is therefore not simply whether an LLM can assign plausible grades, but whether its explanation of those grades provides information about dimensions that are normatively relevant (derived from the rubric) and related to the actual grading.

2 METHODS

2.1 TASK AND DOMAIN

In this study, discretionary judgment is investigated through university exam grading. Exam grading provides a useful setting for studying discretionary judgment because there is no single correct answer; the examiner must weigh several considerations against a grading guideline and justify the resulting grade. Similar features arise in other high-stakes evaluative settings in which multiple considerations must be combined into a single decision.

Two questions were selected from an authentic Autumn 2025 written exam *International Human Rights Law: Institutions and Procedures* (University of Oslo, 2026) to form the grading task used in the study. An examiner is required to combine several aspects of answer quality (legal knowledge, the use of legal authorities, analytical reasoning, and the organization of the response) into a single holistic grade rather than applying a fixed scoring rule. This makes the examination task suitable for investigating how LLM evaluators combine multiple cues when forming a judgment. Synthetic students responded to the same examination task, and all sub-judges and holistic examiners evaluated the resulting fixed set of responses.

2.2 SYNTHETIC STUDENT AND ANSWER GENERATION

A total of 120 synthetic student personas were generated and used to produce examination answers with Anthropic's Claude Opus 4.8 (Anthropic, 2026). This design created a controlled population in which student characteristics and intended answer quality could be varied systematically, while the resulting set of answers were held fixed across all subsequent examiner conditions.

Following the Item Response Theory (IRT) conception of *ability* as a continuous latent trait, each persona was assigned a latent ability parameter, θ (Wainer et al., 2007). Here, θ was used as an IRT-inspired simulation parameter rather than an empirically estimated ability score. To ensure coverage across the intended performance range, personas were stratified into six predefined ability bands: A $[0.80, 0.95]$, B $[0.65, 0.80]$, C $[0.50, 0.65]$, D $[0.35, 0.50]$, E $[0.20, 0.35]$, and F $[0.05, 0.20]$. Twenty personas were sampled uniformly within each band, yielding 120 personas in total. Each persona was also assigned background characteristics including age, faculty, language background, writing style, and a Big Five personality profile (John et al., 1999). These characteristics were sampled independently of θ so that they varied across personas without being systematically associated with intended performance. Preparation level was assigned separately and was positively related to θ by design, such that higher-ability personas were more likely to receive more thorough preparation.

The latent ability parameter θ represented the student's overall intended level of performance. However, students with the same overall ability were not expected to perform equally well in every aspect

of the examination. To introduce this variation, each persona was assigned answer quality targets: ECHR source use, UN/international source use, source variety and integration, systemic understanding, legal knowledge, reasoning and analysis, and structure and coherence. These targets were based on the persona's overall ability but were allowed to vary independently. As a result, two personas with similar values of θ could have different strengths and weaknesses; for example, one could be stronger in legal reasoning but weaker in the use of legal sources.

Each synthetic student persona was then used to generate one examination answer with Claude Opus 4.8 comprising of responses to the two selected exam questions. For each persona, the model received the exam questions, the persona characteristics, the assigned quality targets, and relevant curriculum materials, including applicable legal doctrines, cases, and sources, with instructions to write the answer as that synthetic student persona would write, reflecting the student's characteristics. The answer generation prompt is provided in Listing 1, Appendix A. Generation was performed once, producing 120 answers in total. These answers were then held fixed across all subsequent sub-judge and holistic examiner conditions, so that every model evaluated the same set of answers.

2.3 Sub-Judge Scoring of Answer Quality

To emulate an objective measure of answer quality, we derived ten quality dimensions from the official grading rubric. Six dimensions concerned the use of legal sources: European Convention on Human Rights (ECHR) case count, ECHR case quality, United Nations (UN) sources count, UN sources quality, source types count, and source integration quality. The remaining four captured broader aspects of answer quality: systemic understanding, legal knowledge, reasoning and analysis, and structure and coherence.

Each of the 120 examination answers were scored on all ten dimensions by five LLM sub-judges: Qwen 3.7 Plus, Gemma 3 (27B), Ministral 14B, DeepSeek V4 Pro, and GPT-5.6 Sol. Each scoring call presented a sub-judge with one examination answer and one dimension, together with the definition and scoring range for that dimension. The sub-judge was instructed to evaluate on the specified dimension only. The dimensions included both count and quality dimensions, with separate prompts for the two types of dimension. For count dimensions, the number of relevant legal authorities or source types present in the response was to be recorded. Quality dimensions were to be evaluated using anchored ordinal scales, with written descriptions specifying what each score represented. The prompts for the sub-judges are provided in Listings 2 and 3 in Appendix B.1. The descriptions and scoring scales for all ten dimensions are provided in Table 6 in Appendix B.2.

To reduce variability from stochastic token generation, each answer-dimension pair was evaluated five times by each sub-judge model and averaged, resulting in one model-level score for each answer and dimension. This produced five independently obtained model-level scores for every answer-dimension pair.

2.4 The Examiner

The holistic grading task consisted of two separate stages. In the first stage, the examiner received all 120 examination responses in a single call and evaluated them in the order in which they were presented. For each response, the examiner assigned a holistic grade from A to F and produced a short natural-language justification for that individual grade. These justifications are referred to as *local justifications*. The examiner was given the official grading guidelines but was not shown the ten dimensions used in the sub-judge analysis or any of the sub-judge scores. The prompt for this stage is provided in Listing 5, Appendix C. The full grading guideline is in Listing 4, Appendix C.

The second stage was conducted in a separate call after grading. The examiner was provided with the examination responses together with the grades and local justifications produced in the first stage. At this point, the ten answer-quality dimensions were introduced explicitly. The examiner was asked to retrospectively report the relative importance it believed each dimension had played in its grading decisions by assigning weights across the ten dimensions that summed to 100 as described in Listing 6, Appendix C. It also produced a brief reflection describing its overall grading approach. These weights are referred to as the examiner's *self-reported policy*.

This two-stage procedure separates two forms of explanation. The local justification is produced at the time of an individual grading decision, whereas the self-reported policy is a retrospective account

of the examiner's overall grading behavior, elicited after the examiner had access to its previously assigned grades and justifications.

The holistic grading task was replicated four times, with different LLMs serving as the examiner in each replication: Gemma 3 (27B), Ministral 14B, DeepSeek V4 Pro, and GPT-5.6 Sol. To account for variability arising from answer order and stochastic generation, each examiner graded the same 120 examination responses in a randomized order across 50 repeated runs, yielding approximately 6,000 individual grade assignments per examiner.

3 RESULTS

3.1 EXAMINER'S GRADING BEHAVIOR AND CONSISTENCY

Descriptive statistics for all individual grading observations are reported in Table 1, where grades F–A are coded 1–6. The examiners differed in grading severity. Ministral assigned the highest mean grade, followed by Gemma, GPT-5.6 Sol, and lastly DeepSeek. Ministral's observed grades ranged from 3 to 6 (D–A), whereas the other three examiners assigned grades ranging from 2 to 6 (E–A). Across the 50 repeated grading runs, single-run consistency was high within all four examiners, with $\mathrm{ICC}(C, 1)$ ranging from 0.79 to 0.85. Pairwise Spearman correlations measured whether examiners ranked the 120 answers similarly and showed strong agreement across the models' average ratings ($\rho = 0.83$-0.96, all $p < 0.001$), with the strongest association between DeepSeek and GPT-5.6 Sol ($\rho = 0.96$) and the weakest between Ministral and GPT-5.6 Sol ($\rho = 0.83$).

Table 1: Descriptive statistics of grades (F–A coded 1–6) assigned by each LLM examiner to 120 exam answers across 50 repeated grading runs. Statistics are computed over all individual grading observations. Gemma has three missing observations due to omitted or unparseable grades in three separate runs, each affecting a different candidate.

Examiner	N	Mean	SD	Min	Max
Gemma	5,997	5.01	0.91	2	6
Ministral	6,000	5.33	0.72	3	6
DeepSeek	6,000	4.56	1.18	2	6
GPT-5.6 Sol	6,000	4.74	0.96	2	6

3.2 BEHAVIORALLY RECOVERED CUE-UTILIZATION POLICIES

To estimate which quality dimensions were associated with each examiner's grades, all grades from the 50 runs per answer were retained, yielding approximately 6,000 observations per examiner. Grades were regressed on the ten measured dimensions for each answer using pooled ordinary least squares (OLS), with standard errors clustered per answer to account for the 50-fold grading of the same answer (Table 2). The ten dimensions jointly explained a substantial proportion of grade variation across examiners, with R^2 ranging from 0.685 for Ministral to 0.798 for DeepSeek.

Table 2: Behaviorally recovered cue-utilization policies. R^2 denotes the proportion of variation in individual grades explained jointly by the ten answer dimensions. The three largest dominance-analysis contributions are shown for each examiner.

Examiner	R^2	Three largest contributions (%)
Gemma	0.752	Legal knowledge (23.73), Reasoning (14.54), ECHR count (14.33)
Ministral	0.685	Legal knowledge (23.65), Reasoning (17.69), Structure (17.14)
DeepSeek	0.798	ECHR quality (18.91), ECHR count (17.29), Legal knowledge (14.44)
GPT-5.6 Sol	0.792	ECHR quality (21.88), ECHR count (20.00), Legal knowledge (14.29)

The relative contribution of each dimension to the explained variance was then estimated using dominance analysis (Budescu, 1993). The resulting contributions were normalized to sum to 100%,

producing the behaviorally recovered cue-utilization profile for each examiner. The dominance profiles differed across examiners. Gemma and Ministral placed greater relative importance on legal knowledge and reasoning, whereas DeepSeek and GPT-5.6 Sol showed greater reliance on ECHR case-related dimensions. The full dominance profiles and standardized regression coefficients are reported in Tables 9 and 10, Appendix D.

3.3 Behavioral Policy vs. Local Justifications and Self-reported Policy

The behaviorally recovered cue-utilization profiles were next compared with the two accounts provided by the examiners (Section 2.4): their *local justifications* when grading unaware of the ten quality dimensions and their post-grading *self-reported policy* explicitly on the ten dimensions.

For the former, we examined the natural-language justifications produced for individual grades. Each justification was deductively coded for the presence or absence of the ten answer dimensions used in the behavioral analysis, and mention rates were calculated for each examiner. The deductive coding revealed substantial differences in the dimensions that examiners implicitly invoked when explaining individual grades. DeepSeek and Gemma most frequently invoked legal knowledge, systemic understanding, and reasoning and analysis, whereas Ministral placed a particularly strong emphasis on reasoning and analysis. GPT-5.6 Sol highlighted legal knowledge in almost every justification and also frequently referred to systemic understanding. Source-related dimensions were less common in the justifications from Gemma and GPT-5.6 Sol, but appeared more frequently for Ministral. The coded data files and code-book are provided in the Supplementary Materials.

Regarding the self-reported policy, we averaged the weights each examiner gave across the 50 grading runs, the 'other' category was excluded, and the resulting mean weights for the ten shared dimensions were normalized to sum to 100%.

Figure 1 shows all three judgment profiles side by side, while Table 3 reports their pairwise Spearman rank correspondence. The normalized justification profile is used only to compare the relative emphasis placed on different dimensions; unlike the behavioral and self-reported profiles, it does not represent an estimated weighting policy. The three profiles show only partial correspondence. Across all four examiners, ECHR case count received lower importance in the post-grading self-reports and local justifications than in the behaviorally recovered profiles.

Table 3: Spearman rank correlations between behavioral policy weights, post-grading self-reported weights, and coded local-justification weights for each LLM examiner.

Examiner	Behavior vs. Self-report	Behavior vs. Justification	Self-report vs. Justification
Gemma	0.321	0.697	0.770
Ministral	0.467	0.709	0.855
DeepSeek	0.697	0.309	0.733
GPT-5.6 Sol	0.650	0.394	0.669

3.4 Inductive Themes in Grade Justifications

A stratified sample of the local justifications was additionally analyzed inductively to identify recurring evaluative themes not captured by the predefined dimensions (Terry et al., 2017). The ten predefined dimensions captured most of what the examiners discussed when explaining their grades. However, the inductive analysis identified three recurring aspects of answer quality that were not adequately represented by the original dimensions: professional legal register, economy and non-repetition, and epistemic assurance. A fourth theme, holistic grade calibration and cross-part compensation, concerned how examiners combined strengths and weaknesses when deciding the overall grade rather than a separate quality of the answer itself.

Professional legal register concerned whether the answer was written in an appropriately formal legal-academic style. Examiners, for example, criticized answers for being "too informal" or for using conversational expressions such as "I think." *Economy and non-repetition* concerned whether the answer communicated its points efficiently without unnecessary repetition or excessive wording;

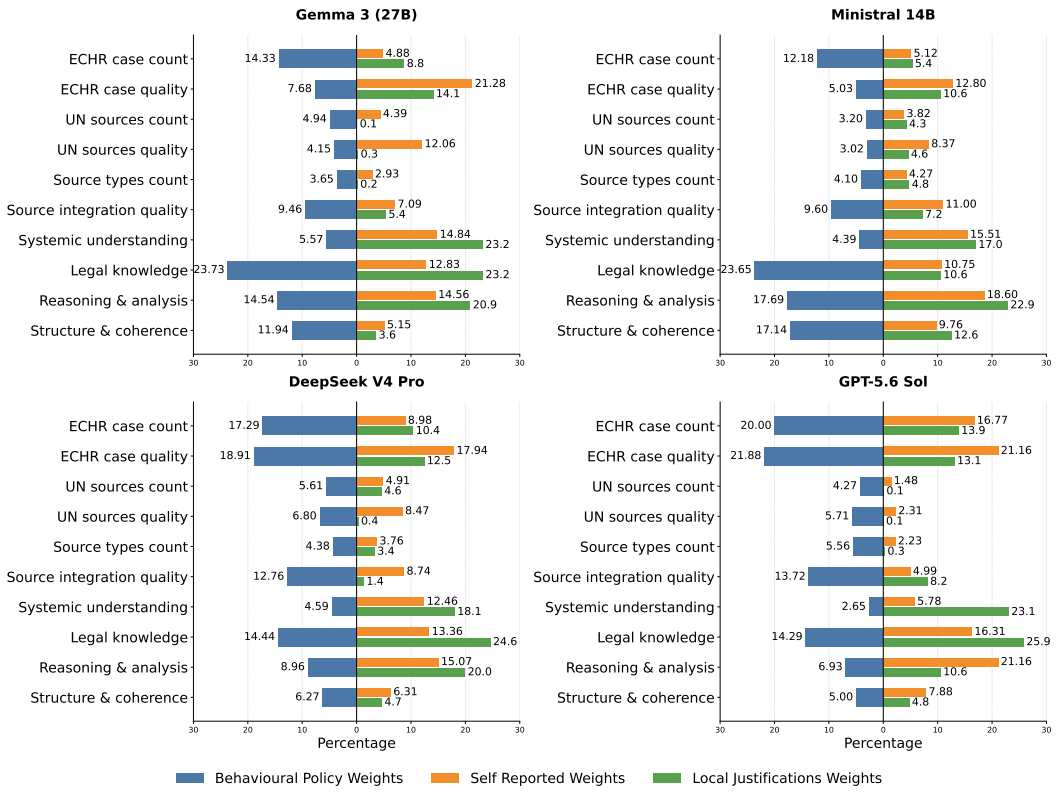


Figure 1: Comparison of the ten answer-quality dimensions across three measures of examiner judgment for each LLM examiner. All three measures are normalized to sum to 100 within each examiner. Behavioral policy weights are shown to the left of the center axis; their displayed values are positive dominance shares of explained variance. Self-reported and coded local-justification weights are shown to the right.

justifications included comments that an answer was "somewhat repetitive" or, conversely, "concise." *Epistemic assurance* concerned how confidently legal propositions were presented when the substantive content itself was otherwise acceptable. Examiners sometimes described answers as "tentative," "hesitant," or "uncertain." All three themes appeared across the four examiners, although epistemic-assurance judgments were particularly visible in the justifications produced by GPT-5.6 Sol and DeepSeek. Details are given in Tables 12–15, Appendix D.

The additional meta-evaluative theme, *holistic grade calibration and cross-part compensation*, captured statements showing how examiners translated multiple strengths and weaknesses into an overall grade. These included references to higher-grade thresholds and judgments about whether strong performance in one part of an answer could compensate for weaknesses elsewhere. For example, an examiner noted that one section was considerably stronger but "cannot compensate fully" for weaknesses in another section. Unlike the preceding themes, this theme describes the grading process itself rather than an additional property of answer quality.

4 Discussion

The central question of this study was to what extent LLM examiners' justifications and explanations are informative about dimensions associated with their actual grading behavior. Across the four models, grading behavior was systematic: the ten measured answer dimensions explained a substantial proportion of variation in assigned grades, and the examiners showed high agreement in their relative assessment of the same answers. However, the relative importance of individual dimensions differed across examiners, showing that similar overall judgments can arise from different cue-utilization profiles.

The examiners' post-grading self-reports showed partial correspondence with the cue-utilization patterns recovered from their grading behavior, but varying across models. In other words, some examiners, Deepseek ($\rho = 0.70$) and GPT-5.6 sol ($\rho = 0.65$) were better than others at reporting answer dimensions that were related to actual grading. Discrepancies were also visible at the level of individual dimensions. Most notably, all four examiners reported mere mention of quantity of the cases as less important than it appeared to be in their behaviorally recovered policies.

According to Rakow (2014) an apparent gap in recovered judgment policy and self-insight could be due to incorrect specification of the policy model and inadequate elicitation of self-insight. In our study, the regression model is not assumed to be a true model of the underlying judgment process, and we cannot establish whether a justification or an explanation accurately reflects the policy underlying the judgment. However, we can assess whether the post hoc weighting and the local justifications correspond to the predictive importance of factors associated with the actual grading behavior. LLM-generated explanations should therefore be treated as accounts of judgment that can be tested against observed behavior, rather than as direct descriptions of the process that produced the decision.

The local grade justifications are particularly important because they are explanations provided alongside the grade, and would therefore be the account directly recieved by an affected user. Rank correlations between justifications and behaviorally recovered cue utilization ranged from $\rho = 0.31$ to $\rho = 0.71$ across models. The justification profiles also corresponded consistently with the post-grading self-reports ($\rho = 0.67$–0.86), indicating some coherence between the two verbal accounts of what mattered. Most evaluative content concerned the same ten dimensions used in the behavioral analysis, while the inductive analysis identified additional considerations, including professional legal register, economy and non-repetition, epistemic assurance, and cross-part compensation. These findings suggest that local justifications contain information about normatively relevant dimensions associated with grade variation and may, therefore, add information beyond the grade itself. Our approach does not establish whether the justification for a particular answer is informative, but it shows that at the aggregate level, the rubric quality dimensions relates to the grading in a meaningful way and that the same dimensions were important in the justifications and the post hoc self-reported weighting. Together, the three measures capture different aspects of examiner judgment: behavioral analysis shows which dimensions systematically predict grades, post-grading self-report shows what the examiner later describes as important, and local justifications show what the examiner chooses to mention when explaining an individual decision.

The study is limited to one legal grading task using synthetic responses and LLM-based scores of answer quality. The recovered cue-utilization policies describe behavioral associations rather than internal model mechanisms, and the local justifications likewise should not be interpreted as direct traces of latent reasoning. The predefined dimensions may also omit features that influence grading, as indicated by the additional considerations identified in the inductive analysis. Future work should examine whether these policies remain stable under changes in batch composition, contrasts between neighboring responses, different exam languages, and other discretionary judgment tasks.

5 Conclusion

This paper examined whether LLM examiners' justifications and explanations can be informative about their grading behavior. We find that much of the variation in grades could be explained by normatively relevant quality dimensions derived from the exam rubric. Both post-grading self-reports of the importance of the quality dimensions and local justifications for individual exam answers were meaningfully related to the dimensions associated with actual grading. Local justifications fit rubric-derived categories, with some references to other evaluative dimensions. The methods are pragmatic and cannot establish whether the explanations accurately explain the process behind the judgments. However, the study suggests that LLM justifications in the context of exam grading provide information beyond the grade itself. In applied contexts, LLM explanations could be useful if they clarify judgments, even without fully revealing how information is systematically weighted.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic examination responses, evaluate responses with LLM sub-judges and examiners, assist with analysis code, debug the code, and draft and polish parts of the manuscript. We have not used generative AI tools to generate ideas or make final decisions about the interpretation of results, and other disclosure categories not relevant to this work are not applicable. We reviewed all AI-assisted work. LLM-generated and debugged code was tested against the underlying data and outputs, reported statistics were manually checked, and manuscript claims and references were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, code, analyses, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study did not involve human participants, real student examination responses, or personal data. All student personas and examination responses were synthetically generated. The work investigates the behavior and explanations of LLMs in a controlled grading setting and does not evaluate or endorse their use as autonomous replacements for human examiners. Given the potential fairness, accountability, and transparency concerns associated with automated decision-making in high-stakes settings, the findings are interpreted as evidence about model behavior under the studied conditions rather than as evidence of deployment suitability.

REPRODUCIBILITY STATEMENT

The experimental design and analysis procedures are described in Section 3. The appendix provides the full prompts used for answer generation, sub-judge scoring, holistic grading, and post-grading self-report, together with the persona-generation procedure, ability-band definitions, scoring dimensions, and model settings. Additional implementation details, analysis code, and intermediate outputs used to produce the reported statistics and figures are provided in the supplementary material. These materials are intended to support the reproduction of the synthetic data generation, grading experiments, and statistical analyses reported in this work.

REFERENCES

Anthropic. Introducing claude opus 4.8. `https://www.anthropic.com/news/claude-opus-4-8`, May 2026. Accessed: 2026-09-16.

Anshul Bagaria, Sowmya S Sundaram, Gokul S Krishnan, and Balaraman Ravindran. Judging LLM-as-a-judge: Concerning rubric artifacts in LLM-based automated text generation evaluation. *arXiv preprint arXiv:2609.02942*, 2026.

David V Budescu. Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychological bulletin*, 114(3):542, 1993.

Imran Chamieh, Torsten Zesch, and Klaus Giebermann. LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th workshop on innovative use of nlp for building educational applications (bea 2024)*, pp. 309–315, 2024.

Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-yi Lee. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 2489–2513, 2024.

Victoria Crisp. Judging the grade: exploring the judgement processes involved in examination grading decisions. *Evaluation & Research in Education*, 23(1):19–35, 2010.

Haotian Deng, Chris Farber, Jiyoon Lee, and David Tang. Rubric-conditioned llm grading: Alignment, uncertainty, and robustness. *arXiv preprint arXiv:2601.08843*, 2025.

James Elander and David Hardman. An application of judgment analysis to examination marking in psychology. *British Journal of Psychology*, 93(3):303–328, 2002.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on LLM-as-a-judge. *The Innovation*, 7(6), 2026.

Samiha Haque, Danula Hettiachchi, and Daniel Beck. Rubric-guided prompting for automatic assessment scoring. In *Proceedings of the Thirteenth ACM Conference on Learning@ Scale*, pp. 428–432, 2026.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.745. URL https://aclanthology.org/2024.acl-long.745/.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4198–4205, 2020.

Oliver P John, Sanjay Srivastava, et al. The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. 1999.

Natalia Karelaia and Robin M Hogarth. Determinants of linear judgment: A meta-analysis of lens model studies. *Available at SSRN 1002859*, 2007.

Esther Kaufmann, Ulf-Dietrich Reips, and Werner W Wittmann. A critical meta-analysis of lens model studies in human judgment and decision-making. *PloS one*, 8(12):e83528, 2013.

Sneha Kola, Pankaj Kumar Sharma, Soumyadeep Dey, Bamdev Mishra, and Mayur Datar. Saja: A simple approach to judge alignment for LLM-as-a-judge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pp. 646–664, 2026.

SB Krishnan. LLM evaluations: a survey of programmatic, human, and LLM-as-judge approaches. *International Journal of Innovative Research in Science, Engineering and Technology*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. Checkeval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15771–15798, 2025.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, 2025.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Arash Marioriyad, Omid Ghahroodi, Ehsaneddin Asgari, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. The judge who never admits: Hidden shortcuts in LLM-based evaluation. *arXiv preprint arXiv:2602.07996*, 2026.

Jerin George Mathew, Sumayya Taher, Anindita Kundu, and Denilson Barbosa. LLMs do not grade essays like humans. *arXiv preprint arXiv:2603.23714*, 2026.

Katie Matton, Robert Ness, John Guttag, and Emre Kiciman. Walk the talk? measuring the faithfulness of large language model explanations. In *International Conference on Learning Representations*, volume 2025, pp. 73212–73277, 2025.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Adam Morris, Ryan W Carlson, Hedy Kober, and Molly J Crockett. Introspective access to value-based multi-attribute choice processes. *Nature communications*, 16(1):3733, 2025.

Tim Rakow. Self-insight research as (double) model recovery. *Behavioral and Brain Sciences*, 37 (1):37, 2014.

Barbara A Reilly and Michael E Doherty. The assessment of self-insight in judgment policies. *Organizational Behavior and Human Decision Processes*, 53(3):285–309, 1992.

D Royce Sadler. Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & evaluation in higher education*, 34(2):159–179, 2009.

Paul Slovic and Sarah Lichtenstein. Comparison of bayesian and regression approaches to the study of information processing in judgment. *Organizational behavior and human performance*, 6(6): 649–744, 1971.

Gareth Terry, Nikki Hayfield, Victoria Clarke, Virginia Braun, et al. Thematic analysis. *The SAGE handbook of qualitative research in psychology*, 2(17-37):25, 2017.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

University of Oslo. JUS5710 – International Human Rights Law: Institutions and Procedures. `https://www.uio.no/studier/emner/jus/humanrights/JUS5710/`, 2026. Accessed: 2026-09-16.

Howard Wainer, Eric T. Bradlow, and Xiaohui Wang. *Item response theory*, pp. 24–43. Cambridge University Press, 2007.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024.

Claire Wyatt-Smith and Val Klenowski. Explicit, latent and meta-criteria: Types of criteria at play in professional judgement practice. *Assessment in Education: Principles, Policy & Practice*, 20 (1):35–52, 2013.

Yusuke Yamauchi, Taro Yano, and Masafumi Oyamada. An empirical study of LLM-as-a-judge: How design choices impact evaluation reliability. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 167–176, 2026.

Khurram Yamin, Jingjing Tang, Santiago Cortes-Gomez, Amit Sharma, Eric Horvitz, and Bryan Wilder. When agents say one thing and do another: Validating elicited beliefs from LLMs. *arXiv preprint arXiv:2602.06286*, 2026a.

Khurram Yamin, Jingjing Tang, Eric Horvitz, and Bryan Wilder. Can revealed preferences clarify LLM alignment and steering? *arXiv preprint arXiv:2605.08556*, 2026b.

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. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

Ruiyang Zhou, Lu Chen, and Kai Yu. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pp. 9340–9351, 2024.

A STUDENT ANSWER CONSTRUCTION

A.1 PROMPT TEMPLATES

Listing 1: Full prompt used to generate the synthetic examination answers. Bracketed variables were populated separately for each synthetic student persona.

```
You are simulating a university student answering an exam.

EXAM: {exam_subject}

QUESTION 1: {question_1}

QUESTION 2: {question_2}

STUDENT PROFILE:
- Age: {age}, Faculty: {faculty}
- Language background: {language_background}
- Preparation: {preparation_level}
- Writing personality: {writing_style_desc}

COURSE KNOWLEDGE BASE (use only what fits the student's quality profile;
    strong students cite precisely, weak students vaguely or not at all)
{course_knowledge}

Write this student's natural exam answers. Just write as this
particular student would naturally answer based on their background.

ANSWER QUALITY PROFILE (write the answer to match these
descriptions exactly):
{factor_prose}

INSTRUCTIONS:
- Write approximately 250-300 words per answer
- Match the quality profile EXACTLY - do not make the answer better
  or worse than described
- Match the writing personality in tone, confidence, and style
- Write in English
- Do NOT mention the student profile or quality descriptions in
  the answer
- Write each answer as flowing prose

Respond ONLY with valid JSON:
{"persona_id": "{pid}",
 "q1_answer": "answer here",
 "q2_answer": "answer here"}
```

A.2 PERSONA CONSTRUCTION

Table 4: Sampling scheme used to construct the 120 synthetic student personas.

Attribute	Sampling procedure
Latent ability (θ)	Twenty personas were sampled uniformly within each of six predefined ability bands: A $[0.80, 0.95]$, B $[0.65, 0.80]$, C $[0.50, 0.65]$, D $[0.35, 0.50]$, E $[0.20, 0.35]$, and F $[0.05, 0.20]$.
Age	Sampled from $\mathcal{N}(23, 3.5)$ and clipped to $[19, 40]$.
Faculty	Law (45%), Political Science (20%), International Relations (15%), Philosophy (10%), and Sociology (10%).
Language background	Native English (35%), fluent non-native English (40%), and intermediate non-native English (25%).
Preparation level	Sampled conditionally on θ. For $\theta > 0.65$: 60% thoroughly prepared, 30% moderately prepared, and 10% minimally prepared. For $0.40 < \theta \leq 0.65$: 25% thoroughly prepared, 50% moderately prepared, and 25% minimally prepared.. For $\theta \leq 0.40$: 10% thoroughly prepared, 35% moderately prepared, and 55% minimally prepared.
Big Five traits	Each trait was sampled independently from $\mathrm{N}(0.5, 0.15)$ and clipped to $[0.05, 0.95]$. Trait scores were converted to prose writing-style descriptions.
Random seed	All persona sampling used a fixed random seed of 42.

B SUB-JUDGES

B.1 PROMPT TEMPLATES

Each sub-judge evaluated one dimension at a time. Count-based dimensions and quality-based dimensions used separate prompt templates. The placeholders shown below were populated with the corresponding dimension definition, scale, and student answer at runtime.

B.1.1 COUNT-BASED DIMENSIONS

Listing 2: Prompt template for count-based dimensions.

```
Count ONE specific dimension in this student exam answer.

Count ONLY this. Ignore everything else.

WHAT TO COUNT: {dimension_description}

VALID RANGE: {minimum} to {maximum}

STUDENT ANSWER:

{student_answer}

Respond ONLY with valid JSON:

{"score": <integer {minimum}-{maximum}>,
 "reason": "one sentence listing what you counted"}
```

B.1.2 QUALITY-BASED DIMENSIONS

Listing 3: Prompt template for quality-based dimensions.

```
Score ONE specific aspect of this student exam answer.

Score ONLY this aspect. Ignore everything else.

ASPECT: {dimension_name}

MEANING: {dimension_description}

SCALE: {minimum} to {maximum} ({scale_description})

STUDENT ANSWER:

{student_answer}

Respond ONLY with valid JSON:

{"score": <integer {minimum}-{maximum}>,
 "reason": "one sentence"}
```

B.2 TEN QUALITY DIMENSIONS

Table 5: Dimensions used by the LLM sub-judges to evaluate the student exam answers. Each dimension was evaluated independently.

Dimension	Type	Range	Definition
ECHR case count	Count	0–10	Number of distinct ECHR cases cited by name. The quality of their use is not considered.
ECHR case quality	Quality	1–5	How well ECHR cases are used, considering correctness, relevance, and substantive application rather than mere name-dropping.
UN sources count	Count	0–10	Number of distinct UN or international sources cited, including treaty bodies, charter bodies, general comments, and instruments.
UN sources quality	Quality	1–5	How well UN sources are used, considering correct identification, distinction between treaty and charter bodies, and substantive application.
Source types count	Count	1–5	Number of five legal source types appearing in the answer: treaties/conventions, customary international law, general principles, soft law, and case law.
Source integration quality	Quality	1–5	How well different legal source types are integrated into a coherent argument rather than presented separately.
Systemic understanding	Quality	1–5	Extent to which the answer demonstrates a holistic understanding of the international human rights protection system, including global and regional dimensions and how they interact, overlap, or diverge.
Legal knowledge	Quality	1–6	Accuracy and depth of knowledge of international law as applied to human rights, including legal sources, jurisdiction, legal standing, and applicable law.
Reasoning and analysis	Quality	1–5	Analytical depth, including application of law beyond description, independent analysis, legal precision, and completeness in addressing the core issues.
Structure and coherence	Quality	1–6	Organization and coherence of the answer, including logical development of the argument and consistency of quality throughout.

Table 6: Scoring anchors used for the seven quality-based sub-judge dimensions.

Dimension	Scoring anchors
ECHR case quality	1 = none cited, or all wrong/irrelevant; 2 = name-dropped, no real application; 3 = correct but surface-level use; 4 = correctly integrated into the argument; 5 = precise – each case relevant, correct, and drives the argument.
UN sources quality	1 = none, or all vague/incorrect; 2 = mentioned in passing, no specifics or clear errors; 3 = cited but not distinguished or integrated; 4 = correctly identified and meaningfully applied; 5 = accurate, well-differentiated, treaty/charter distinction clear.
Source integration quality	1 = listed in isolation, no connection; 2 = mentioned together but not integrated; 3 = partial, uneven integration; 4 = well-integrated and mutually reinforcing throughout; 5 = seamless – all sources build one unified argument.
Systemic understanding	1 = no systemic awareness, treats systems in isolation; 2 = basic awareness but no meaningful connection between systems; 3 = some systemic understanding with partial interaction discussion; 4 = solid systemic understanding with meaningful interaction analysis; 5 = excellent holistic understanding with precise illustration of interactions.
Legal knowledge	1 = major misunderstandings or no legal knowledge; 2 = basic knowledge with significant errors; 3 = generally correct with some gaps or errors; 4 = accurate and well-grounded legal knowledge; 5 = comprehensive and precise legal knowledge throughout; 6 = exceptional – outstanding precision, depth, and command of international human rights law.
Reasoning and analysis	1 = purely descriptive, no analysis, major gaps; 2 = mostly descriptive with limited analytical moments; 3 = some analytical depth but uneven or incomplete; 4 = solid analytical depth with mostly complete coverage; 5 = excellent independent analysis with comprehensive and precise coverage.
Structure and coherence	1 = incoherent, no logical organisation; 2 = some structure but disjointed or very uneven; 3 = adequate structure with some weak sections; 4 = clear and well-constructed argument throughout; 5 = excellent structure, coherent flow, consistently high quality; 6 = masterful – structure is invisible, every section builds the legal argument with no gaps or redundancy.

C EXAMINER MATERIALS

This is the exact grading guideline taken from International Human Rights Law: Institutions and Procedures Autumn 2025 Examination.

Listing 4: Assessment criteria provided to the examiner LLM.

```
Assessment criteria:

Purpose:

This course examines the international system of human rights
protection, focusing on both global and regional mechanisms.
Students are expected to demonstrate understanding of how
different.

Expected competence:

Candidates should:

- Apply ECHR case law and other international instruments,
  accurately including decisions and general comments from
  UN bodies.

- Differentiate UN treaty and charter bodies,

- Demonstrate a systemic and holistic understanding of the
  international human rights protection system, in both its
  global and regional dimensions.

- Explain general aspects of international law as they apply
  to human rights, including:
    - the legal nature and effects of different sources
      (treaties, custom, general principles, and soft law),
    - questions of jurisdiction, legal standing, and
      applicable law.

- Show analytical ability, legal precision, and
  well-structured argumentation supported by relevant
  sources.

Particular emphasis:

ECHR jurisprudence is given particular focus, but answers
that ignore UN treaty monitoring and other mechanisms will
be considered incomplete. Candidates should illustrate how
these systems interact, overlap, or diverge.

Assessment criteria (A-F):

- A (Excellent): Comprehensive and precise use of ECHR and
  UN materials; strong systemic understanding; clear legal
  reasoning and independent analysis.

- B (Very good): Solid integration of ECHR and UN sources;
  few weaknesses in structure or analysis.

- C (Good): Adequate discussion of core issues; generally
  correct but somewhat descriptive or uneven.

- D (Satisfactory): Limited depth; incomplete source use;
  weak systemic analysis.

- E (Sufficient): Basic knowledge with notable gaps;
  descriptive and lacking structure.
```

```
- F (Fail): Major misunderstandings; no engagement with ECHR
  or UN sources; lacks legal reasoning.
```

Listing 5: Prompt used to elicit holistic grades and local justifications from the examiner LLM.

```
You are an examiner. Grade all {N} exam answers below.

{EXACT GRADING GUIDELINES}

Exam questions the candidates were answering:
Question 1: {QUESTION_1}
Question 2: {QUESTION_2}

Here are all {N} answers:

{ANSWERS}

Grade each answer one by one in the order presented.

For each give one grade (A-F) and a 2-3 sentence justification.

Respond ONLY with a valid JSON array, no markdown:

[
  {"candidate": "CANDIDATE_001",
   "grade": "X",
   "justification": "..."},
  ...
]

Include ALL {N} candidates.
```

Listing 6: Post-grading prompt used to elicit the examiner's self-reported policy.

```
You just finished grading {N} exam answers.

Here are all your grades and justifications:

{GRADES_AND_JUSTIFICATIONS}

Reflect on your grading. For each dimension below, estimate
what percentage of your decisions it influenced. You should
also include an "other" field for anything that influenced
your grading but is NOT captured by the dimensions below.

All values including "other" must sum to exactly 100.

{DIMENSION_LIST}

- other: anything that influenced your grading not covered by
  the dimensions above (e.g. overall impression, writing style,
  tone, etc.)

Respond ONLY with valid JSON:

{
  "dimension_weights": {
    {DIMENSION_KEYS},
    "other": <0-100>
  },
  "sum_check": <must equal 100, sum of all dimension_weights
                including other>,
  "comparison_effects":
```

```
    "did you compare answers to each other? describe",
  "external_factors":
    "any factors outside the rubric that influenced grading",
  "overall_reflection":
    "how consistent were you and what drove your grades most"
}
```

D ANALYSIS TABLES

Table 7: Within-examiner reliability across 50 repeated grading runs. ICC(C,1) indicates the consistency of a single grading run, whereas ICC(C,$k = 50$) indicates the reliability of the mean across the 50 runs.

Examiner	Answers	Runs	ICC(C,1)	ICC(C,50)
Gemma	120	50	0.831	0.996
Ministral	120	50	0.784	0.995
DeepSeek	120	50	0.860	0.997
GPT-5.6 Sol	120	50	0.866	0.997

Table 8: Pairwise Spearman rank correlations between LLM examiners. Correlations are computed across the 120 examination responses using examiner-level mean grades across repeated runs.

Examiner 1	Examiner 2	ρ	p
Gemma	Ministral	0.943	< 0.001
Gemma	DeepSeek	0.945	< 0.001
Gemma	GPT-5.6 Sol	0.921	< 0.001
Ministral	DeepSeek	0.849	< 0.001
Ministral	GPT-5.6 Sol	0.825	< 0.001
DeepSeek	GPT-5.6 Sol	0.962	< 0.001

Table 9: Standardized regression coefficients for the ten answer-quality dimensions by examiner. Standard errors are cluster-robust at the response level.

Dimension	DeepSeek	GPT-5.6 Sol	Gemma	Ministral
ECHR case count	0.252	0.258	0.245	0.240
ECHR case quality	0.418	0.393	0.004	-0.114
Legal knowledge	0.146	0.249	0.434	0.374
Reasoning and analysis	0.107	0.068	0.088	0.116
Source integration quality	-0.110	0.027	0.016	0.008
Source types count	0.006	-0.058	0.083	0.044
Structure and coherence	0.047	-0.087	0.061	0.204
Systemic understanding	0.101	-0.005	0.087	0.072
UN sources count	0.051	0.069	-0.022	0.028
UN sources quality	0.120	0.124	0.068	-0.038

Table 10: LMG dominance shares for the ten answer-quality dimensions by examiner. Values represent each dimension's percentage contribution to the variance explained by the full regression model and sum to approximately 100% within each examiner.

Dimension	DeepSeek	GPT-5.6 Sol	Gemma	Ministral
ECHR case count	17.29	20.00	14.33	12.18
ECHR case quality	18.91	21.88	7.68	5.03
Legal knowledge	14.44	14.29	23.73	23.65
Reasoning and analysis	8.96	6.93	14.54	17.69
Source integration quality	12.76	13.72	9.46	9.59
Source types count	4.38	5.56	3.65	4.11
Structure and coherence	6.27	5.00	11.94	17.14
Systemic understanding	4.59	2.65	5.57	4.39
UN sources count	5.62	4.27	4.94	3.20
UN sources quality	6.80	5.71	4.15	3.02

Table 11: Additional themes identified in the inductive analysis of local grade justifications

THEME	TYPE	DEFINITION	RECURRENCE
Professional legal register	Answer quality	Evaluation of whether the answer adopts an appropriately formal and professional legal-academic style, including avoidance of conversational, casual, or excessively personal phrasing.	259 justifications across 115 distinct examiner-answer combinations; observed for all four examiners.
Economy and non-repetition	Answer quality	Evaluation of whether the answer communicates its points efficiently without unnecessary repetition or excessive wording.	142 justifications across 84 distinct examiner-answer combinations; observed for all four examiners.
Epistemic assurance	Answer quality	Evaluation of the degree of confidence, decisiveness, or excessive hedging with which the student presents legal propositions, independently of whether those propositions are substantively correct.	Observed across all four examiners, particularly GPT-5.6 Sol and DeepSeek.
Holistic grade calibration and cross-part compensation	Meta-evaluative	Explicit statements showing how the examiner converts multiple strengths and weaknesses into an overall grade, including higher-grade thresholds, comparison with stronger responses, and whether a strong section can compensate for a weak one.	11 of 339 sampled justifications (3.2%): Sol 6, Ministral 3, Gemma 2, DeepSeek 0.

Table 12: Representative excerpts for professional legal register

EXAMINER	REPRESENTATIVE EXCERPT
Gemma	"the style is too informal"
Ministral	"occasionally uses informal phrasing, such as 'I think' and 'I would say"'
DeepSeek	"slightly informal but analytically sound"
GPT-5.6 Sol	"the legal analysis is basic and informal"
	"remains somewhat informal"

Table 13: Representative excerpts for economy and non-repetition

EXAMINER	REPRESENTATIVE EXCERPT
Ministral	"The answer is correct but somewhat repetitive…"
Gemma	"The writing is somewhat repetitive."
DeepSeek	"somewhat repetitive and lacks depth…"
	"Concise but accurate and well-structured."
Ministral	"a clear, concise explanation"

Table 14: Representative excerpts for epistemic assurance

EXAMINER	REPRESENTATIVE EXCERPT
Gemma	"Vague and hesitant phrasing ('one might argue', 'I am not fully certain')."
Ministral	"…contains some inaccuracies or uncertainties, such as 'I am somewhat uncertain here."'
GPT-5.6 Sol	"…although some analysis is tentative."
	"The first answer is substantively correct …although the discussion is somewhat tentative and descriptive."
DeepSeek	"somewhat tentative"
	"somewhat uncertain"

Table 15: Representative excerpts for holistic grade calibration and cross-part compensation

EXAMINER	REPRESENTATIVE EXCERPT
GPT-5.6 Sol	"The admissibility section is considerably better …but cannot compensate fully for Question 1."
	"Question 1 is strong …Question 2 is less precise …so the overall answer is uneven."
Ministral	"…would be expected for a higher grade."
Gemma	"…required for a higher grade."
GPT-5.6 Sol	"…preventing a higher grade under the stated criteria."
	"…narrower than in stronger responses."