BENCHMARK RELIABILITY: A CONTROLLED COMPARISON OF HUMAN PREFERENCE, RUBRIC SCORING, AND LLM JUDGES ON ENTERPRISEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM judges are validated by their agreement with expert ratings, a practice that treats expert reliability as a fixed property of the raters. We show that it depends substantially on what those raters are asked. On EnterpriseVAL, a benchmark of expert-authored finance and accounting tasks, we hold tasks, outputs and rubrics fixed and vary only what the judgement asks for: blinded preference between an LLM output and the expert reference deliverable; a score on each criterion of the task rubric; and a binary critical-failure gate, where a single disqualifying error fails the output. Expert agreement rises as the claim under judgement becomes more specific: Krippendorff's $\alpha = 0.20$ on preference, $\alpha = 0.55$ on criterion grading, Cohen's $\kappa = 0.90$ on the critical-failure gate. Two consequences follow for validation: a judge's agreement with experts is uninterpretable without the expert–expert baseline on the same instrument, and preference can favour outputs the gate disqualifies. Reliability trades against discriminative power. The critical-failure gate, our most reliable instrument, separates the evaluated LLMs least, because a criterion every LLM passes, or every LLM fails, carries no comparative information. Reliability and discriminative power are separate claims about an instrument, and a benchmark tuned to maximise the first can end up measuring nothing about the LLMs it exists to compare. Where a single wrong value disqualifies a deliverable, categorical correctness belongs in a benchmark's results as a claim of its own, reported separately from graded and preference measures.

1 INTRODUCTION

Expert evaluation does not scale, so the field increasingly substitutes automated judges for experts. This is justified by early findings that an LLM judge agrees with human raters most of the time (Zheng et al., 2023), and larger, more careful studies are now qualifying that reassurance. An LLM judge, however, cannot be validated against a human rating without knowing how reliable the measurement instrument that produced those ratings is, and how much of its variance the judge is being asked to reproduce. Agreement with humans is a comparison against a moving baseline, and the baseline is rarely reported.

Two lines of work bear on this. The first argues that inter-annotator agreement does not bound achievable LLM performance (Boguslav & Cohen, 2017; Richie et al., 2022), a position placed on a theoretical footing by Xu et al. (2023), who derive an upper bound on the average annotator's accuracy against an unobserved oracle and certify several classifiers as superhuman on SST and SNLI. That work distinguishes an LLM's accuracy against the oracle from its agreement with human annotations, and notes that the two are frequently confounded. Our claim concerns agreement only. We do not treat the expert panel as an approximation to truth, and we make no claim that its agreement bounds a judge's achievable accuracy. The second line of work concerns what a single metric can be asked to do. Krishnappa et al. (2026) show that automatic metrics for open-ended question answering are asked to perform two jobs at once: establishing validity, or distinguishing genuine content alignment from surface coincidence, and exercising discriminative power, or ranking a better system above a worse one. On opinion-driven tasks, no metric does both well. Cosine similarity separates real answers from random ones at Cohen's $d \approx 2$ yet cannot rank the five LLMs they evaluate.

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.

We observe the same structure among human and judge instruments, between the reliability of a measure and its power to separate systems.

How benchmarks are judged is changing in a related way. Hua et al. (2026) propose a work-centered representation with four fields (represented activity, tested setting, required work product, and evaluated result) and apply it to released cases including GDPval (Patwardhan et al., 2025). In their GDPval case the evaluation ends at deliverable quality, leaving workflow linkage, approval records and audit trail outside the episode; they present this as a case-level demonstration rather than a characterisation of the benchmark as a whole. GDPval's authors report the complementary result: on an under-contextualised variant with prompts 42% of their original length, LLM performance fell, as the LLMs had to discover the constraints and locate the data themselves (Patwardhan et al., 2025).

This paper asks the adjacent question of how the judgement is elicited. Answering it requires tasks hard enough to produce genuine expert disagreement. Expert-authored rubrics have become the standard instrument for capturing expert judgement, and several recent benchmarks apply them to professional work. ProfBench pairs more than 7,000 response-criterion items with expert annotations across Physics and Chemistry PhD, Finance MBA and Consulting MBA, and builds affordable LLM judges that mitigate self-enhancement bias, with GPT-5-high reaching 65.9% overall (Wang et al., 2026b). PRBench provides large-scale expert rubrics for high-stakes professional reasoning in finance and legal domains (Akyürek et al., 2025). BigFinanceBench pairs each of 928 expert-authored financial-research tasks with a point-weighted rubric that decomposes the derivation into independently checkable steps across 36,241 rubric points, supporting partial fail and the localisation of failures across the analyst workflow (Wang et al., 2026a). APEX-Agents releases 480 long-horizon, cross-application tasks authored by investment banking analysts, management consultants and corporate lawyers, each carrying an expert-written rubric of one to ten binary criteria graded independently by an LLM judge, with a task passing only if every criterion is met; the strongest of eight agents reaches 24.0% Pass@1 (Vidgen et al., 2026).

EnterpriseVAL, our benchmark of high-complexity enterprise tasks, takes finance and accounting (F&A) as its first domain, with tasks authored by expert practitioners from their own daily workflows and reviewed by an independent expert panel. It sits alongside these other benchmarks on task design and it differs in two respects. First, in what is done with the rubric: no previous work examines a comparison between elicitation formats on the same outputs, which is what our design is built to do. Second, in the artefact evaluated: deliverables are multi-file and in native formats, so a failure can be a specific wrong value in a named location rather than a judgement about a document as a whole.

The methodological commitment is that each task-response pair goes through three evaluations. The first draws on two established traditions: comparative judgement, which builds reliable measurement from many pairwise comparisons rather than absolute ratings (Thurstone, 1927; Pollitt, 2012), and its modern form in open arena-style leaderboards (Chiang et al., 2024; Zheng et al., 2023). The second applies the expert-authored rubrics described above. To these we add a third, a critical-failure gate within the same rubric. The second and third are each applied twice: once by human experts, once by an LLM judge. Where GDPval distinguishes expert preference from automated grader and leaves the two separate, our design is built to measure how they relate. The three differ only in how specifically they pose the question: which of two documents is preferred; how far each rubric criterion is met; and whether a single named outcome is present. Because tasks, outputs and rubrics are held fixed across all three, the form of the judgement is the only thing that varies, which makes the instruments themselves the object of measurement.

We contribute:

- A measurement of elicitation format as a determinant of expert agreement. Holding tasks, outputs and rubrics fixed, chance-corrected inter-rater agreement rises with the specificity of the instrument, and we identify which criterion properties drive it within a single instrument.
- A reframing of expectations for LLM judges. Because agreement with a human expert is uninterpretable without the reliability of the evaluation instrument, a judge's validity cannot be established without first characterising expert performance on that instrument. The same judge validated against our preference instrument, where experts agree at $\alpha = 0.20$.
- Evidence that a measure's inter-rater agreement trades against its power to separate LLMs. Our most reliable measure is the least able to distinguish the systems under comparison,

because criteria that every system passes or every system fails are perfectly consistent and carry no comparative information.

- EnterpriseVAL, a benchmark of complex finance and accounting tasks authored and independently reviewed by expert practitioners, in which every task-response pair carries three evaluations over identical outputs: blinded expert preference, expert rubric grading with a critical-failure gate, and an LLM judge on the same rubric.

2 BUILDING THE BENCHMARK: TASK CREATION, EXPERT REVIEW, RUBRIC DESIGN

Task-response pairs. The building block of EnterpriseVAL is a task-response pair. The task comprises three elements: a prompt setting out an enterprise-realistic scenario; input files, which provide everything required to complete the task; and a purpose-built rubric which defines the scoring criteria applied by human or LLM judges. The response is the 'gold standard' that answers the prompt in full, our reference deliverable.

The prompt and input files are given to the LLM being benchmarked, which produces an output: the model's counterpart to the reference deliverable.

A task is admissible only if it meets three conditions: it demands real judgement, calling for trade-offs, interpretation and prioritisation rather than mere extraction or summary; it produces a digital deliverable that can be graded and scored; and it can be broken down into checkable true-or-false assertions. Admissible tasks are also generic across industries, hold their relevance for at least six months, and take a qualified professional two to six hours to complete, excluding data gathering.

Expert authoring and review. Authoring begins with a use case proposal within a practitioner's area of expertise. The candidate topic (task prompt) is approved by the expert group before building starts. Review of the task-response pair is carried out by a separate team, organised by sub-domain expertise, in five phases: an overview pass, an independent solve, a peer challenge session, a remediation meeting with the author, and final validation and closure. The strict independence of author and reviewer, with staged quality control, is what backs the claim that EnterpriseVAL is expert-verified.

Rubric design. Every task carries a purpose-built rubric written by its author (see example in Appendix B). A criterion is a single assertion that an expert or LLM grades against the output, scored 1 if met and 0 if not, with a third outcome, a partial fail worth half the weight, where the author has pre-specified the condition for it (Appendix B). Each criterion also carries a dimension (see Appendix C) and a weight from 1 to 5. Grading is the act of judging an output against the rubric's criteria and a score is the value it produces. An output's graded score is the sum of its weighted criterion scores, reported as a proportion of the rubric's total weight so that tasks with differently sized rubrics can be pooled. The same rubric is graded by the expert panel and by the automated LLM judge, which is what makes cross-evaluations directly comparable.

In each rubric the author flags one or more criteria as critical in a dedicated column. A critical criterion captures the outcome the whole task is built towards, such as the final value of a calculation, the decision or classification the task exists to produce, or the conclusion the task depends upon. The critical-failure gate is the rule that applies them: failing any one critical criterion fails the gate, and failing the gate disqualifies the output outright. This prevents a polished but wrong answer from scoring well, however well it scores elsewhere. The frequency of critical criteria failures is reported as a separate result rather than blended into the average. Critical criteria are scored as binary pass/fail, since being wrong on one is not a matter of degree; partial fail, whether from a grader or the LLM judge, is resolved to a pass or a fail before the critical-failure gate is applied (Appendices D and E).

Eighty-one finance and accounting tasks were produced in four core F&A sub-domains: Record-to-Report, Order-to-Cash, Financial Planning and Analysis, and Procure-to-Pay; and two additional sub-domains, Treasury and Risk & Compliance. Each was authored and reviewed by experts with an average of 22 years' experience, and each has passed the five-phase expert review with all findings closed. Twenty-five tasks, drawn by stratified random sampling, form the shared evaluation core for our studies (Appendix F).

3 STUDY DESIGN

Each sampled task-response pair is evaluated three times. The three evaluations are ordered by how specifically they pose the question to the rater. The three studies share one stratified sample of 25 tasks. Study 1 evaluates six LLMs, Studies 2 and 3 five; all matched analyses use the five shared LLMs, giving 125 outputs each carrying a preference verdict, an expert rubric score and a critical-failure gate outcome, and an LLM judge score and critical-failure gate outcome.

Study 1: blinded expert preference. Each task is set against six LLMs, one pairwise comparison per LLM, each rated by two independent evaluators (300 ratings over 150 events across 25 tasks). The six LLMs are: Claude Fable 5; Gemini 3.6-Flash; GLM-5.2-FP8; GPT5.6-SOL-high; GPT5.6-SOL-max; and Muse Spark 1.1. Evaluators see an anonymised expert response against an anonymised LLM output side by side, rendered in native file types, with left-right position randomised and the map retained for unblinding. Blinding serves two purposes: an evaluator who knows which output is the reference might favour it, and blinding leaves room for the more interesting result, that an expert sometimes prefers the LLM output over the expert response. Each evaluator scores both outputs 1–5 on five attributes (Structure, Style, Format, Accuracy and Factual Faithfulness, Professional Judgement and Reasoning), then gives a mandatory free-text rationale and a preference verdict. Disagreements are retained rather than reconciled, so agreement reflects genuine rater variation.

Study 2: expert rubric grading & critical-failure gate. The same tasks and outputs are graded, criterion by criterion, against each task's rubric by a different rater group: experts who already hold the task context; they are referred to as graders. Grading a single output against a multi-item rubric requires familiarity with the input files and the intended deliverable, which makes fully independent raters infeasible at scale. Only outputs from five LLMs are graded: Claude Fable 5; Gemini 3.6-Flash; GLM-5.2-FP8; GPT5.6-SOL-high; and Muse Spark 1.1. GPT5.6-SOL-max is evaluated in Study 1 only, so all matched analyses use the five shared LLMs. Two experts forming a grader panel grade each output independently with no reconciliation. LLM identity is masked behind labels randomised per task and per grader. The rubrics comprise 691 criteria (13–69 per task) carrying 1,804 weighted points, of which 50 are critical criteria, so every task is gated. Five LLMs per task gives 125 outputs and 6,910 criterion judgements. Weights concentrate on substance: Relevance and Accuracy carries 70% of available points, Structure 15%, Format 6%, Aesthetics 5%, Style 4%.

Study 3: automated judge. An LLM judge (Claude Opus 4.8, see Appendix E for self-referential bias tests) grades the identical rubric on the identical outputs, with the critical-failure gate applied afterwards exactly as on the human side. Each output is graded in a single pass, so the cross-grader resolution used to address discrepancies between two graders in Study 2 does not apply: a critical criterion fails the critical-failure gate only when the judge awards no credit. One pass also means no judge-side reliability estimate is possible, and every figure reported for Study 3 is judge against grader panel. We predetermine the success requirements: for the judge to substitute for expert grading at scale, judge–human agreement on the critical-failure gate must approach the human–human chance-corrected inter-rater agreement reported in Study 2. Agreement on rubric-criterion scores is reported as a secondary outcome.

3.1 ANALYSIS PLAN

The object of study is the instrument rather than a ranking of LLMs, so the analysis establishes that the instruments measure consistently before it is used to measure anything else. Four questions are addressed in sequence, and inter-rater reliability leads the result for every study. Definitions are collected in Appendix A.

Do raters agree? Krippendorff's α is used wherever the rater pool varies or judgements are ordinal, covering the criterion-level rubric scores and the preference verdicts. Cohen's κ is used for two-rater binary decisions, the critical-failure gate, scored both expert against expert and judge against the panel's verdict. Raw percentage agreement accompanies every chance-corrected coefficient, because the two diverge sharply when the marginal distribution is skewed.

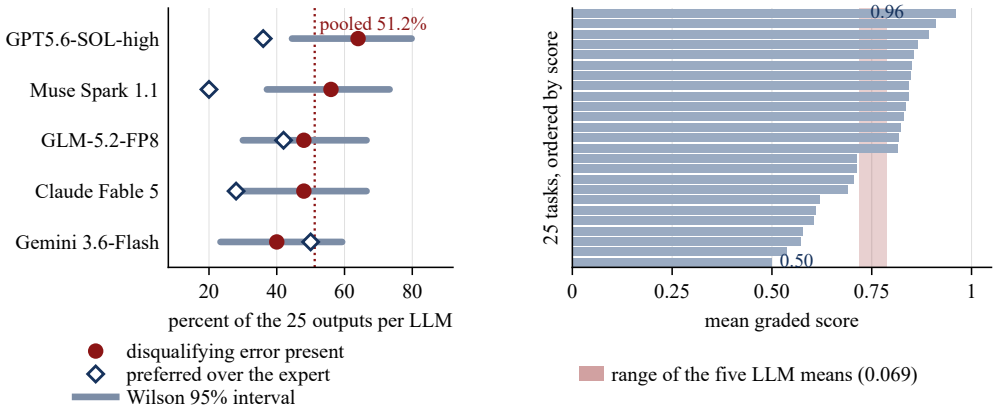
What drives agreement? Each of the 691 criteria contributes one observation: its exact-agreement rate is regressed by ordinary least squares on properties of the criterion, with predictors standardised so the coefficients are comparable. This converts a reliability figure into rubric-design guidance.

Does the judge track the panel? The LLM judge is what makes the benchmark usable, and the panel is what makes it correct. We need evidence that the cheap instrument reproduces the expensive one. Judge and panel graded scores are compared by Pearson and Spearman correlation, with a Wilcoxon signed-rank test on the paired differences, since correlation cannot detect directional bias. The judge's verdict is additionally treated as a classification against the panel verdict and reported as sensitivity and specificity, which separates the two error directions that raw agreement pools. Preference win rate is compared across gate outcomes by Mann-Whitney U with a rank-biserial effect size, and thematic code prevalence by Fisher's exact test.

Does the instrument separate LLMs? We need evidence that the instrument registers real differences rather than returning much the same verdict for every LLM. LLM comparisons use the repeated-measures test appropriate to each outcome: Friedman for the five attributes in Study 1, averaging the two evaluators' scores within each event, and for the weighted graded score, Cochran's Q for the binary gate. Post-hoc contrasts are Holm-corrected within each family, and no contrast is interpreted unless it survives that correction. Win-rate intervals come from 5,000 task-clustered bootstrap resamples with a fixed seed, clustered at task level because ratings within a task are not independent; gate failure rates use Wilson intervals.

4 RESULTS

Figure 1 summarises the benchmark. Panel (a) gives LLM performance on both human measures: expert rubric grading finds a disqualifying error in 51.2% of the 125 outputs, and blinded expert preference favours the LLM over the expert response on between 20% and 50% of comparisons, with no LLM reaching parity. Panel (b) gives the spread of task difficulty against which LLMs are compared: mean graded score by task spans 0.50 to 0.96, six times the spread between LLM means. The benchmark separates tasks far more sharply than it separates these five systems, which is a property of the sample to keep in view when reading the LLM comparisons below.



(a) LLM performance

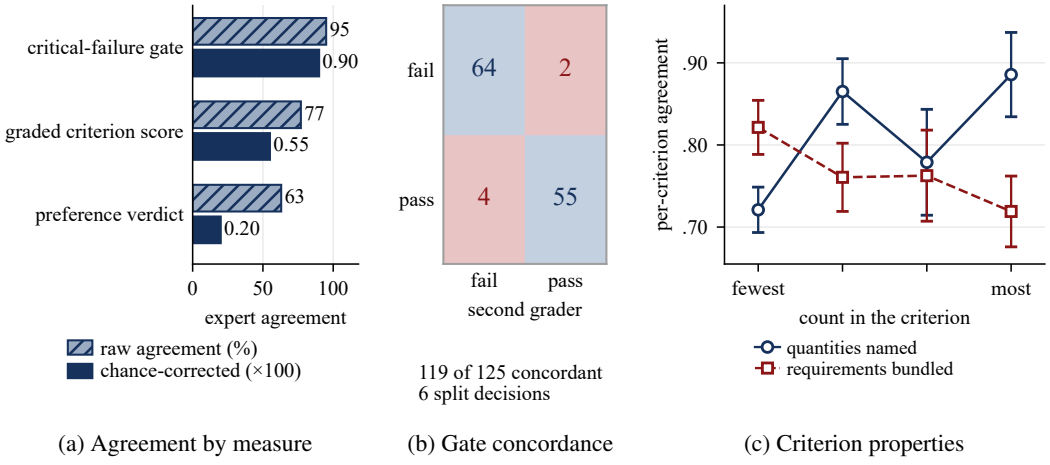
(b) Task difficulty

Figure 1: EnterpriseVAL on the 25-task evaluation core, five shared LLMs, 125 outputs. (a) Per LLM, the share of outputs with a disqualifying error and the share preferred over the expert response; the dotted line marks the pooled failure rate. (b) Mean graded score for each of the 25 tasks, ordered; the shaded band spans the five LLM means.

4.1 STUDY 1 RESULTS: BLINDED EXPERT PREFERENCE

Reliability of the preference verdict is low (Figure 2a). Two evaluators agreed on 63.3% of comparisons, Krippendorff's $\alpha = 0.197$, with 55 of the 150 rating events split one to one. Against that, no LLM performed better than the human expert: across 300 ratings the expert response was preferred in 65% of comparisons, giving a pooled LLM win rate of 35% (task-clustered bootstrap 95% CI [26.3, 43.3]). Differences between LLMs reach significance on the pooled verdict (Friedman

$\chi^2(5) = 14.010$, $p = 0.0155$), but with agreement this low the ordering is reported for completeness rather than relied upon.



(a) Agreement by measure (b) Gate concordance (c) Criterion properties

Figure 2: Reliability of the expert instrument, all panels on the same 125 outputs. (a) Agreement between the two independent experts on each measure, raw and chance-corrected. (b) The two graders' gate verdicts cross-tabulated; grader order is assigned alphabetically, so the split of the off-diagonal is arbitrary and only the diagonal and off-diagonal totals are meaningful. (c) Per-criterion exact agreement across all 691 criteria, binned by two criterion properties, with 95% intervals.

Preference ratings discriminate on presentation and not on substance. A Friedman test across the six LLMs, blocking by task and averaging the two evaluators' scores within each event, found significant differences on the three presentation attributes, Style ($\chi^2(5) = 31.452$, $p < 0.001$), Format ($\chi^2(5) = 24.223$, $p = 0.0008$) and Structure ($\chi^2(5) = 16.788$, $p = 0.015$), but not on the two substance attributes, Judgement ($\chi^2(5) = 4.962$, $p = 0.841$) and Accuracy ($\chi^2(5) = 2.047$, $p = 0.843$); p values are Holm-corrected across the five attributes. Preference ratings separate the LLMs on how the deliverable is organised, worded and presented, but not on the substance attributes.

Thematic coding of the 300 rationales (Appendix G) shows LLMs losing on accuracy and winning on completeness and decision-readiness. These are compatible rather than opposing properties, and together they are the mechanism behind "persuasive but wrong" results: an evaluator reading a fuller, better documented package credits the coverage without always catching the error inside it. Neither "structure and navigability" nor "audience fit" separates winning from losing rationales, so the presentation features that distinguish the LLMs from one another play no part in whether the expert or the LLM is preferred.

The six F&A sub-domains and their composition are set out in Appendix F. Among them, LLMs are least competitive on FP&A, where the expert is preferred in 79.2% of ratings, and most competitive on Procure-to-Pay, where LLMs are preferred on 52.8% of ratings. This finding is consistent with FP&A being the most judgement-intensive sub-domain and P2P the most procedural. The theme "completeness" appears in 80% of FP&A rationales where the LLM won against 8% where it lost, roughly a 72-point gap, against an average gap of 25 points for other F&A sub-domains. The route to winning on FP&A is comprehensiveness.

4.2 Study 2 results: expert rubric grading & critical-failure gate

Graded scores are reported as the ratio of weighted points achieved to weighted points available, so they are comparable across tasks of differing rubric size. The mean graded score pooled across LLMs was 0.753 (SD 0.149), with a range of 0.718 (Muse Spark 1.1) to 0.787 (Gemini 3.6-Flash). The five LLMs significantly differ on graded score, Friedman $\chi^2(4) = 11.374$, $p = 0.0227$; no pairwise contrast survives Holm correction, so the ordering is indicative. Task difficulty dominates LLM identity as a source of variance: task means span 0.50 to 0.96, six times the between-LLM spread (Figure 1b).

On the critical-failure gate, 64 outputs failed under both graders, 55 passed under both, and 6 received opposite decisions, resolved by the rule in Appendix D. This gives 64 fails of 125 outputs, a pooled rate of 51.2% (Wilson 95% CI [42.5, 59.8]). Per-LLM critical criterion failure rates ranged from 40.0% (Gemini 3.6-Flash) to 64.0% (GPT5.6-SOL-high).

In contrast to the graded score, Cochran's Q test on the binary outcome does not reach significance ($Q(4) = 7.172$, $p = 0.127$). The critical-failure gate is reported for the absolute level it establishes rather than as a discriminator between LLMs: on this class of professional work, roughly half of all frontier-LLM outputs contain at least one error an expert could not sign off on.

Reliability rises sharply as the criterion becomes more specific (Figure 2a). Across the 3,455 paired criterion judgements, ordinal Krippendorff's α is 0.553 with exact agreement of 77.2%. On the binary critical-failure gate, raw agreement is 95.2% and Cohen's κ 0.904, with the two graders splitting on only 6 of 125 outputs (Figure 2b). Within the instrument the same relationship holds across 691 criteria: agreement rises with the number of quantities a criterion names and falls with the number of separate requirements it bundles, while length has no effect (Figure 2c).

Each grader recorded a free-text note per output: 250 notes and 19,716 words, thematically coded with their prevalence compared on gate-failed versus gate-passed outputs by Fisher's exact test with Holm correction across the nine output codes. One code separated the gate-failed group of outputs from gate-passed outputs significantly: criticism of a wrong value, classification or decision, present in 56% of notes on failed outputs against 28% on passed (Holm $p < 0.001$). No other code distinguished them, including all four presentation codes, and criticism of excess content was slightly more common on the outputs that passed. What the critical-failure gate detects is wrongness, and the graders' language locates that wrongness in the substance of the answer rather than its presentation.

4.3 Study 3 results: the automated judge

On the per-output decision, i.e. whether the deliverable contains a critical criterion failure, the judge sits below the band of human disagreement (Figure 3). It agrees with the panel verdict on 86.4% of the 125 outputs ($\kappa = 0.729$), whereas the two expert graders agree with each other on 95.2% ($\kappa = 0.904$). It flags 57 failures against the panel's 64, with sensitivity 81.2% and specificity 91.8%.

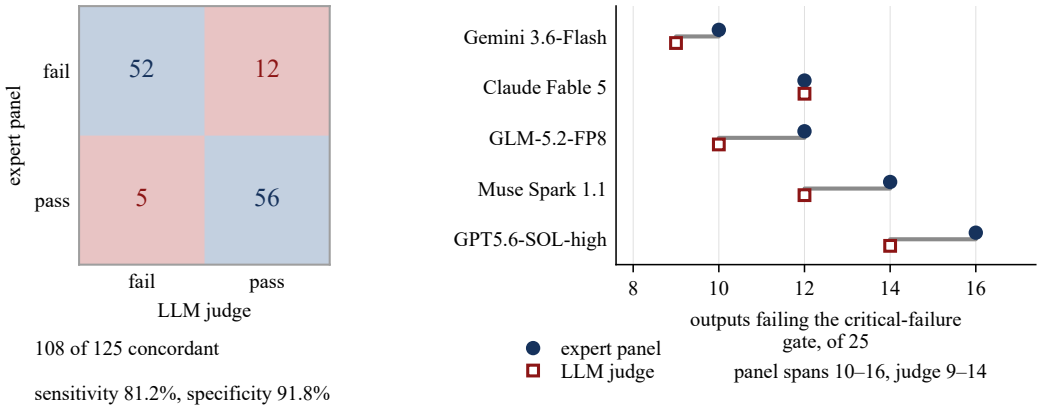
The judge does not reproduce the human grader's inferential result. On panel data the five LLMs differ on the weighted graded score (Friedman $\chi^2(4) = 11.374$, $p = 0.0227$); on judge data they do not ($p = 0.182$), and neither instrument separates the LLMs on the gate. This is compression rather than inaccuracy (Figure 3b). Per-LLM agreement is uniform between 80.0% and 92.0%, so no one LLM is mishandled; every LLM is pulled toward the average. The LLM judge's per-LLM failure counts span 9 to 14 of 25 against the expert panel's 10 to 16, failing fewer of Gemini 3.6-Flash's outputs than the panel does and fewer of GPT5.6-SOL-high's, i.e. the LLMs with the fewest and the most panel failures. Because Friedman and Cochran's Q both depend on the spread of those per-LLM totals, a judge can be accurate on each output and still lose the difference between LLMs, which is the measurement most often wanted from it. The direction is predictable, so a judge-measured difference is better read as a lower bound than an estimate, and a judge-measured null as inconclusive: judge scores can screen and detect large differences but cannot certify small ones.

5 Discussion

The three studies run over the same 125 outputs, graded against the same rubrics. This design removes the usual explanations for disagreement between evaluation methods. Where two benchmarks disagree, it is normally impossible to discern whether the material differed, the annotators differed, or the measure differed. Here, only the measure differed by design. What follows is therefore a comparison of instruments rather than of frontier LLMs.

5.1 Reliability depends on the question, not only on the raters

Two experts, asked which of two finance deliverables is better, agreed on roughly two comparisons in three. The same population, asked whether a specific stated figure appears in a specific place, agreed on nineteen outputs in twenty. Chance-corrected, that is $\alpha = 0.20$ rising to $\kappa = 0.90$, almost the whole usable range of the coefficient; raw agreement rises monotonically over the same



(a) Per output

(b) Across LLMs

Figure 3: LLM judge against the expert panel. (a) Gate verdicts where the LLM judge and the experts agreed and disagreed. (b) Outputs failing the gate per LLM, panel and judge.

three formats (63.3%, 77.2%, 95.2%), so the ordering does not depend on which coefficient is used. Across all 691 criteria, agreement rose with the number of quantities a criterion named and fell with the number of separate requirements it bundled, while criterion length made no difference: what raises reliability is anchoring a judgement to something countable, not writing at greater length. Low expert agreement is often reported as an unavoidable cost of using experts, but much of it is an outcome of the instrument.

5.2 "Our judge agrees with humans" is not a claim until the human baseline is stated

Our judge reached $\kappa = 0.73$ against the expert panel's critical-failure verdict, a figure that reads as strong in isolation and as a shortfall once the panel's $\kappa = 0.90$ is placed beside it. The same number supports opposite conclusions depending on a baseline that is rarely reported. Where the judge's errors fall is informative: of the criteria the panel failed, it awarded partial fail on nineteen and full credit on six, and because partial fail passes the gate those twenty-five account for all its misses. The judge is conservative on individual criteria, and its failures are concentrated in graduated rather than categorical judgements, which is where the expert panel was least consistent. Improving the criterion is therefore a prerequisite for improving the judge rather than an alternative to it.

5.3 Preference has a blind spot for verifiable error

On 27 of the 125 outputs at least one expert preferred the LLM output to a qualified expert practitioner's reference deliverable while the rubric panel found a disqualifying error in the LLM output; on six, both experts did. This is not noise. It concentrates in Record-to-Report and Risk & Compliance, where failures are specific postings, totals and classification decisions, and it reproduces when the judge's verdict replaces the panel's, so it is a property of the outputs rather than of one rater group. The mechanism is visible in what the evaluators wrote: LLMs lost on accuracy and won on completeness. Those are compatible properties, so a fuller and better-documented package attracts credit for its coverage from a reader who does not catch the error inside it.

5.4 Reliability trades against discriminative power

This last point cuts against the obvious inference. Our most reliable instrument is the least informative about differences between LLMs: 35 of the 50 critical criteria return the same verdict for all five LLMs, and 14 of 25 tasks are decided unanimously. A criterion every LLM passes, or every LLM fails, is perfectly reliable and carries no comparative information. The outputs that clear the gate show the same thing, their graded scores falling within a 0.048 band across all five LLMs

while the number clearing it ranges from 9 to 15 of 25. The LLMs are near-indistinguishable on the quality of work that survives review and differ substantially in how much survives. A benchmark tuned to maximise agreement risks selecting for criteria that no longer separate the LLMs it exists to compare.

5.5 LIMITATIONS

Measurement precision. The coefficients are not strictly commensurable: ordinal α on a three-point scale and κ on a binary decision are different statistics over different category structures, so the ordering across formats rather than the arithmetic difference between them is what should be relied upon. Nor is specificity the only thing that changes between formats. Moving from a preference verdict to a criterion score to a gate decision also reduces the answers available to a rater, from a free comparison to three options to two, and this design cannot separate how much of the rise in agreement comes from a more specific question and how much from fewer ways to answer it.

The reference standard is expert judgement, not ground truth. The panel's verdict is the standard against which both the preference verdict and the judge are assessed, and it is itself a human judgement. On the 27 disputed outputs we read the discrepancy as preference missing a verifiable error; that the panel was too strict cannot be ruled out, though critical criteria state whether a specified value or decision appears, so a failure is checkable against the input files, and an independent judge scoring the same criteria reached the panel's verdict on 86% of outputs. Settling them would need third-party adjudication against the source data. Our preference protocol pairs each output against an expert-authored reference with attribute ratings and a written rationale, so these findings do not transfer directly to arena-style comparison or to preference annotation for LLM training.

Sample and scope. Three judges were compared at selection, but only one was run over the full rubric set, so no judge-side reliability is estimable for the reported analysis and the compression we report could be per-output noise rather than a pull toward the centre.

Twenty-five tasks support the five-LLM comparisons but not the pairwise ones, so orderings are indicative and the absence of a gate difference reflects roughly 54% power against the observed effect rather than equivalence. The results come from one domain, one rubric design and one panel: we expect the ordering of reliability across formats and the direction of the criterion-level effects to generalise, not the specific coefficients.

5.6 CONCLUSION AND DIRECTIONS

EnterpriseVAL is a continuing evaluation programme rather than a fixed dataset. The full pool of 81 expert-authored finance and accounting task-response pairs supports ongoing benchmarking of frontier LLMs on the three instruments described here and on extensions to them. Twenty-five pairs, the evaluation core of these studies, are released as a public sample with their task definitions, input files, expert reference deliverables and rubrics, together with the preference verdicts, rubric scores and gate outcomes recorded over them.

If reliability is largely set by the form of the question, several things usually treated as fixed become design variables. Criterion writing becomes an object of study in its own right, and the two properties we identify are among many that could be tested by posing the same underlying judgement in systematically varied formats. Instrument reliability becomes a quantity to report alongside scores, so that a reader can calibrate any agreement figure a judge is validated against. And the conventional target for an automated judge, maximising agreement with human raters, becomes ambiguous wherever the human instrument is itself weak. The last of these is the most immediate opportunity: a judge validated against a well-specified instrument and one validated against a poorly specified instrument are not comparable claims, yet the literature rarely reports which is which.

The position we reach is a reporting discipline rather than a ranking. Wherever a single wrong value disqualifies a deliverable, categorical correctness should be reported separately from graded and preference measures rather than averaged into them, because the three measure different things and one of them has a blind spot for what the others measure. An instrument's own reliability belongs in a benchmark's results as a first-class claim, reported separately from its discriminative power.

AI USE STATEMENT

We disclose all LLM use in accordance with the ICLR 2027 policy and Code of Ethics; the authors are fully responsible for this paper's contents. LLMs appear in three roles. Six frontier LLMs are the objects of study, producing the deliverables evaluated here: 150 in Study 1, of which 125 from five shared LLMs carry all three evaluations (Section 3). One LLM, Claude Opus 4.8, is the LLM judge of Study 3, after three judges were screened. And the authors used LLM assistance for statistical computation, thematic analysis, figure production and prose editing. All reported statistics were recomputed from the released data and manually checked; all cited literature was verified against primary sources. No LLM generated or altered expert ratings, rubric criteria, task definitions or reference deliverables, which are the work of the named practitioners.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. All tasks were authored, reviewed, evaluated and graded by practising finance and accounting professionals within the authors' organisation, as part of their professional duties, having been informed of the purpose of the evaluation and the intention to publish aggregate results. Identities are not reported; identifiers in the released data are pseudonymous and the mapping key is withheld. Task scenarios and input files are synthetic, modelled on the authors' own workflows but containing no real client data, personally identifying information or confidential material; entity names, account identifiers and figures are fabricated. All evaluated LLMs are commercially available and were accessed under the applicable terms of service; the performance limitations we report should be read as measurements of those systems as accessed under the documented protocol, not as fixed properties of any model family.

REPRODUCIBILITY STATEMENT

Twenty-five expert-authored task-response pairs, the evaluation core of all three studies, are released with task prompts, input files, expert reference deliverables, complete rubrics, per-criterion scores from both graders, Study 1 preference verdicts and attribute ratings, free-text rationales and grader notes, and the LLM judge's per-criterion output. The remaining 56 pairs are withheld to support continued benchmarking. Analysis code for every reported statistic is released, including the criterion-level regression and its bin definitions, the gate resolution rule of Appendix D, and the interval procedures of Section 3.1. Seeds are fixed: 5,000 bootstrap resamples with seed 20260828. The single judge pass means judge-side variance is not estimable, though the model version, prompt and decoding parameters are included so the pass can be repeated.

REFERENCES

A. F. Akyürek, A. Gosai, C. B. C. Zhang, et al. PRBench: Large-scale expert rubrics for evaluating high-stakes professional reasoning. *arXiv preprint arXiv:2511.11562*, 2025.

M. Boguslav and K. B. Cohen. Inter-annotator agreement and the upper limit on machine performance: Evidence from biomedical natural language processing. *Studies in Health Technology and Informatics*, 245:298–302, 2017. doi: 10.3233/978-1-61499-830-3-298.

V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa.

W.-L. Chiang, L. Zheng, Y. Sheng, et al. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 8359–8388. PMLR, 2024.

Y. Hua, H. Na, C. Ayubcha, and L. Lian. Design and report benchmarks for knowledge work. *arXiv preprint arXiv:2605.23262*, 2026.

P. Krishnappa, A. Das, V. Jain, et al. RECOM: A validity–discrimination tradeoff in automatic metrics for open-ended Reddit question answering. *arXiv preprint arXiv:2606.19218*, 2026.

T. Patwardhan, R. Dias, E. Proehl, et al. GDPval: Evaluating AI model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374*, 2025.

A. Pollitt. The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3):281–300, 2012. doi: 10.1080/0969594X.2012.665354.

R. Richie, S. Grover, and F. Tsui. Inter-annotator agreement is not the ceiling of machine learning performance: Evidence from a comprehensive set of simulations. In *Proceedings of the 21st Workshop on Biomedical Language Processing*. Association for Computational Linguistics, 2022.

L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927. doi: 10.1037/h0070288.

B. Vidgen, A. Mann, A. Fennelly, et al. APEX-Agents. *arXiv preprint arXiv:2601.14242*, 2026.

A. Wang, G. Meinhardt, J. Katz, et al. BigFinanceBench: A workflow-grounded benchmark for financial-research agents. *arXiv preprint arXiv:2606.03829*, 2026a.

Z. Wang, J. Jung, X. Lu, et al. ProfBench: Multi-domain rubrics requiring professional knowledge to answer and judge. In *International Conference on Learning Representations*, 2026b. arXiv:2510.18941.

Q. Xu, C. Walder, and C. Xu. Humanly certifying superhuman classifiers. In *International Conference on Learning Representations*, 2023. arXiv:2109.07867.

L. Zheng, W.-L. Chiang, Y. Sheng, et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.

APPENDICES

APPENDIX A: GLOSSARY AND STATISTICAL ANALYSES

Table A1: Glossary of benchmark and study terminology

Term	Definition as used in this paper
Benchmark construction	
Task	One enterprise-realistic scenario. Admissible if it demands genuine judgement rather than extraction or summary, yields a digital deliverable that can be scored, and decomposes into checkable true-or-false assertions. Admissible tasks are generic across industries, hold relevance for at least six months, and take a qualified professional two to six hours.
Task–response pair	The building block of the benchmark. The task comprises three elements, a prompt setting out an enterprise-realistic scenario; input files, which give the LLM solving the task everything the task needs; and a purpose-built rubric which defines the scoring criteria applied by human or LLM-as-judge evaluators. The response is the expert deliverable that answers the prompt in full as the 'gold standard'.
Task response	The expert-authored deliverable that answers the scenario in full – the gold standard. In Study 1 it is the reference against which each LLM output is judged.
LLM output	One LLM's attempt at one task. The matched analyses use 125 outputs: five shared LLMs on the 25 sampled tasks.
Instrument	The form in which a judgement is elicited from an evaluator, covering the question posed, the response options available, and the scope of the comparison required. In this paper, three instruments are compared: blinded pairwise preference, graduated criterion scoring, and binary critical-failure gate.

Term	Definition as used in this paper
Evaluation core	The 25 tasks drawn from the approved pool by stratified random sampling, shared by all three studies so that differences between results are differences between instruments rather than between datasets.
Rubric and scoring	
Rubric	The task-specific scoring checklist written by the task author, made up of criteria (13–69 in our studies).
Criterion	A single assertion, scored against one output by a grader or LLM judge. 1 if met, 0 if not, and 0.5 only where the author has specified the partial fail condition. Every item is defined inside the criterion, with no reference back to the input files, so scoring never depends on re-deriving the answer. Each criterion also carries a dimension (Appendix C) and a weight from 1 to 5.
Graded score	The sum of an output's weighted criterion scores, reported as a proportion of the rubric's total weight so that tasks with differently sized rubrics can be pooled. Graded score is reported by the expert panel and automated LLM judge.
Critical criterion (gate item)	A criterion designated critical by the task author at the time of writing, flagged in its own yes/no column. It captures the outcome the whole response is built towards: the final value of the calculation, the decision or classification the task exists to produce, or the conclusion everything else depends on.
Critical-failure gate	The rule applied before overall score is computed: an output that fails any critical criterion is recorded as a critical failure. This prevents a polished but wrong answer from grading well.
Study 1: blinded expert preference	
Pairwise comparison	The Study 1 elicitation design: one LLM output judged against the expert reference response, the two shown side by side in native file types, anonymised, with left–right position randomised and the map retained for unblinding. Distinct from a post-hoc contrast (below).
Evaluator	Experts who record Study 1 preference ratings. An evaluator judges two outputs as wholes and states a preference; they do not grade against the rubric.
Rating	One evaluator's completed judgement of one pairwise comparison: 1–5 attribute scores for both responses, a mandatory free-text rationale, and a preference verdict. Study 1 comprises 300 ratings.
Event	One LLM–task comparison, each rated by two independent evaluators. Study 1 comprises 150 events (six LLMs $\times$ 25 tasks).
Win rate	LLM-preferred ratings over total ratings.
Studies 2 and 3: rubric grading and the automated judge	
Graders	Experts who grade outputs against rubric criteria in Study 2. Two graders score each output independently using the rubric criteria.
Panel	The two independent expert graders of Study 2, whose resolved verdict is the human reference against which the automated judge is assessed. LLM identity is masked behind labels randomised per task and per grader, and the two grade without reconciliation.
LLM Judge	The automated rater of Study 3: a single LLM (Claude Opus 4.8) grading the same outputs against the same rubrics as the panel, one pass per output, with the critical-failure gate applied afterwards.

Term	Definition as used in this paper
Judge-against-panel	The reporting direction of every Study 3 figure. Because there is a single judge on a single pass, no judge-side reliability can be estimated, so the judge is compared only to the panel.
Analysis and reporting	
Post-hoc contrast	A comparison between two LLMs' aggregate results, reported where an omnibus test rejects and subject to Holm correction of the familywise error rate. A distinct procedure at a different level of analysis from a pairwise comparison (above).

Table A2: Statistical procedures, by family of inference

Procedure	What it estimates	Where used
Raw percentage agreement	Share of paired judgements on which two raters recorded the same value. Not corrected for agreement arising by chance, so reported only alongside a corrected coefficient.	All three studies
Krippendorff's alpha (ordinal)	Agreement corrected for chance on a graduated scale, weighting a one-step disagreement as less severe than a two-step. 0 is chance, 1 is perfect.	Criterion-level rubric scores
Krippendorff's alpha (nominal)	The same coefficient without ordinal weighting, used where the judgement is a two-way choice. On a binary outcome ordinal and nominal forms coincide.	Preference verdict
Cohen's kappa	Chance-corrected agreement between exactly two raters on a categorical decision.	Critical-failure gate, expert against expert and judge against panel
Pearson and Spearman correlation	Strength of linear and of rank association between two continuous measures. Both are reported because agreement in ordering does not require agreement in level, and neither detects a constant offset.	Judge against panel, graded scores
Wilcoxon signed-rank test	Paired test on the differences between two measures of the same outputs, without assuming normality. Reported alongside correlation, which cannot detect a directional bias.	Judge-against-panel graded score bias
Sensitivity and specificity	Share of panel-identified failures the judge also fails, and share of panel-passed outputs it also passes. Separates the two error directions that a single agreement figure pools.	Judge against panel
Fisher's exact test	Tests association in a 2×2 table without large-sample approximation. Used because several thematic codes are sparse.	Thematic prevalence
Mann-Whitney U with rank-biserial effect size	Tests whether values in one independent group are systematically higher than in another without assuming normality. The effect size is reported with it because the test statistic gives significance but not magnitude.	Preference win rate across gate outcomes

Procedure	What it estimates	Where used
Friedman test	Repeated-measures test for ranked outcomes across more than two conditions measured on the same blocks. Used because every LLM attempts every task.	Study 1 attribute ratings; weighted graded score
Cochran's Q	The equivalent for binary outcomes, and the extension of McNemar's test beyond two conditions.	Critical-failure gate
Holm correction	Controls the familywise error rate across multiple comparisons without assuming independence. No post-hoc contrast is interpreted unless it survives.	All post-hoc contrasts
Permutation test	Compares an observed statistic against the distribution produced by relabelling the same data, so no distributional assumption is required. Used where the quantity of interest is one LLM's departure from the others.	Self-referential bias of the judge (Appendix E)
Task-clustered bootstrap	Resamples whole tasks, 5,000 times with a fixed seed, because ratings within a task are not independent.	Preference win-rate intervals
Ordinary least squares regression	Linear model fitting criterion-level exact agreement to standardised properties of the criterion, one observation per criterion, so coefficients are comparable across predictors measured on different scales.	Criterion properties against agreement, all 691 criteria
Wilson interval	Binomial confidence interval that remains valid near 0 and 1, unlike the normal approximation.	Gate failure rates

APPENDIX B. EXAMPLE OF PART OF A RUBRIC (TAKEN FROM UC01)

Some criteria are single claims with internal structure, such as a value that is correct only given a stated adjustment, or a set of required items. An output can satisfy these in part, where both full credit and zero would misreport it. We therefore allow a third outcome, a partial fail, worth half the criterion's weight. It is available only where the author has specified the exact condition for it in the rubric's Notes/Tolerance field, for example the right figure reached by a sub-optimal route or k of n named items identified. Criteria without such specification stay binary. Table B1 shows the field in use: criterion 1 awards half credit where the deliverable uses more than one tab, while criterion 2 states that any deviation from the template ordering is a full fail.

Table B1: Example of part of a rubric (taken from UC01).

Total possible score: A deliverable must be created with the following five dimensions in mind:						65
#	**Dimension**	**Criterion Description**	**Weight**	**Tag**	**Notes / Tolerance**	**Critical fails**
1	Format	Deliverables are two .xlsx workbooks named ZF61_Income Tax Accrual_Apr302026.xlsx and Monthly Tax Calculation_Apr302026.xlsx with a single tab	2	TRUE	a single tab is needed Scores 1/2 if the LLM creates 2 or more tabs	No
2	Structure	ZF61_Income Tax Accrual_Apr302026.xlsx should be structured as in the template, i.e. row 8 should have columns called: Account, Amount, Assignment, Cost center, WBS Element, Profit, Cocd, Text.	3	TRUE	Ordering of columns is a must - any deviation from what is stated in criterion is a full fail as it should take exactly from a template	No
3	Relevance & Accuracy	ZF61_Income Tax Accrual_Apr302026.xlsx workbook should contain 1011 in the Cocd field	2	TRUE		No
4	Relevance & Accuracy	ZF61_Income Tax Accrual_Apr302026.xlsx workbook should contain EUR in the correct Currency field	2	TRUE		No
5	Relevance & Accuracy	ZF61_Income Tax Accrual_Apr302026.xlsx workbook should contain 260000 in one of the Account fields	2	TRUE		No
6	Relevance & Accuracy	ZF61_Income Tax Accrual_Apr302026.xlsx workbook should contain 780000 in one of the Account fields	2	TRUE		No
7	Relevance & Accuracy	ZF61_Income Tax Accrual_Apr302026.xlsx workbook should contain 184,000 debit for Account 260000	5	TRUE		Yes

APPENDIX C: DEFINITION OF DIMENSIONS

Table C1: Definition of dimensions.

Dimension	Definition
Structure	Logical sequence and completeness of organisation
Style	Tone and professional voice appropriate to persona and audience
Format	Correct file type, layout conventions, domain-standard organisation
Aesthetics	Visual clarity: labels, units, figures; distribution-ready presentation
Relevance & Accuracy	Every factual claim traceable to the input files; no hallucination

APPENDIX D. CRITICAL-FAILURE GATE

The gate is the one element of the rubric instrument that is not a scalar, and it is applied before any score is computed. This appendix states the procedure and the results that depend on it rather than on the scores.

Designation. Fifty of the 691 criteria are designated critical by the task author at the time of writing the rubric, between one and five per task, so every task is gated. These are the criteria on which being wrong is not a matter of degree: a posting to the wrong account, a materially wrong total, the wrong recommendation. They are not graded on a separate scale and they carry author-assigned weights averaging 3.9 points against 2.5 for the remainder, so criticality was expressed through weight as well as through the flag.

Resolution across two graders. Each critical criterion carries two independent scores that must be reduced to one verdict before the gate is applied. A criterion fails when either grader records a zero and the other records nothing better than a half; it passes otherwise. The treatment of the 0 + 1.0 combination is deliberate: a flat contradiction as to whether a defect exists is treated as an unreliable judgement rather than as evidence of a defect. Table D1 gives the rule with the observed frequency of each combination.

Table D1: Resolution of a critical criterion across the two graders, with observed frequencies over the 250 critical criterion-output pairs (50 criteria $\times$ 5 LLMs).

Both graders' scores	Verdict	Pairs observed
0 + 0	Fail	88
0 + 0.5	Fail	0
0 + 1.0	Pass	9
0.5 + 0.5	Pass	3
0.5 + 1.0	Pass	0
1.0 + 1.0	Pass	150
Total	88 fail / 162 pass	250

Application. An output failing any critical criterion is recorded as a critical failure rather than given a low value. Averaging a disqualifying error into a presentation score is what allows a plausible but wrong output to pass, which is the design choice the gate exists to avoid. Table D2 gives the per-LLM consequence.

Why the gate discriminates weakly. Each critical criterion is judged five times, once per LLM. The distribution of failures is bimodal: 26 of the 50 criteria are never failed by any LLM and nine are failed by all five, leaving 15 that discriminate. Criteria that every system passes or every system fails are informative about task difficulty and carry no comparative information, which is the mechanism behind the reliability and discrimination trade-off discussed in Section 5.

Table D2: Graded score and critical-failure rate by LLM. Graded score averages 25 tasks × 2 graders (n = 50 per LLM); SD is across the 25 task-level means. Failures are counted at output level (n = 25 per LLM); intervals are Wilson 95% binomial intervals. Inferential tests are reported in Section 4.

LLM	Graded score	SD	Failures	Rate	95% CI
Gemini 3.6-Flash	0.787	0.146	10 / 25	40.0%	[23.4, 59.3]
GLM-5.2-FP8	0.760	0.163	12 / 25	48.0%	[30.0, 66.5]
GPT5.6-SOL-high	0.750	0.143	16 / 25	64.0%	[44.5, 79.8]
Claude Fable 5	0.750	0.157	12 / 25	48.0%	[30.0, 66.5]
Muse Spark 1.1	0.718	0.139	14 / 25	56.0%	[37.1, 73.3]
Pooled	0.753	—	64 / 125	51.2%	[42.5, 59.8]

Table D3: Distribution of failure across the 50 critical criteria.

LLMs failing the criterion (of 5)	Critical criteria
0	26
1	2
2	3
3	5
4	5
5	9
Total	50

APPENDIX E. STUDY 3: AUTOMATED JUDGE

Method. A single judge LLM (Claude Opus 4.8) scored each task's full rubric on all six LLMs' outputs, one pass per output, replicating the Study 2 instrument, with the critical-failure gate applied afterwards. Every Study 3 figure reported here is based on the 125 outputs from the five shared LLMs, since GPT5.6-SOL-max has no panel grades to compare against. Because the judge is run once there is no cross-grader resolution: a critical criterion fails only when the judge awards no credit. If the critical criterion is awarded partial fail it passes the critical-failure gate, mirroring the treatment of a half score on the expert side. One pass also means no judge-side reliability is estimable, so all figures are judge against panel. Both sides run the same 50 critical criteria; fifteen item judgements the judge had additionally flagged were de-flagged to match the expert set. Item-level judge output is available for all 25 tasks.

Table E1: Judge against expert panel, at output and criterion level. The panel-against-panel figures are the reliability the judge is measured against, not a competing result.

	Output level (n = 125)	Criterion level (n = 250)
Raw agreement	86.4%	86.8%
Cohen's kappa	0.729	0.697
Panel failures	64	88
Judge failures	57	71
True positives	52	63
False negatives	12	25
False positives	5	8
True negatives	56	154
Sensitivity	81.2%	71.6%
Specificity	91.8%	95.1%
Panel against panel	95.2% (κ 0.904)	—

Where the judge misses. Of the 88 criteria the panel judged failed, the judge awarded partial fail on 19 and full credit on 6. Because partial fail passes the gate, those 19 account for most of the shortfall in sensitivity, and they fall in the partial-fail region of the scale. The judge inherited the instrument's ambiguity rather than the panel's disagreements.

Comparison with a binary-criterion judge. APEX-Agents validates its judge against the same quantity we report here, on a rubric whose criteria are short self-contained statements graded true or false with no partial credit. Against 747 human-labelled criterion judgements, with the human labellers blind to the judge's grades, their judge reaches 98.1% recall and a 1.9% false negative rate (Vidgen et al., 2026). Our judge reaches 71.6% sensitivity at criterion level. The instruments differ in one respect that bears directly on that gap: our criteria admit a partial fail where the author has pre-specified the condition, and partial fail passes the critical-failure gate. Of the 88 criteria our panel failed, the judge awarded partial fail on 19 and full credit on 6, so the partial-credit option accounts for most of the shortfall. The comparison is not like-for-like, since the two benchmarks differ in domain, task length and rubric size, but it is consistent with the argument of Section 5.2 that a judge's measured agreement is a property of the question it is asked rather than of the judge alone.

Graded score. Pearson r = 0.793 against the panel, with a mean bias of +0.035 (panel 0.753, judge 0.788; Wilcoxon signed-rank $p = 2.0 \times 10^{-6}$), present in all five LLMs. Correlation alone does not detect a directional shift of this kind, which is why the paired test is reported alongside it.

Inference. On panel data the five LLMs differ on the weighted graded score (Friedman $p = 0.023$); on judge data they do not ($p = 0.182$). Neither instrument separates the LLMs on the gate (panel $p = 0.127$, judge $p = 0.107$). The judge's per-LLM failure counts span 9 to 14 of 25 against the panel's 10 to 16, with per-LLM agreement uniform between 80.0% and 92.0%, so the shortfall is a compression of between-LLM differences rather than an error concentrated on any one LLM.

Test for Self-Referential Bias. Three candidate judges were screened before Study 3, and Claude Opus 4.8, which showed no self-referential lift on critical criteria, was selected as the sole automated judge; because that judge shares a developer with one of the evaluated LLMs, we tested whether it graded that LLM unusually favourably. The screen ran on a wider base than Study 3 itself: all six LLMs from Study 1 across the 25-task evaluation core, scored on all 691 rubric criteria, giving 4,146 item-level scores. For each LLM, item-level lift was calculated as its score minus the mean score of the other five LLMs. A permutation test then assessed whether the Anthropic LLM's observed lift was unusually large relative to the between-LLM differences within the same Opus-scored data.

The same concern arises in APEX-Agents, where the judge model also appears on the leaderboard. Their mitigation is structural, withholding the agent trajectory from the judge, and their check is the false positive rate on the judge's own outputs: 1 of 84 criteria, or 1.2%, in line with the other systems they evaluate (Vidgen et al., 2026). Our design differs in that the judge and the evaluated LLM share a developer rather than being the same model, and our check is a permutation test on item-level lift rather than a false positive rate.

Table E2: Self-referential bias analysis of the Opus 4.8 automated judge.

Result	Claude Fable 5	Comparator or interpretation
Mean rubric-item score	0.830	Ranked 2nd of six; Gemini 3.6-Flash ranked 1st at 0.841
Mean item-level lift	+0.033	Gemini 3.6-Flash had a larger lift of +0.046
Permutation test	p = 0.333	Claude's lift was not unusually large
Bootstrap 95% CI	[+0.017, +0.049]	Positive performance difference, but not evidence that the magnitude was unique to Claude
Critical-criterion pass rate	0.642	Below Gemini at 0.717 and GLM at 0.670
Critical-criterion lift	0.000; p = 0.50	No advantage on the most consequential criteria

Although Claude Fable 5 scored above the average of the other LLMs, its uplift was not unusual under permutation testing, and Gemini 3.6-Flash achieved both the highest overall score and the larger item-level lift. Claude also received no uplift on critical criteria. These results provide no clear evidence of self-referential bias sufficient to materially distort the reported findings, although a small LLM-specific effect cannot be excluded because the evaluation included only one Anthropic LLM.

APPENDIX F: SAMPLING FRAME

Eighty-one task-response pairs were produced across six F&A sub-domains, each having passed the five-phase expert review with all findings closed. Twenty-five pairs, about thirty-one percent of the pool, form the shared evaluation core for all three studies. Stratified random sampling was applied across sub-domains as the pool stood at the draw, using largest-remainder rounding to the target count, with difficulty applied as a second stratification layer inside each sub-domain where the ratings allowed.

Authoring continued after the sample was drawn, so the composition of the final pool differs from the composition the sample was drawn against. This is recorded because it bears on how the sample should be described rather than on any reported result: the sample is representative of the pool at the time of drawing, not of the pool as it now stands.

Table F1: Evaluation core by sub-domain.

Sub-domain	Tasks sampled	Critical criteria	Criteria
Record-to-Report	7	16	219
Order-to-Cash	6	11	110
FP&A	4	7	168
Procure-to-Pay	3	4	58
Risk & Compliance	3	7	84
Treasury	2	5	52
Total	25	50	691

APPENDIX G. THEMATIC ANALYSIS

Method. *Both corpora were analysed following Braun and Clarke's six phases (Braun & Clarke, 2006), as codebook thematic analysis: an inductively derived frame was developed from the corpus itself rather than imported, then applied systematically. Coding was carried out with explicit lexical indicators for each code rather than by independent human coders, so no inter-coder reliability is available and prevalence should be read as indicative. Prevalence is reported at the level of the rationale or note, as the share containing at least one criticism-valenced instance of a code, and compared by Fisher's exact test with Holm correction across the codes in each frame.*

Study 1 rationales. *Three hundred rationales, 25,087 words, median 64 words, from 15 evaluators. Because each rationale explains a preference verdict, prevalence can be compared between rationales where the LLM won and where it lost, which distinguishes correlates of winning from correlates of losing. Ten codes were derived. Five separate the outcomes after correction.*

Table G1: *Prevalence of each code in Study 1 rationales, by whether the LLM won or lost the comparison (n = 105 won, 195 lost). Fisher exact, Holm-corrected across ten codes.*

Code	LLM won	LLM lost	Diff	p (Holm)
Completeness and coverage	48%	18%	+30	<0.001
Decision-readiness	69%	50%	+19	0.015
Analytical depth and reasoning	64%	47%	+17	0.045
Evidence and audit trail	45%	30%	+15	0.064
Audience fit and register	57%	52%	+5	0.934
Structure and navigability	74%	73%	+1	0.891
Readability and clarity	42%	50%	−8	0.677
Task fit and brief adherence	11%	23%	−12	0.070
Concision versus verbosity	20%	46%	−26	<0.001
Numerical and factual accuracy	48%	74%	−26	<0.001

The winning correlates are not the mirror image of the losing ones. LLMs lost on accuracy and on verbosity, and won on completeness, decision-readiness and analytical depth. Those are compatible properties: an output can be thorough, well organised and wrong. Leave-one-evaluator-out resampling across 15 folds holds every effect within six percentage points, so none rests on a single evaluator.

Study 2 grader notes. *Two hundred and fifty notes, 19,716 words, median 58, one per grader per output, from eight graders. Nine output-level codes were derived. One separates the gate outcomes after correction.*

Table G2: *Prevalence of criticism by code in grader notes on outputs that failed versus passed the critical-failure gate (n = 128 failed, 122 passed). Fisher exact, Holm-corrected across nine codes.*

Code	Failed	Passed	Diff	p (Holm)
S1 Wrong value, classification or decision	56%	28%	+28	<0.001
S9 Evidence cited from benchmark or output	25%	12%	+13	0.117
S3 Missing required element	36%	24%	+12	0.274
S7 Register, narrative and commentary	33%	25%	+8	1.00
S8 Professional or accounting judgement	19%	11%	+8	0.700
S2 Calculation mechanics	16%	11%	+5	1.00
S6 Navigability and visual clarity	41%	39%	+2	1.00
S5 Deliverable form and artefact fit	30%	29%	+1	0.783
S4 Unrequested or excess content	17%	20%	−3	1.00

Criticism of a wrong value, classification or decision is the only code whose prevalence separates gate-failed from gate-passed outputs. Every presentation-side code is flat, and criticism of excess content is marginally more common on outputs that passed. The gate detects wrongness, and the graders' own language locates that wrongness in the substance of the answer rather than in its presentation.