WHERE LLM GRADERS SUCCEED AND BREAK: EVIDENCE FROM TWO COMPUTER-SCIENCE EXAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

One long-form exam in a large course costs hundreds of grader-hours, and qualified graders are scarce; LLM graders are a tempting alternative. To show its pitfalls we grade a practical Computer Vision exam (570 dual-graded students) under 171 configurations spanning closed and open-weights models; the best reaches mean absolute error $1.64/35$, below the $2.61/35$ two human graders achieve against each other. The catch is the prompt: a short "strict grader" preamble drives 14 of 17 open-weights models out of the graded band (MAE ≥ 8), three stopping grading altogether. The damage traces to the preamble's two credit-withholding sentences, not to tone or model scale; one of them, "never give partial credit", alone makes two of three probed models stop grading. The closed flagships of three vendors shift calibration under it but stay in the band. In 162 further configurations on a second, independent Machine Learning exam from another course (1,038 dual-graded students), the preamble worsens ten models, moving three out of the band into collapse and one into refusal, yet improves seven whose neutral prompts over-mark: the vulnerability replicates, but its direction is exam-specific. Light LoRA fine-tuning repairs it: one adapter on the two exams' pooled $\sim 3{,}900$ graded examples brings five small open models to parity or better with a human grader in agreement with the grader pair, and sensitivity to the three harsh personas nearly vanishes (≤ 0.32 MAE). We release the anonymised dataset, full ablation grid, and grading, fine-tuning and analysis pipelines.

1 INTRODUCTION

Large language models (LLMs) are being used in computer-science education (Prather et al., 2023), grading included (Phung et al., 2023): hundreds of grader-hours per cycle and scarce qualified graders make the economic case. The methodological case is murkier: the usual metrics hide which configuration of model, prompt and persona grades usably, and how it survives an instructor's small prompt edits. On a practical Computer Vision (CV) exam, dual-graded by fixed grader pairs, we run 171 configurations with bootstrap CIs: closed and open-weight models (7B–480B), prompt components, personas, few-shot demonstrations, temperatures. Six full-cohort configurations, four Gemini and two GPT-5.5, match or beat the exam's inter-grader floor of $2.61/35$.

The central finding is an asymmetry: for closed models the logical prompt (reference solution, grading guidelines, rubric breakdown, neutral persona) is already at a joint optimum (Section 6), while open models are one sentence away from failure. A "strict grader" preamble knocks 14 of 17 models out of the graded band in all six tested open-weight families, three stopping grading entirely, the damage is unordered by scale and attributed to its two policy sentences, not its tone; across three closed vendors only Gemini's cheapest tier leaves the band, and only on the second exam (Sections 5 and 6.1). Failure takes three forms MAE alone cannot separate (zeroed submissions, blanket mark-downs, one structured field dismantled while the rest grades on); a conflicting-instruction account fits only the last, and only as a hypothesis (Section 5.3).

The collapse is not an artifact of one exam: under the same preamble, a 162-configuration replication on another Machine Learning (ML) exam from a different course worsens 10 of 17 models, three leaving the graded band and one refusing outright, while the best closed and open configurations again land below its floor. The direction does not carry over: seven models *improve*, the ML exam's over-marking cancelling against the persona, calibration masquerading as robustness (Section 5.4).

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.

The collapse trains away. One LoRA adapter on the two exams' pooled $\sim 3{,}900$ examples, supervised by per-question grader marks, takes five open models (4B–30B; Qwen, Llama, Gemma) from significantly worse than one human grader to parity or better under a paired third-grader test on *both* exams; an adapter trained on one exam already improves the unseen other. The persona sweep costing base models up to 20 MAE points moves the pooled adapters by at most 0.32 under the three harsh personas, and by at most 0.39 under *lenient* bar Qwen3-Coder-30B-A3B (up to 1.81; Section 8).

2 Related Work

LLMs as judges and graders. LLM judging is standard since Zheng et al. (2023) showed GPT-4 matches expert labellers on MT-Bench and Liu et al. (2023) formalised prompt-plus-chain-of-thought scoring. Programming auto-grading mostly unit-tests functional correctness (Messer et al., 2024), without rubric-aligned partial credit, or correlates LLM with human raters on ~ 100-submission datasets; the nearest benchmark, Phung et al. (2023), compares ChatGPT and GPT-4 with human tutors. We apply the same scaffolding to *student code grading* on two real, released exams ($n = 570$ and $1{,}038$) whose ground truth is two independent human graders, not a curated gold reference. Our few-shot arm supplies two worked examples per question labelled with D01's scores and rationales, not the graders' marks — in-context distillation of the best closed grader in the format of Brown et al. (2020) (Appendix D).

Prompt brittleness and personas. Sclar et al. (2024) show that semantically meaningless formatting changes shift benchmark performance by tens of percentage points and, with Mizrahi et al. (2024), urge reporting distributions over prompt variants; Deshpande et al. (2023) report persona-dependent toxicity rises up to $6\times$ in ChatGPT. Our perturbation instead carries meaning, and we replay it on a second exam, where the vulnerability replicates but its direction and failure mode do not (Section 5.4).

Instruction hierarchies and conflicts. Section 5.3's failure mode — abandoning one of two incompatible instructions rather than negotiating them — connects to instruction arbitration. Models treat instructions as equally privileged unless trained on an explicit hierarchy (Wallace et al., 2024), fail a non-trivial fraction of simple verifiable instructions (Zhou et al., 2023), break stated rules even on straightforward tests (Mu et al., 2023), degrade sharply when instructions conflict (Zhang et al., 2025), and rarely flag the conflict (He et al., 2026). Here the conflict arises from an ordinary instructor edit, not an attack, with a clean behavioural signature: the structured score channel surrenders while the prose channel keeps following the rubric.

3 Experimental Setup

Exams, graders and grid. 570 students took the CV exam's four code questions (**Q1** transfer learning, **Q2** CNN-from-scratch classification, **Q3** semantic segmentation, **Q4** bonus colorization); **Q1**–**Q3** form the 35-point base scale (weights $12/11/12$) used throughout; bonuses graded but excluded. Two of 20 graders in 10 fixed pairs (≈ 57 students each) grade every submission independently; the grader-average total is the ground truth (Section 4). The ML exam — another course, $1{,}038$ students, three questions, ≈ 65-point scale, dual-graded by 49 graders in non-fixed pairs — carries a 162-configuration replication (Section 5.4, Appendix H). We evaluate 4 Gemini models (Gemini Team et al., 2023; Comanici et al., 2025) via the Vertex API, GPT-5.5, GPT-5.4 and Claude Opus 5 via their batch APIs (Section 6.1), and 17 open-weights variants from six vendors, from 7B to 480B, dense and mixture-of-experts, served locally with vLLM (Kwon et al., 2023) on A100 nodes. The 171 CV configurations (33 closed, 138 open-weights) span per-model baselines, prompt-component removals, thinking, five-persona sweeps, few-shot prompting, temperature and mechanism probes (Section 5.2); Appendix A documents roster, serving stack, reproducibility, thinking configuration and exclusions; Appendices I and K tabulate every run.

The prompt. The default prompt supplies the question, its rubric, the reference solution, the student notebook, grading guidelines and a breakdown of each scorable item; the model returns a score, any bonus and a rationale. Each question is graded in its own call (three or four per student) with only its own notebook, serialised cell by cell; prompts run to $\approx$ 22k characters on the CV exam, ≈ 16k

on the ML exam, half to two thirds shared reference material. Personas are short preambles; *strict* reads "*You are a HARSH teaching assistant. Award the MINIMUM defensible score for any task that is incomplete, buggy, or deviates from the rubric. Never give partial credit if the task does not run correctly.*", the *neutral* default "*You are a strict but fair teaching assistant.*", with *lenient*, *rigorous* and *exacting* analogous.

Metrics. The headline metric is MAE of the AI total against the grader-average total, with 95% percentile bootstrap Confidence Intervals (CIs) (2,000 resamples) (Efron, 1979; Efron & Tibshirani, 1994), alongside the *mean signed error* (bias); the human floor is computed identically, MAE $= \mathbb{E}[\,|G_1 - G_2|\,]$ over the grader pair, and both are recomputed unchanged on the ML exam's own scale. Because a total can hide per-question errors of opposite sign, Appendix F decomposes the bias per question for the headline runs and Table 2 (Appendix A) gives the item-level Spearman ρ per question alongside the total. Students whose grading call permanently failed are excluded from that run's metrics — none in most runs, at most 3 elsewhere, exceptions in Appendix A.

4 The Human-Grader Floor

Across all 570 dual-graded submissions of the CV exam the two graders' totals agree to MAE $= 2.61/35$ (95% bootstrap CI $[2.37, 2.85]$; Pearson $r = 0.868$). We treat 2.61 as the human-grader floor: an AI configuration whose MAE sits at or below it makes, in aggregate, no more error against the grader average than the two graders make against each other. The floor is comparative only, and leans in the AI's favour because averaging two graders cancels part of their noise. Beating it yields a parity claim of the kind Section 8.2 retests with a paired third-grader test. The floor pools across pairs whose internal disagreement varies by $4.2\times$: per-pair MAE runs 0.85 to 3.55 over the 10 fixed pairs, so which pair a student draws is a source of variability (Appendix G).

5 The Strict-Persona Collapse

5.1 The collapse, quantified

Table 3 (Appendix B) pairs each model's *strict* and *neutral* runs, identical bar the persona sentence; Figure 2 renders all 51 strict-flavoured runs. MAE alone cannot separate the failures (zeroing every student scores MAE ≈ 26, the distance from the grader mean), so we add a *behaviour* class: a **refusal** zeroes at least 90% of students with awarded-total standard deviation below 0.5, a *near-refusal* keeps some variation at that zero-rate, a **collapse** reaches $3.07\times$ its own exam's inter-grader floor while still discriminating between students, and anything else is *graded*. The CV exam's MAE ≥ 8 fixes that multiple (3.07×2.61) and gives MAE ≥ 15.7 on the ML exam (3.07×5.13), the same severity on both scales.

Every open-weight family breaks. Prepending a short "strict grader" preamble to the otherwise-best prompt takes 14 of the 17 config-matched pairs out of the graded band, all six families represented, MAE rising $\times 1.69$ to $\times 7.11$. Three models stop grading outright: Llama-3.1-8B and Mistral-Small-24B award $0.00/35$ on average, GLM-4-9B 0.56; Qwen2.5-Coder-32B still marks students (MAE $= 20.29$, bias -20.28, mean awarded $5.75/35$). No Gemini configuration collapses under any wording here (Appendix B): `gemini-3.1-pro-preview` moves $1.86 \rightarrow 2.75$ under *strict* (F03, still at the floor), Flash-Lite's strict-flavoured runs sit at 3.73–5.75 — an imperfect calibration knob, not a failure mode — and GPT-5.5, GPT-5.4 and Claude Opus 5 stay in band (Section 6.1).

Scale does not order the damage. Parameter count and the strict/neutral MAE ratio correlate at $\rho = -0.21$ ($p = 0.42$) across the 17 sized models. The two largest ratios belong to a 24B model that stops grading and a 14B that zeroes 88% of submissions, while 106B GLM-4.5-Air collapses to 21.02; GLM *improves* from 9B to 32B before worsening at 106B while Gemma *worsens* from 12B to 27B — opposite directions over matched rungs. Robustness does appear at the top, Qwen3-235B-A22B and Qwen3-Coder-480B staying in band, but both are Qwen, the only family with rungs above 106B, confounding family with scale exactly where it matters; the third in-band model, Qwen2.5-Coder-7B, clears the threshold by 0.08 from an already-weak baseline.

The damage is specific to one preset. Two of 13 models leave the band under *rigorous* and two under *exacting*, against 14 of 17 under *strict* (Table 4, Appendix B); only *strict* carries the two policy sentences, so the gap is about policy content, not vocabulary.

5.2 ATTRIBUTION: THE POLICY SENTENCES, NOT THE ADJECTIVE

The *strict* preset is three sentences: a frame ("You are a HARSH teaching assistant.") and two policy sentences — **S1**, "Award the MINIMUM defensible score for any task that is incomplete, buggy, or deviates from the rubric.", and **S2**, "Never give partial credit if the task does not run correctly." Varying the adjective alone — STRICT, RIGOROUS, FAIR — over verbatim S1 and S2 moves MAE by 0.94 to 2.15 across five models, in no consistent direction (Table 5, Appendix B); **FAIR** with the harsh policy leaves Mistral-Small-24B at 25.05, still not grading.

A 2×2 crosses S1 with S2 on three models, frame fixed: the frame alone moves each less than two points off neutral, both sentences take them to 20.26–26.04, S2 alone to outright refusal on Llama-3.1-8B and Mistral-Small-24B. S1 is not benign — alone it takes Llama-3.1-8B to 23.45 and outdoes S2 on Qwen2.5-Coder-32B (15.55 against 12.92) — so dominance varies by family: S1 for the 32B, S2 for the other two, on both exams, where S2 again alone stops graders grading (Appendix H). S2 contradicts the partial-credit scale the rubric scaffold mandates — hence the safer wordings, and advice about one instruction, not a tone.

5.3 MECHANISM: THREE FAILURE MODES, ONE MAE RANGE

Three explanations fit the headline numbers: refusal, calibration drift, conflicting instructions ("be strict" vs. "award partial credit per rubric"). All three occur, on different models, indistinguishably by MAE: Gemma-3-27B reaches MAE $= 11.55$ having zeroed Q1 on *zero* of 570 students, GLM-4-32B a lower 9.66 by zeroing 273. One question separates them — of students scored 0 on Q1, what fraction received a non-zero Q2? — and it splits the 19 matched strict runs into **blanket zeroing** (five), **selective field collapse** (five) and **uniform severity** (nine), both thresholds in empty bands (Table 6, Appendix B).

In the selective group the prompt's two output channels disagree: on the two hand-annotated runs, 20% and 43% of zeroed submissions carry a rationale itemising partial credit while the score field reads 0 (Appendix B) — one channel obeys S2, the other the rubric breakdown. But this covers neither other group, and its clearest prediction fails: withdrawing the breakdown *worsens* the collapse in five of five families against a clean neutral control (Appendix B). An alternative fits that failure — the breakdown may anchor the score field to non-zero sub-totals, so withdrawing it lets the credit-withholding sentence run unopposed — and our grid cannot separate the two; we treat the mechanism as classified rather than explained (Section 10).

5.4 THE COLLAPSE REPLICATES ON THE ML EXAM; ITS DIRECTION DOES NOT

We replicate it on the ML exam (Section 3; floor 5.13 computed identically) over 162 configurations: neutral and *strict* runs for all 17 open-weights models, five-persona sweeps, attribution probes, a closed-model arm (Appendix H). Under *strict* four models cross from the graded band to above that exam's (MAE ≥ 15.7); Llama-3.1-8B again refuses. A fixed MAE ≥ 8, only $1.56\times$ this floor, would instead put ten strict runs out of band and eleven *neutral* baselines with them, against one here (DeepSeek-Coder-V2-Lite, 16.41). The collapse is thus no artifact of one rubric, cohort or scale — nor of open weights alone: Flash-Lite worsens under *strict* ($7.53 \rightarrow 9.02$) and leaves the band under *lenient* (17.86), while `gemini-3.1-pro-preview` stays below the floor ($3.40 \rightarrow 3.67$, IG23).

The *direction* does not transfer (Figure 1): seven of 17 improve under the same sentence — calibration, not robustness. The ML exam's neutral prompt over-marks for 15 models, and across the 16 non-refusal pairs neutral bias predicts the strict effect at $r = -0.73$ (permutation $p = 0.002$); on the CV exam, whose neutral biases straddle zero (none above $+1.68$), zero of 17 improve. The decomposition also splits the two big Qwens' apparent robustness: the 480B over-marks by 9.52 points at neutral (MAE $= 10.10$, twice its floor) and is pulled back by the sentence; the 235B is the one model above 106B in band under both personas on both exams. Parity replicates on both sides (`gemini-3.1-pro-preview` at $0.66\times$ the ML exam's floor, Qwen3-Coder-Next at 4.47), and the failure mode wanders: Qwen2.5-Coder-32B, selective here, zeroes whole submissions there.

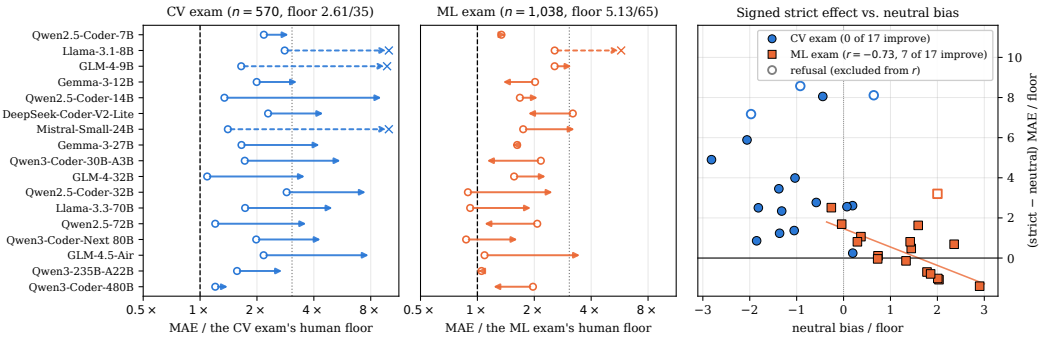

Figure 1: The strict persona on the 17 models run on both exams, floor-relative. *Left, centre:* neutral (circle) → strict MAE; dashed = floor, dotted = each exam's band ($3.07\times$ floor: MAE ≥ 8 on the CV exam, ≥ 15.7 on the ML exam), dashed arrows = refusals (ceiling MAE). *Right:* the signed strict effect $(\text{MAE}_{\text{strict}} - \text{MAE}_{\text{neutral}})/\text{floor}$ against neutral bias, negative exactly where the persona cancels over-marking.

One preamble, two remedies. One preamble separates competitive open-weights graders (Section 7) from 14 out of band, a gap single-prompt evaluation cannot see (Sclar et al., 2024; Mizrahi et al., 2024), with two remedies: never forbid partial credit, or fine-tune lightly on in-house labels (Section 8).

6 Closed-Model Configurations

On the closed side the baseline recipe sits at the joint optimum: no prompt-component removal reliably improves on it.

Six full-cohort configurations at or below the floor. Six full-cohort configurations meet or undercut the human floor of 2.61 with CI upper bounds below it (Table 10, Appendix C): D01 (`gemini-3-flash-preview`, neutral, $t = 0$) at MAE $= 1.64$ (bound 1.79); F02 and D02 (the 3.1-pro under *lenient* and *neutral*); B03 (Flash-Lite with thinking), an order of magnitude cheaper; and `gpt-5.5` under *lenient* and *neutral* (O03 2.25, O01 2.43; Section 6.1). The floor comparison is not like-for-like (Section 4), so we retest all six with the paired third-grader statistic of Eq. (1) (Table 11): three are better than a human grader — D01 $\bar{d} = -0.54$ $[-0.71, -0.36]$, F02 -0.37 $[-0.56, -0.18]$, D02 -0.37 $[-0.55, -0.18]$ — while B03 (-0.19 $[-0.39, +0.01]$) and both `gpt-5.5` runs (-0.01 $[-0.22, +0.21]$ lenient, $+0.19$ $[-0.02, +0.41]$ neutral) are at parity. The Flash-Lite baseline A01 is significantly worse ($+1.03$), as are `claude-opus-5` ($+1.13$) and all 17 open-weights baselines' closest analogues (Section 7). Within Gemini the newer Flash beats the older Pro (D03, 4.10); D01's scatter (Figure 3) is essentially unbiased, its residual error concentrated in the low-score band where the graders also disagree more.

On Flash-Lite, thinking helps and dropping the solution hurts. The B-series varies one prompt component at a time on Flash-Lite (Table 12). Dropping the reference solution flips a small under-grading bias (-1.78) into over-grading ($+3.30$) costing ≈ 0.8 MAE; dropping the guidelines or breakdown moves MAE by at most 0.12; thinking (B03) is the one large move, closing about three-quarters of the gap to D01 on a far cheaper model. Two combination runs cross a persona with another component: F01 (Flash-Lite, *strict* + thinking) at 3.73 recovers most of the *strict* penalty (5.75); F02 (3.1 Pro + *lenient*) reaches 1.79 against 1.86 neutral (Table 24). The G-series ablations (`gemini-3-flash-preview`, first 100 students; Appendix C) beat D01 only as a sample-size artifact; the $t = 0$ variance probes are in Appendix C.

6.1 The asymmetry holds across vendors

Gemini is one vendor, so we replayed the persona sweep on two more via batch APIs — same prompt bytes, thinking off, temperature 0 where exposed (Appendix C, Table 9): OpenAI's `gpt-5.5` and

`gpt-5.4` under *neutral*, *strict* and *lenient* on both exams, Anthropic's `claude-opus-5` under *neutral* on both and *strict* on the CV exam. None leaves the graded band under *strict* — `gpt-5.5` $2.43 \to 4.43$ (CV exam) and $3.54 \to 3.70$ (ML exam), `claude-opus-5` $3.54 \to 5.17$, `gpt-5.4` $4.69 \to 6.90$ and $4.30 \to 5.11$, zero-total rates at most 0.7%, no refusal — the calibration shift the Gemini models show ($1.86 \to 2.75$, $3.34 \to 5.75$), against 14 of 17 open-weights models leaving it under the identical sentence. The cheaper tiers are the fragile ones: `gpt-5.4` is worst under *lenient* on the ML exam ($4.30 \to 8.30$, bias $+7.94$) but stays in band, where Flash-Lite reaches 17.86 and leaves it; on the CV exam *lenient* nearly halves its error ($4.69 \to 2.69$) by cancelling a -4.45 under-marking — calibration masquerading as robustness (Section 5.4), now on a closed model. At neutral `gpt-5.5` is below the floor on both exams, `claude-opus-5` on the ML exam only (4.37).

7 Open-Weights Configurations Under Neutral Prompts

Under neutral prompts nothing stands out: the 17 baselines span 2.85 to 7.49 MAE, no family dominating or collapsing (Table 14, Appendix D); none of it predicts which stop grading under a persona sentence. Scaling is not monotone: Qwen2.5-Coder improves from 7B to 14B ($5.68 \to 3.51$) then regresses at 32B (7.49, bias -7.35); dropping the reference solution recovers 3.12, consistent with over-anchoring on it, though exam-locally: the same removal moves it $+0.16$ on the ML exam. Within Qwen3-Coder the 30B-A3B MoE beats the 80B Coder-Next (4.50 vs. 5.17), reversing at 480B (3.14).

GLM-4-32B takes four of the five best open-weights slots. Its best (rubric breakdown removed, neutral; L-BD07), 2.68, is the grid's strongest open-weights grader (Table 24); L-X04 (*rigorous*) reaches 2.74, essentially unbiased; the fifth slot, Qwen3-Coder-Next + *lenient* (2.81) is a bias-correction artifact, not capability. Two gaps to D01 (1.64): just over one point on best achievable, 1.42 on best surviving a persona sentence (Qwen3-Coder-480B under *rigorous*, 3.06; L-R04), since GLM-4-32B collapses to 9.66 under *strict*.

Model choice beats size. The 30B-A3B beats the 80B, the 14B the 32B, the 72B dense the 235B MoE; the 480B's edge is persona robustness: at neutral it only ties the 72B. *Which* model is a property of the exam, not the model: across the 17 run on both, neutral MAE is uncorrelated ($\rho = -0.07$) and Qwen2.5-Coder-32B, weakest here, ties statistically for best there. Open-weights graders are practically viable, but only zero-shot on a prompt guaranteed never to carry a strict-flavoured persona (Section 5), which Section 8 removes with in-house labels.

8 Fine-Tuning Repairs the Strict-Persona Collapse

So far we have changed only the prompt. We have better material than prompts: two humans graded every submission in both exams. In this section we use their grades to fine-tune the models with LoRA (Hu et al., 2022). We fine-tune five small open models (4B to 30B). Each model gets three adapters: one per exam, and one trained on both exams pooled (about 3,900 examples). Three things happen. With the pooled adapter, every model grades as well as a human or better, on both exams. The skill transfers: an adapter trained on one exam also gets better at the other, without ever seeing it. And the persona problem is gone. The prompt that adds as much as 20 MAE points to a base model barely moves the fine-tuned one.

8.1 Setup

We fine-tune five open models across three families and sizes from 4B to 30B parameters with LoRA: Qwen2.5-Coder-7B and 14B (Hui et al., 2024), Qwen3-Coder-30B-A3B (Yang et al., 2025b), Llama-3.1-8B (Grattafiori et al., 2024), and Gemma-4-E4B (Gemma Team et al., 2026). Training hyperparameters are in Appendix E. Each exam uses a seed-fixed $80/20$ student split: $456/114$ on the CV exam, giving 1,494 training examples, and $830/208$ on the ML exam, giving 2,413. The ML exam records one grade per question with the bonus folded in, so we split each grade into a score up to the base cap and the remainder as bonus. We train three adapters per model: one per exam and one pooled. The pooled training set is the union of the two train splits, so no adapter ever sees a held-out

student. We evaluate all three, plus the untuned base, on both held-out sets. On the held-out students, the two human graders disagree with each other by 2.83 marks on average on the CV exam and 5.19 on the ML exam. These are the floors: a grader that disagrees with the humans by less than this agrees with them better than they agree with each other. We use two training recipes with identical prompts. The *marks* recipe averages the two graders' scores into a single target, one example per student and question, with a {score, bonus} output format. The *bd* (breakdown) recipe additionally distils a per-task breakdown from Gemini 3 Flash Preview (run D01), without its reasoning text.

8.2 One adapter reaches human parity or better on both exams

Every untuned base grades worse than a human on both exams and mis-calibrates in opposite directions: Llama-3.1-8B reads 19.8 on the CV exam where the grader average is 25.3, yet 40.5 on the ML exam where it is 30.9. With a score standard deviation of 2.1, it barely distinguishes the CV exam's students at all. The pooled fine-tune produces better results on both exams: it lands at 1.75–2.01 MAE on the CV exam and 3.29–3.53 on the ML exam (marks recipe; Table 1), revives the collapsed score spread on the CV exam (σ from $2.1 \rightarrow 7.3$–7.5, against 8.0 for the grader average), and replaces the bases' bias of -5.9 to $+9.5$ marks with under one mark of generosity ($+0.2$ to $+0.9$) on both exams.

Because MAE against a two-grader *average* is easier than any single grader faces (Section 4), the verdicts come from a paired third-grader test:

$$d_i \;=\; \tfrac{1}{2}\big(|\mathrm{AI}_i - \mathrm{G}_{1,i}| + |\mathrm{AI}_i - \mathrm{G}_{2,i}|\big) \;-\; |\mathrm{G}_{1,i} - \mathrm{G}_{2,i}|. \tag{1}$$

Pairing removes per-student exam difficulty; averaging d over the held-out students with a 95% CI, an interval containing 0 means the AI is statistically indistinguishable from a human grader, entirely below 0 that it disagrees with the humans *less than they disagree with each other*, entirely above, worse. Below zero is possible because $|\mathrm{G}_1 - \mathrm{G}_2|$ carries *two* graders' idiosyncratic noise while $|\mathrm{AI} - \mathrm{G}_i|$ carries one human's plus the model's; the AI beats the floor exactly when its own noise is smaller than a single human's. With no ground truth beyond the graders, *better* here means more reliable, not more correct: added to the grading pool, the model would introduce less disagreement than another human.

All base configurations are significantly worse than a human grader on both exams. The pooled marks adapter is significantly *better* in 9 of 10 (model $\times$ exam) cells: all five models on the ML exam ($\bar{d}$ from -0.94 to -1.13, every CI below zero) and four of five on the CV. The pooled bd adapter reaches at least parity in all 10 cells and is better in two. No cell on either exam reads worse (Table 19, Appendix E). The ML exam sets a lower bar because its two graders disagree more (floor 5.19 vs. 2.83), but every point estimate there is negative regardless.

8.3 Grading skill transfers across exams

An adapter fine-tuned on one exam improves grading on the other, which it never saw: mean MAE on the ML exam falls from 7.95 (base) to 5.36 under the CV-exam adapter, and on the CV exam from 5.54 to 3.57 under the ML-exam adapter. Every model improves in both directions, with one exception discussed in Section 8.5. Transfer stops short of in-domain performance (3.54 and 1.90), but pooling closes the gap for free: the pooled adapter matches or beats each specialist *on the specialist's own exam* in all eight (exam $\times$ recipe) columns (Table 1).

8.4 Fine-tuning immunises against the strict-persona collapse

We re-run the persona sweep for base and pooled models on both exams (Tables 16 and 18, Appendix E). The bases reproduce the collapse on both exams: under *strict*, Qwen2.5-Coder-14B goes $4.06 \rightarrow 23.81$ on the CV exam, and Llama-3.1-8B reaches MAE 30.9 on the ML with bias -30.9: it scores nearly every student zero. The pooled adapters are flat everywhere: at most $+0.32$ (marks) and $+0.62$ (bd) across three personas, two exams and five models, and the catastrophic tail the collapse creates is gone: under a neutral prompt the worst base mis-grades 43 of 114 CV-exam students by more than 10 marks, while every pooled model is at 0–1. On the CV exam the *strict* preset is the most destructive on every base, as Section 5.2 predicts: it alone carries the two policy sentences. The per-exam adapters had one immunity failure, and pooling fixes it: Llama-bd under *strict* sat at 4.08 when tuned on the CV exam alone and drops to 2.67 with pooled training.

Table 1: Held-out MAE vs. the grader average by training set (marks recipe; bd in Table 17, Appendix E). The cross columns are transfer; pooled matches or beats each specialist on its own exam. ‡: the one negative transfer cell (Section 8.5).

	CV exam (n=114)				ML exam (n=208)			
Model	base	CV	ML	pooled	base	CV	ML	pooled
Qwen2.5-Coder-7B	6.60	1.93	3.22	**1.84**	6.48	6.37	3.67	**3.40**
Qwen2.5-Coder-14B	4.06	1.89	3.29	**1.89**	7.70	5.41	3.35	**3.29**
Qwen3-Coder-30B-A3B	4.76	1.80	2.59	**1.75**	8.60	5.22	3.67	**3.45**
Gemma-4-E4B	4.29	1.88	3.79	**1.88**	4.66	4.79‡	3.51	3.53
Llama-3.1-8B	7.99	1.99	4.95	**2.01**	12.30	5.02	3.48	**3.37**
mean	5.54	1.90	3.57	**1.87**	7.95	5.36	3.54	**3.41**

On the ML exam *strict* sometimes improves a base model (Qwen3-Coder-30B: $8.60 \rightarrow 6.34$). This is not the persona working. The bases over-grade the ML exam, and the induced harshness happens to cancel part of the bias, the accidental calibration of Section 5.4 again. So the same wording that destroys a model on one exam helps it on the other. An effect that flips sign between exams is not a calibration knob.

8.5 What matters, what does not

Size and family remain nearly irrelevant. The five pooled marks fine-tunes converge to 1.75–2.01 MAE on the CV exam and 3.29–3.53 on the ML, and starting quality does not predict the tuned result: Llama-3.1-8B starts worst on both exams (7.99, 12.30) and finishes with everyone else (2.01, 3.37). Human labels alone suffice: pooled marks matches or beats pooled bd on every (model, exam) cell, decisively on the ML exam (3.41 vs. 4.15 mean). Nine of the eleven better verdicts in the paired test are marks verdicts, and marks needs no closed-model teacher. Which exam the labels come from matters less than having labels at all, with one exception. Gemma-4-E4B is the strongest base on the ML exam (4.66) and the only model that cross-exam transfer fails to help there ($4.79^{\ddagger}$), and in-domain labels move it only to 3.51: fine-tuning buys the most where the base grades worst, and Gemma had the least room to improve. The residual errors concentrate on the students the two human graders also disagree on (correlation $+0.35$ to $+0.48$ for every pooled model, against no consistent relationship for the bases; Appendix E): the tuned models are unsure exactly where the rubric is ambiguous.

9 Discussion

What the collapse is, and is not. Of the three failure modes only selective field collapse looks like a failure to negotiate conflicting instructions, and only as a hypothesis (Section 5.3); the others are calibration drift and refusal, which never appears on a Qwen model, so a single-family roster misses that mode. Spanning all six families unordered by scale, the failure belongs to open-weights instruction following: no closed model of the three vendors leaves the band on either exam, and robustness above 106B is half illusory once the ML exam decomposes it. Fine-tuning shows the failure is zero-shot, not small-model.

Why single-prompt evaluation misses it. One well-behaved prompt leaves the best open-weights configuration just over one MAE point behind Gemini's best (2.68 vs. 1.64), one instructor-style persona preamble nearly fourteen behind (15.58): single-prompt evaluation would have told a sunnier story (Sclar et al., 2024; Mizrahi et al., 2024).

Implications for deployment. **(1) Never instruct a grader to withhold credit**: both policy sentences collapse graders on their own (Section 5.2), unordered by parameter count, and under-grading harms students who attempted the work in good faith. Verify the model against the prompt *and the exam*: on the ML exam the same preamble helps seven of 17 models purely by cancelling their over-marking, yet nearly doubles the error of the ML exam's best. **(2) With in-house dual-graded examples, fine-tune rather than prompt-engineer — and pool the courses.** One adapter

on the two exams' pooled $\sim 3{,}900$ examples brings every 4B–30B model to parity or better with a human grader on both exams and removes the collapse on both (persona drift ≤ 0.32 MAE), with no closed-model teacher and no cost to either course. **(3) Evaluate several prompt variants**, including instructor-written ones. **(4) Report bias alongside MAE**: Flash-Lite's lenient and strict have comparable MAE with opposite-signed bias ($+4.01$ vs. -5.44): MAE alone does not show which way the harm runs.

10 LIMITATIONS

Coverage. Both exams are from one university: replication spans courses, cohorts, rubrics and scales, not institutions. Fine-tuning's two transfer directions are not size-matched (456 vs. 830 training students), confounding their comparison; Gemma-4's ML-exam baseline, unlike the other four, lacks independent corroboration. We varied three strict-flavoured wordings and two preset policy sentences, not every credit-withholding instruction. Both models above 106B are Qwen, entangling scale with family where it matters (Section 5.4 separates their robustness, not that confound). The `gemini-3-flash-preview` ablations used the first 100 students (matched-subset baselines, wider CIs). Closed coverage: four Gemini models; three personas on two OpenAI tiers and one Anthropic model (Section 6.1; Anthropic *strict* on the CV exam only), no *rigorous* or *exacting* outside Gemini. Grading is static: nothing is executed (Appendix A; no test suites, minutes-to-hours GPU cost, no stored outputs), so "does not run" verdicts are unaudited and their error rate on valid but unconventional code unmeasured. We know of no public corpus of long-form notebooks with dual human rubric grades for external benchmarking; automated-grading corpora are unit-test based (Messer et al., 2024) or score short per-task programs (Phung et al., 2023).

ML-exam ground truth. Nine of the ML exam's 1,038 rows record 0.0 for one grader against real marks from the partner; they are retained in the headline floor but excluded from the maximum-disagreement figure; excluding them everywhere moves no behaviour class, no table ordering and no held-out verdict (Appendix H). Its pairing is hybrid rather than fixed, so per-pair statistics are scoped to the 18-pair backbone covering 79% of the cohort.

Few-shot demonstrations. The K-series demonstrations carry D01's labels, and their five students remain in the evaluated cohort ($5/570$), graded with their own worked answer in the prompt. The ML exam's demonstrations likewise carry IG08's labels, with three students remaining; its Flash-Lite arm suffers biased dropout from prompt-length failures, so only its matched subsample is comparable (Appendix H).

Claims. Section 5.3's three-way split classifies rather than explains: withdrawing the rubric breakdown contradicts the conflicting-instruction reading, leaving the causal story open. The few-shot uplift is capability, not held-out generalisation: the demonstrations carry D01's own labels. One seed-fixed split and one run per configuration make the paired third-grader test load-bearing; its normal-approximation CIs leave three CV-exam *better-than-grader* verdicts marginal: upper bounds -0.044 (14B marks), -0.020 (Gemma marks) and -0.001 (7B bd), all strictly below zero under a 2000-resample percentile bootstrap, though the weakest would contain zero under a t-interval ($p = 0.052$ paired, 0.069 Wilcoxon); the ML exam's are not (upper CIs ≤ -0.30). Appendix J details the remaining caveats.

11 CONCLUSION

For a dual-graded exam the recipe is short. The stronger closed models (Gemini 3 Flash and 3.1 Pro, GPT-5.5) grade at or below the human floor untuned, from a neutral prompt; an open model gets there only with light fine-tuning on the exam's own grades, and only then survives the instructor edits that break it zero-shot. Between exams the vulnerability and its repair transfer; the model ranking, the persona's direction and $t = 0$ determinism do not. So verify a grader on the exam it will grade, under an instructor's own prompts.

AI USE STATEMENT

We used generative AI tools (Opus 4.8, Opus 5.0 and Fable 5) as coding and writing assistants throughout this work. Specifically: for implementing and refactoring the data pipeline, the grading harness, the fine-tuning scripts, and the analysis code; for drafting and editing prose in this paper; and for exploratory literature search. We did not use generative AI to generate research ideas, to produce experimental results, or to write any of the analysis numbers reported here — every number in this paper is computed by the released scripts from the released data, and the ledgers those scripts emit are the authority for the manuscript. All AI-assisted code was reviewed and executed by the authors, and the statistical results were independently re-derived before being reported. We take responsibility for the final content of this work, including all text, claims, and artifacts.

Note that generative models are also the *object* of study here rather than a tool: the LLM graders evaluated in Sections 6–8 produced the scores we analyse, and those outputs are released in full.

REPRODUCIBILITY STATEMENT

Every number in the paper is regenerated from released artifacts by released code. The two anonymised exam datasets, the rubrics and reference solutions, one result workbook per configuration, and the analysis, grading, and fine-tuning pipelines accompany the submission as supplementary material. Section 3 and Appendix A specify the prompt, the serving stack, the decoding parameters, and the exclusion rules; Appendix I tabulates all 171 CV-exam configurations with their bootstrap confidence intervals; Appendix E gives the fine-tuning hyperparameters and the seed-fixed split. Training uses HuggingFace with PEFT and every evaluated number is served by vLLM, Gemma-4's adapter included. One caveat is stated where it arises: evaluation is not bitwise reproducible because vLLM batching reorders reductions (~ 0.01 MAE); Gemma-4's vLLM serving path was cross-validated against HuggingFace generation to within ~ 0.1 MAE (Appendix E).

ETHICS STATEMENT

The data are student exam submissions and grader marks from two university courses, released with the course instructors' permission. Both datasets are anonymised before release: student identifiers, names, and email addresses are removed from notebook code, markdown, outputs, and file metadata, and graders appear only as opaque identifiers. No demographic attributes are collected or released, and no student is identifiable from the released artifacts.

The application this paper studies carries real risk to the people being graded, which is why we report the failure mode rather than only the headline accuracy. An under-grading collapse of the kind documented in Section 5 harms students who attempted the work in good faith, and it is invisible to single-prompt evaluation. We therefore recommend against deploying an LLM grader on the basis of one prompt's measured accuracy, and we report bias alongside MAE throughout so that the direction of harm is visible. We regard human oversight of consequential grading decisions as a requirement, not an option, and none of the results here should be read as licensing unsupervised automated assessment.

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 9079–9093, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25cc3adf8c85f7c70989cb8a97a691a7-Paper-Conference.pdf.

Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

DeepSeek-AI et al. DeepSeek-Coder-V2: Breaking the barrier of closed-source models in code intelligence, 2024. URL https://arxiv.org/abs/2406.11931.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in ChatGPT: Analyzing persona-assigned language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1236–1270, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.88. URL https://aclanthology.org/2023.findings-emnlp.88/.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Bradley Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, May 1994. ISBN 9780429246593. doi: 10.1201/9780429246593. URL https://doi.org/10.1201/9780429246593.

Gemini Team et al. Gemini: A family of highly capable multimodal models, 2023. URL https://arxiv.org/abs/2312.11805.

Gemma Team et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Gemma Team et al. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

GLM-4.5 Team et al. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models, 2025. URL https://arxiv.org/abs/2508.06471.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Xingwei He, Qianru Zhang, Pengfei Chen, Guanhua Chen, Linlin Yu, Yuan Yuan, and Siu-Ming Yiu. ConInstruct: Evaluating large language models on conflict detection and resolution in instructions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):30969–30977, March 2026. doi: 10.1609/aaai.v40i37.40356. URL https://doi.org/10.1609/aaai.v40i37.40356.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Binyuan Hui et al. Qwen2.5-Coder technical report, 2024. URL https://arxiv.org/abs/2409.12186.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pp. 611–626. ACM, October 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

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 Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Marcus Messer, Neil C. C. Brown, Michael Kölling, and Miaojing Shi. Automated grading and feedback tools for programming education: A systematic review. *ACM Transactions on Computing Education*, 24(1):1–43, February 2024. doi: 10.1145/3636515. URL https://doi.org/10.1145/3636515.

Meta. Llama 3.3 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md, December 2024. 70B Instruct released 6 December 2024.

Mistral AI. Mistral-Small-24B-Instruct-2501. Hugging Face model card, https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501, January 2025.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tacl_a_00681. URL https://aclanthology.org/2024.tacl-1.52/.

Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Basel Alomair, Dan Hendrycks, and David Wagner. Can LLMs follow simple rules?, 2023. URL https://arxiv.org/abs/2311.04235.

Aditya Pathak, Rachit Gandhi, Vaibhav Uttam, Arnav Ramamoorthy, Pratyush Ghosh, Aaryan Raj Jindal, Shreyash Verma, Aditya Mittal, Aashna Ased, Chirag Khatri, Yashwanth Nakka, Devansh, Jagat Sesh Challa, and Dhruv Kumar. Rubric is all you need: Improving LLM-based code evaluation with question-specific rubrics. In *Proceedings of the 2025 ACM Conference on International Computing Education Research V.1*, ICER '25, pp. 181–195. ACM, 2025. doi: 10.1145/3702652.3744220. URL https://doi.org/10.1145/3702652.3744220.

Tung Phung, Victor-Alexandru Pădurean, José Cambronero, Sumit Gulwani, Tobias Kohn, Rupak Majumdar, Adish Singla, and Gustavo Soares. Generative AI for programming education: Benchmarking ChatGPT, GPT-4, and human tutors. In *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 2*, ICER 2023, pp. 41–42. ACM, 2023. doi: 10.1145/3568812.3603476. URL https://doi.org/10.1145/3568812.3603476.

James Prather, Paul Denny, Juho Leinonen, Brett A. Becker, Ibrahim Albluwi, Michelle Craig, Hieke Keuning, Natalie Kiesler, Tobias Kohn, Andrew Luxton-Reilly, Stephen MacNeil, Andrew Petersen, Raymond Pettit, Brent N. Reeves, and Jaromir Savelka. The robots are here: Navigating the generative AI revolution in computing education. In *Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education*, ITiCSE 2023, pp. 108–159. ACM, December 2023. doi: 10.1145/3623762.3633499. URL https://doi.org/10.1145/3623762.3633499.

Qwen Team. Qwen3-Coder-Next technical report. Technical report, Alibaba Cloud, 2026. URL https://github.com/QwenLM/Qwen3-Coder/blob/main/qwen3_coder_next_tech_report.pdf. Model card: https://huggingface.co/Qwen/Qwen3-Coder-Next.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6c0e99d736da621403018ca7b32b1a4d-Paper-Conference.pdf.

Team GLM et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools, 2024. URL https://arxiv.org/abs/2406.12793.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions, 2024. URL https://arxiv.org/abs/2404.13208.

An Yang et al. Qwen2.5 technical report, 2025a. URL https://arxiv.org/abs/2412.15115.

An Yang et al. Qwen3 technical report, 2025b. URL https://arxiv.org/abs/2505.09388.

Z.ai. GLM-4-32B-0414. Hugging Face model card, https://huggingface.co/zai-org/GLM-4-32B-0414, April 2025a.

Z.ai. GLM-4-9B-0414. Hugging Face model card, https://huggingface.co/zai-org/GLM-4-9B-0414, April 2025b.

Zhihan Zhang, Shiyang Li, Zixuan Zhang, Xin Liu, Haoming Jiang, Xianfeng Tang, Yifan Gao, Zheng Li, Haodong Wang, Zhaoxuan Tan, Yichuan Li, Qingyu Yin, Bing Yin, and Meng Jiang. IHEval: Evaluating language models on following the instruction hierarchy. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8374–8398, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.425. URL https://aclanthology.org/2025.naacl-long.425/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

A EXPERIMENTAL DETAILS

The CV exam in full. The CV exam's 570 submissions answer four code questions: **Q1** transfer learning with EfficientNetV2, **Q2** a CNN from scratch, **Q3** semantic segmentation with a pretrained encoder, **Q4** a colorization-dataset bonus. Beyond the 35 base points (weights 12 / 11 / 12) sit up to 13 bonus points (Q1, Q2 and an all-or-nothing 5-point Q4), graded but excluded from every reported metric.

Model roster and serving. On the closed-model side we use Google's `gemini-2.5-pro` (Comanici et al., 2025) and `gemini-flash-lite`, `gemini-3-flash-preview`, and `gemini-3.1-pro-preview`, for which no technical report was available at the time of writing, via the Vertex API (Gemini Team et al., 2023). Two further vendors enter only in the cross-vendor replication of Section 6.1: OpenAI's `gpt-5.5` and `gpt-5.4` and Anthropic's `claude-opus-5` (Appendix C, Table 9). On the open-weights side we use 17 variants from six vendors. From Qwen: Qwen2.5-Coder at 7B / 14B / 32B (Hui et al., 2024) and the Qwen2.5 72B dense model (Yang et al., 2025a), and from the Qwen3 family the Qwen3-Coder 30B-A3B MoE, the 235B-A22B MoE, and the 480B Qwen3-Coder MoE in FP8 (Yang et al., 2025b), and the 80B Qwen3-Coder-Next MoE (Qwen Team, 2026). From Meta: Llama-3.1-8B (Grattafiori et al., 2024) and Llama-3.3-70B (Meta, 2024). Z.ai: GLM-4-9B and GLM-4-32B, the 0414 releases (Team GLM et al., 2024; Z.ai, 2025b;a), and the 106B GLM-4.5-Air MoE (GLM-4.5 Team et al., 2025). Google: Gemma-3-12B and Gemma-3-27B (Gemma Team et al., 2025). Mistral: Mistral-Small-24B (Mistral AI, 2025). DeepSeek: the 16B DeepSeek-Coder-V2-Lite MoE (2.4B active) (DeepSeek-AI et al., 2024). Serving is vLLM (Kwon et al., 2023) on one $4\times$ A100 node, or $8\times$ for Qwen2.5-72B, 235B-A22B and the 480B FP8 model.

Open-weights generation is *unconstrained*: the grader requests JSON in the prompt and parses the reply, with no grammar or schema enforcement at decode time (Vertex calls do pass a JSON response schema). A controlled A/B confirms that grammar-constrained decoding changes no score and costs $6.9\times$ in latency. Every open-weights grid run uses vLLM `0.19.1` or later, greedy decoding at temperature 0 — the $t = 0.5$ probes L-E01/L-E02 (Appendix C) excepted — and one request in flight.

Notebook serialisation, parsing and retries. Every call receives one question's notebook as plain text — a `-- Cell` k `[code|markdown] --` header then the cell's source — with every output dropped: training logs, printed metrics and rendered plots or masks (base64 media) never reach the model. Nothing is truncated in the grid; the fine-tuning harness alone head-and-tail-truncates to 24,000 characters. Replies are parsed as JSON; the open-weights path falls back from strict parsing to a lenient parse and then to `json_repair`, the Vertex path parses strictly against its response schema. A Vertex call is retried up to three times with a $2s\times$ attempt back-off, except that a `MAX_TOKENS` finish is not retried because it is deterministic at $t = 0$; the open-weights path does not retry a truncated or unparseable reply at $t = 0$ for the same reason. A question with no score after the run's retries and resumes is a permanent failure, and the student is excluded from that run's metrics as described under *Exclusions*.

Ablation-grid legend. Appendix I tabulates the grid's 33 closed-model (25 Gemini, 6 OpenAI, 2 Anthropic) and 138 open-weights runs; the machine-readable master comparison ships in the supplementary material. Gemini series: **A** = baseline (n=1); **B** = prompt-component ablations on Flash-Lite (n=4: B01 drop solution, B02 drop guidelines, B03 thinking on, B04 drop breakdown); **C** = strictness persona on Flash-Lite (n=2); **D** = model swap (n=3); **E** = temperature variance (n=3); **F** = combos and the 3.1 Pro strict run (n=3); **G**, **M** = prompt ablations and persona on `gemini-3-flash-preview` (n=4, n=2); **K** = few-shot (n=1); **P** = alternate-wording personas on Flash-Lite (n=2). Cross-vendor closed series (batch APIs; Section 6.1): **O** = OpenAI, `gpt-5.5` (O01–O03) and `gpt-5.4` (O11–O13) under *neutral*, *strict* and *lenient* (n=6); **N** = Anthropic `claude-opus-5` under *neutral* and *strict* (n=2).

Open-weights series (prefixed **L-**) fall in three groups. *Qwen ablations:* **L-A/L-AN** = per-model baselines (n=5); on Qwen2.5-Coder-32B, **L-B** = prompt-component ablations (n=4), **L-C** = persona (n=2) and **L-P** = alternate wordings (n=2); **L-D/L-DN** = +thinking on Qwen3 MoE (n=2); **L-E** = temperature variance (n=2); **L-G**, **L-H** = prompt ablations replayed on the 30B and 7B (n=6); **L-K** = few-shot (n=1); **L-M** = persona on additional Qwen models (n=7); **L-N**, **L-Q**, **L-R** = full

five-persona sweeps on Qwen2.5-72B, Qwen3-235B-A22B and Qwen3-Coder-480B FP8 (n=15). *Cross-family sweeps*, each five personas on all 570 students: **L-S** = DeepSeek-Coder-V2-Lite; **L-U** = Llama-3.1-8B; **L-V** = Llama-3.3-70B; **L-W** = GLM-4-9B; **L-X** = GLM-4-32B; **L-J** = GLM-4.5-Air; **L-F** = Gemma-3-12B; **L-Y** = Gemma-3-27B; **L-Z** = Mistral-Small-24B (n=45). *Mechanism probes:* **L-MP** = minimal pairs isolating the persona's headword from its policy sentences, on five models (n=25); **L-PS** = the 2×2 over the two policy sentences, on three models (n=12); **L-BD** = rubric-breakdown removal crossed with persona, on five models (n=10).

G- and M-series Gemini configurations run on the first 100 students, `gemini-3-flash-preview` throughput being the binding constraint; comparisons against full-n baselines restrict the baseline to the matching subset.

Gemini thinking configuration. The Vertex API exposes an internal reasoning budget, `thinking_budget`. We run 21 of the 25 Gemini ablations with thinking *disabled* (`thinking_budget` $= 0$) and three (B03, F01, G03) *enabled* at the dynamic budget (-1; the model picks the per-call budget). The exception is D03: `gemini-2.5-pro` rejects `thinking_budget` $= 0$, so it ran under that model's default dynamic policy and reads as thinking-enabled (Section 10).

Exclusions. A student with any of Q1–Q3 ungraded after retries is excluded from that run's metrics rather than scored with a partial total: no students in most runs, at most 3 elsewhere, the exceptions being P02 (13) and the temperature-variance E-series, where a student is kept only if all of Q1–Q3 were graded in at least one rerun (21 in E01, 7 each in E02/E03); per-run n accompanies the affected tables. P02's 13 are genuine permanent failures, one question each (6 on Q1, 7 on Q2); E01's 21 are 14 such failures plus the 7 students who submitted no notebook for at least one of Q1–Q3, and E02/E03's 7 are those non-submitters alone — neither run suffered a grading failure. Only the E-series rerun rule drops a non-submitter: every single-call run keeps them and scores the missing question 0, and restoring them that way moves the E-series MAE by -0.03 ($3.38 \rightarrow 3.36$, $3.51 \rightarrow 3.48$, $3.60 \rightarrow 3.57$). No log survives for these four runs, so we cannot quote a `finish_reason`; where logs do survive (the ML exam's Gemini grid and F03), 705 of 709 permanently failed cells are `MAX_TOKENS` truncations and the other 4 JSON parse errors, with 429, 503 and dropped connections retryable and no safety block at all. The artifacts agree: a failure burns 149–316 s (P02, one call) or 2,095–2,529 s (E01, five reruns) against a 5–6 s normal call; E01's 14 cells fail $5/5$ at $t = 0$ yet grade $5/5$ at $t = 0.5$ and 0.7; and the longest prompt, 39,418 characters ($\approx$ 10,000 tokens) against a million-token window, leaves only the 65,535-token output cap to bind. Unlike the ML exam's few-shot Flash-Lite run (Appendix H), the dropout is neither length- nor ability-biased: excluded notebooks are no longer than survivors' (P02 29,433 vs. 31,625 characters, Welch $p = 0.35$; E01's 14, 31,657 vs. 31,772, $p = 0.90$) and their grader-average totals match (26.9 vs. 26.0 out of 35, $p = 0.69$; 26.0 vs. 26.2, $p = 0.91$). E-series per-question scores are the mean over all recorded reruns (five per cell by design; some Gemini cells logged more) and the total is the Q1–Q3 base sum as elsewhere.

Item-level agreement. Table 2 gives per-question and total-level Spearman ρ between the AI mark and the grader-average mark for the headline configurations of both exams.

Table 2: Item-level agreement for both exams' headline configurations: Spearman ρ between the AI mark and the grader-average mark, per question and at total level, over the same students each run's MAE uses. The total is the optimistic number: across the 40 CV and 38 ML runs measured here (the rows below plus each exam's 17 matched neutral/strict pairs; Table 3, Section 5.4), it exceeds the weakest question's ρ in all 37 CV and 37 ML runs where every correlation is defined — median gap 0.20 (CV), 0.12 (ML), and at least 0.15 in 30 CV and 12 ML runs. The other three CV and one ML runs give every student the same mark, leaving their correlations undefined rather than zero. The weak item is the same within an exam: Q2 on CV (30 of 37), Q1 on ML (27 of 37). Pearson r differs from ρ by a median of 0.02 on both exams. Generated by `analysis/checks/per_question_rank_correlation.py`.

Run	Model	Persona	n	Q1 ρ	Q2 ρ	Q3 ρ	Total ρ
CV exam							
D01	gemini-3-flash-preview	*neutral*	570	0.92	0.90	0.91	0.95
D02	gemini-3.1-pro-preview	*neutral*	570	0.92	0.90	0.90	0.94
F02	gemini-3.1-pro-preview	*lenient*	570	0.92	0.90	0.91	0.94
B03	gemini-flash-lite + thinking	*neutral*	570	0.89	0.88	0.87	0.92
A01	gemini-flash-lite	*neutral*	570	0.77	0.65	0.83	0.84
C01	gemini-flash-lite	*strict*	570	0.78	0.59	0.81	0.83
ML exam							
IG08	gemini-3.1-pro-preview	*neutral*	1038	0.88	0.91	0.93	0.93
IG07	gemini-3-flash-preview	*neutral*	1038	0.86	0.91	0.94	0.93
IG01	gemini-flash-lite	*neutral*	1013	0.77	0.85	0.85	0.88
IG05	gemini-flash-lite	*strict*	1008	0.78	0.81	0.86	0.88

Table 3: The *strict* persona against each model's own *neutral* baseline at an identical prompt configuration. MAE in points out of 35; ratio is strict ÷ neutral; behaviour classes as in Section 5.1; a refusal's MAE is a ceiling artifact and must not be read as severity. Every row is $n = 570$. Generated by `analysis/computer_vision_make_paper_tables.py`.

Run	Model	Family	Params	Neutral	**Strict**	Ratio	Behaviour
Open weights — ordered by total parameters							
L-M05	Qwen2.5-Coder-7B	Qwen	7B	5.68	**7.92**	×1.39	graded
L-U02	Llama-3.1-8B	Llama	8B	7.31	**26.04**	×3.56	**refusal**
L-W02	GLM-4-9B	GLM	9B	4.31	**25.48**	×5.91	near-refusal
L-F02	Gemma-3-12B	Gemma	12B	5.20	**8.77**	×1.69	collapse
L-M07	Qwen2.5-Coder-14B	Qwen	14B	3.51	**24.52**	×6.99	collapse
L-S02	DeepSeek-Coder-V2-Lite	DeepSeek	16B	5.98	**12.09**	×2.02	collapse
L-Z02	Mistral-Small-24B	Mistral	24B	3.66	**26.03**	×7.11	**refusal**
L-Y02	Gemma-3-27B	Gemma	27B	4.32	**11.55**	×2.67	collapse
L-M01	Qwen3-Coder-30B-A3B	Qwen	30B	4.50	**14.91**	×3.31	collapse
L-X02	GLM-4-32B	GLM	32B	2.85	**9.66**	×3.39	collapse
L-C01	Qwen2.5-Coder-32B	Qwen	32B	7.49	**20.29**	×2.71	collapse
L-V02	Llama-3.3-70B	Llama	70B	4.52	**13.52**	×2.99	collapse
L-N02	Qwen2.5-72B	Qwen	72B	3.14	**9.82**	×3.13	collapse
L-M03	Qwen3-Coder-Next 80B	Qwen	80B	5.17	**11.70**	×2.26	collapse
L-J02	GLM-4.5-Air	GLM	106B	5.66	**21.02**	×3.71	collapse
L-Q02	Qwen3-235B-A22B	Qwen	235B	4.09	**7.32**	×1.79	graded
L-R02	Qwen3-Coder-480B	Qwen	480B	3.14	**3.78**	×1.20	graded
Closed — reference							
C01	Flash-Lite	Gemini	—	3.34	5.75	×1.72	graded
F03	3.1 Pro Preview	Gemini	—	1.86	2.75	×1.48	graded

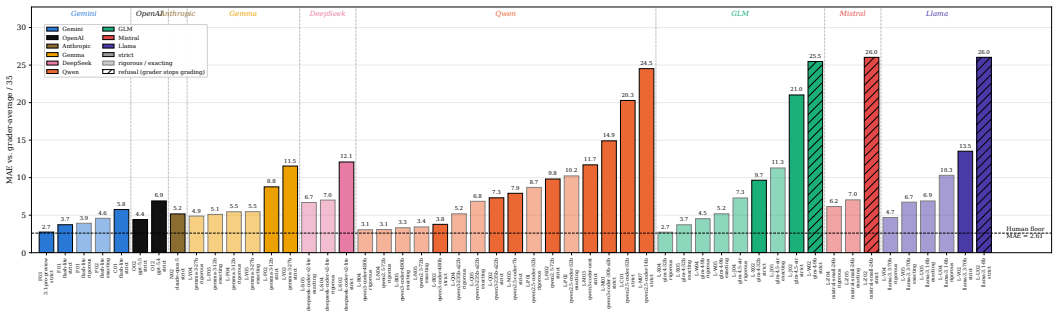

Figure 2: Every strict-flavoured persona run on the full cohort (51 runs; 50 at the default prompt configuration, plus F01, which additionally enables thinking), coloured by family and by wording (*strict* darker; *rigorous/exacting* lighter). The dashed line is the human inter-grader floor (MAE = 2.61). Runs classified as refusals are marked: their bar height is the distance to the grader mean and is not a severity measurement (Section 5.1). The two default-configuration strict runs not shown are M01 ($n = 99$, below the full-cohort cutoff) and F03 (`gemini-3.1-pro-preview`, 2.75; Table 3).

B Strict-Persona Collapse: Additional Tables

This appendix carries the full exhibits behind Section 5.

M01 ($n = 99$, the only strict run on `gemini-3-flash-preview`) has a CI of $[3.60, 4.88]$ spanning most of the Gemini cluster of Section 5.1, so we do not rank it within that cluster.

The three near-total zeroers have awarded-total standard deviations of 0.00, 0.09 and 2.42 and AI–grader correlations undefined, −0.01 and 0.19. GLM-4.5-Air blanket-zeroes 63% of students and still grades the rest ($r = 0.53$), which is why Table 3 classes it a collapse rather than a refusal. Qwen2.5-Coder-14B (L-M07) is the second intermediate case: it zeroes 88% of whole submissions

Table 4: The three strict-flavoured presets on the models that received all of them. Bold marks a run outside the graded band. These presets are **not** minimal pairs: only *strict* carries the two policy sentences (Section 5.2); *rigorous* and *exacting* ask for rubric precision with neither. Qwen2.5-Coder-7B, Qwen3-Coder-30B-A3B and Qwen3-Coder-Next received only *strict* in the full-cohort grid; Section 8.4 covers the first two.

Model	Family	Params	Neutral	*strict*	*rigorous*	*exacting*
Flash-Lite	Gemini	—	3.34	5.75	3.92	4.58
Llama-3.1-8B	Llama	8B	7.31	**26.04**	**10.31**	6.89
GLM-4-9B	GLM	9B	4.31	**25.48**	4.53	5.18
Gemma-3-12B	Gemma	12B	5.20	**8.77**	5.45	5.10
DeepSeek-Coder-V2-Lite	DeepSeek	16B	5.98	**12.09**	7.01	6.69
Mistral-Small-24B	Mistral	24B	3.66	**26.03**	6.15	7.04
Gemma-3-27B	Gemma	27B	4.32	**11.55**	4.88	5.47
GLM-4-32B	GLM	32B	2.85	**9.66**	2.74	3.72
Qwen2.5-Coder-32B	Qwen	32B	7.49	**20.29**	**8.70**	**10.23**
Llama-3.3-70B	Llama	70B	4.52	**13.52**	4.69	6.74
Qwen2.5-72B	Qwen	72B	3.14	**9.82**	3.09	3.43
GLM-4.5-Air	GLM	106B	5.66	**21.02**	7.29	**11.29**
Qwen3-235B-A22B	Qwen	235B	4.09	7.32	5.18	6.85
Qwen3-Coder-480B	Qwen	480B	3.14	3.78	3.06	3.33

Table 5: Attribution of the collapse to the policy sentences rather than the headword. *Left:* S1 and S2 held verbatim, only the adjective varies; the *nharsh* control is the neutral frame with its one adjective swapped to "harsh". *Right:* the HARSH frame held fixed, crossing S1 with S2 (both quoted in Section 5.2). All rows $n \geq 569$, default prompt configuration, $t = 0$. Spreads are max − min of the unrounded MAEs, so they can differ from the difference of the printed cells by 0.01. Bold marks cells whose damage matches the full preset's.

	headword varied, policy fixed					policy varied, frame fixed			
Model	STRICT	RIGOROUS	FAIR	spread	*nharsh*	frame	+S1	+S2	both
Qwen2.5-Coder-32B	16.41	17.74	18.55	2.15	8.25	9.37	15.55	12.92	**20.26**
Llama-3.3-70B	12.67	11.37	11.41	1.30	4.60	—	—	—	—
Gemma-3-27B	10.49	9.23	8.60	1.89	4.56	—	—	—	—
GLM-4-32B	7.02	7.43	6.41	1.02	2.88	—	—	—	—
Mistral-Small-24B	25.99	25.87	25.05	0.94	4.23	5.62	10.33	**26.04**	**26.03**
Llama-3.1-8B	—	—	—	—	—	8.46	23.45	**26.04**	**26.04**

(mean awarded total 1.52, $r = 0.25$), just under the 90% refusal cut. Q1-zero rates fall at 0–5% or 24–100% with nothing between, and among the latter the selective fraction is 0–15% or 51–77%.

Representative example. The clearest case in our sample is L-C01 student $\#2$, where both graders gave Q1 $= 12.0/12$ and the model emitted Q1 $= 0$. The reasoning blob begins:

> *The student completed most tasks correctly, including defining the transforms, creating DataLoaders, … However, the backbone was not properly frozen, and the bonus task for Test Time Augmentation was not implemented.*

The same blob then itemises a breakdown ("Complete transforms: 1.5", "Create DataLoaders: 0.5", …) summing to 10.0 points awarded in prose — the largest contradicted total in Table 7.

A prediction of the conflict account that fails. If the collapse is driven by a conflict between S2 and the rubric breakdown, withdrawing the breakdown should remove one side of the conflict and *relieve* the collapse. Removing the breakdown under *strict* worsens MAE by $+6.42$ on Llama-3.3-70B ($13.52 \to 19.94$), $+5.92$ on GLM-4-32B ($9.66 \to 15.58$), $+3.93$ on DeepSeek-Coder-V2-Lite ($12.09 \to 16.02$), $+1.40$ on Qwen2.5-Coder-32B ($20.29 \to 21.69$) and $+0.15$ on Gemma-3-27B ($11.55 \to 11.69$) — five of five in the wrong direction. The same manipulation under *neutral* is a

Table 6: How each matched *strict* run fails. Q1 $= 0$ is the count of students awarded zero on Q1; the next column is how many of *those* still received a non-zero Q2. Generated by `analysis/computer_vision_make_paper_tables.py`.

Run	Model	Family	MAE	Q1 $= 0$	... of which Q2 > 0
blanket zeroing — consistent with refusal					
L-U02	Llama-3.1-8B	Llama	26.04	570 / 570 (100%)	0 (0%)
L-Z02	Mistral-Small-24B	Mistral	26.03	570 / 570 (100%)	0 (0%)
L-W02	GLM-4-9B	GLM	25.48	570 / 570 (100%)	29 (5%)
L-M07	Qwen2.5-Coder-14B	Qwen	24.52	533 / 570 (94%)	31 (6%)
L-J02	GLM-4.5-Air	GLM	21.02	361 / 570 (63%)	55 (15%)
selective field collapse — consistent with conflicting instructions (a hypothesis; Section 5.3)					
L-C01	Qwen2.5-Coder-32B	Qwen	20.29	437 / 570 (77%)	284 (65%)
L-M01	Qwen3-Coder-30B-A3B	Qwen	14.91	357 / 570 (63%)	274 (77%)
L-V02	Llama-3.3-70B	Llama	13.52	291 / 570 (51%)	183 (63%)
L-N02	Qwen2.5-72B	Qwen	9.82	136 / 570 (24%)	97 (71%)
L-X02	GLM-4-32B	GLM	9.66	273 / 570 (48%)	139 (51%)
uniform severity — consistent with calibration drift					
L-S02	DeepSeek-Coder-V2-Lite	DeepSeek	12.09	1 / 570 (0%)	1 (100%)
L-M03	Qwen3-Coder-Next 80B	Qwen	11.70	19 / 570 (3%)	8 (42%)
L-Y02	Gemma-3-27B	Gemma	11.55	0 / 570 (0%)	n/a
L-F02	Gemma-3-12B	Gemma	8.77	1 / 570 (0%)	0 (0%)
L-M05	Qwen2.5-Coder-7B	Qwen	7.92	8 / 570 (1%)	6 (75%)
L-Q02	Qwen3-235B-A22B	Qwen	7.32	5 / 570 (1%)	1 (20%)
C01	Flash-Lite	Gemini	5.75	31 / 570 (5%)	21 (68%)
L-R02	Qwen3-Coder-480B	Qwen	3.78	1 / 570 (0%)	1 (100%)
F03	3.1 Pro Preview	Gemini	2.75	8 / 570 (1%)	5 (62%)

Table 7: Self-contradiction between the structured Q1 score and the prose breakdown, on the two strict-persona runs whose rationales we annotated by hand. A row is self-contradicting when the structured score is 0 but the rationale itemises positive partial credit. Both runs sit in the selective-field-collapse group of Table 6 and match its Q1 $= 0$ counts; the second column differs because this one requires reading the prose, which we did not do at scale.

Run	Q1 score = 0	...with prose > 0	mean contradicted prose-credit
L-C01 (Qwen2.5-Coder-32B + strict)	437 / 570 (77%)	86 (20%)	2.60 pts (max 10.0)
L-M01 (Qwen3-Coder-30B-A3B + strict)	357 / 570 (63%)	152 (43%)	2.75 pts (max 8.5)

clean control, with a largest shift of $+0.85$ and four of five models within ± 0.25, so this is specific to the persona rather than an artifact of a shorter prompt.

The fairest reading is that this weakens the causal story without settling on an alternative; the structural-scaffolding reading of Section 5.3 fits the L-BD result but is itself untested. Withdrawing the breakdown leaves the rubric table in the prompt, so the conflict is attenuated rather than eliminated; a fully clean test would strip the rubric too, which we have not run. We flag this as open rather than resolved (Section 10).

Prompt-only baselines. Before fine-tuning, we asked whether the *strict*-persona collapse can be undone by prompting alone. We tried the two repairs the literature suggests on the four CV-exam models whose *strict* runs collapse or refuse (Qwen2.5-Coder-32B, GLM-4-32B, Mistral-Small-24B, Llama-3.1-8B), over the first 100 students (the G-series subset), Q1–Q3, with the paper's vLLM stack, and compared each repair against the same model's own *strict* and neutral runs restricted to the same students (Table 8).

The first repair is *decomposition* (Pathak et al., 2025): instead of grading a whole question in one call, the grader makes one call per rubric task row (16/13/11 rows for Q1/Q2/Q3; penalty and bonus rows handled) and awards only that row's marks; the row awards are summed and clamped to the question maxima. The persona, guidelines, reference solution and submission are exactly those of the paper's prompt. The hope is that a small, concrete criterion leaves the model less room to zero a whole question. It does not. Qwen2.5-Coder-32B moves from 21.24 to 20.00 MAE, Mistral-Small-24B from 27.25 to 26.27, and Llama-3.1-8B refuses on every row exactly as it refused on every question (27.26). Only GLM-4-32B re-enters the graded band ($8.90 \rightarrow 5.01$), and even then it sits at $2.2\times$ its neutral error. The credit-withholding sentences bind on each criterion just as they bind on the question as a whole.

The second repair is *grade-then-arbitrate* (Chan et al., 2024): the paper's *strict* prompt and the paper's neutral prompt each grade the question, and a third call under the neutral persona receives both JSON grades together with the rubric and must resolve every disagreement on the rubric's partial-credit scale. The arbiter's grade is the run's grade, at three calls per question instead of one. This does bring every model back into the graded band, but only to where its own neutral prompt already was: GLM-4-32B 2.64 against 2.24 neutral, Mistral 3.13 against 3.59, Llama 7.46 against 7.73 (on 99 students: one arbiter reply, student 87, Q2, was truncated at 4,096 tokens), Qwen 9.29 against 8.02. The arbiter defers to the neutral grade, so arbitration repairs the persona by removing it, at three times the cost, and gains nothing beyond the neutral prompt.

Neither repair therefore reaches what fine-tuning reaches. The neutral level the repairs recover is 2.2–8.0 MAE on these four models; the pooled adapters of Section 8 start from that level and end at 1.75–2.01 on the held-out CV students, below the human floor. Scripts: `prompt_baselines_grade.py`, `analysis/checks/prompt_baselines_vs_paper.py`.

Table 8: Prompt-only repairs under the *strict* persona, CV exam, first 100 students, 35-point base scale; the paper's strict and neutral runs are restricted to the same students. MAE with 95% bootstrap CI (2000 resamples); zero = share of students awarded a total of 0; behaviour classes as in Section 5.1. Human floor on these students: 3.43.

Model	Run	MAE [CI]	Bias	Zero	Behaviour
Qwen2.5-Coder-32B	paper neutral (L-A32)	8.02 [7.38, 8.68]	−7.96	0%	collapse
	paper strict (L-C01)	21.24 [20.31, 22.14]	−21.24	18%	collapse
	decomposed, strict	20.00 [19.10, 20.93]	−20.00	3%	collapse
	arbitrated, strict	9.29 [8.58, 9.98]	−9.27	0%	collapse
GLM-4-32B	paper neutral (L-X01)	2.24 [1.91, 2.59]	−0.17	0%	graded
	paper strict (L-X02)	8.90 [7.81, 10.03]	−8.81	5%	collapse
	decomposed, strict	5.01 [4.50, 5.51]	−4.95	0%	graded
	arbitrated, strict	2.64 [2.23, 3.10]	−1.53	0%	graded
Mistral-Small-24B	paper neutral (L-Z01)	3.59 [3.13, 4.08]	−2.81	0%	graded
	paper strict (L-Z02)	27.25 [25.88, 28.64]	−27.25	99%	**refusal**
	decomposed, strict	26.27 [24.98, 27.55]	−26.27	56%	collapse
	arbitrated, strict	3.13 [2.74, 3.53]	−2.31	0%	graded
Llama-3.1-8B	paper neutral (L-U01)	7.77 [6.88, 8.61]	−6.24	0%	graded
	paper strict (L-U02)	27.26 [25.89, 28.65]	−27.26	100%	**refusal**
	decomposed, strict	27.26 [25.89, 28.65]	−27.26	100%	**refusal**
	arbitrated, strict ($n = 99$)	7.46 [6.60, 8.28]	−5.77	0%	graded

Table 9: Closed models from three vendors under *neutral*, *strict* and *lenient* on both exams (Section 6.1): MAE against the grader average with 95% bootstrap CI, and bias. Gemini rows are grid runs (D02/F03/F02 and A01/C01/C02 on the CV exam; IG08/IG23/IG14 and IG01/IG05/IG06 on the ML exam, with their n as in Tables 24 and 25). OpenAI and Anthropic rows were graded on the full cohorts through the vendors' batch APIs with the identical prompt, reasoning off and temperature 0 (OpenAI) or thinking disabled (Anthropic, whose API exposes no temperature); every reply parsed. Bold marks a run outside the graded band; no zero-total rate exceeds 1.1%. Generated by `analysis/closed_vendor_table.py`.

Model	Persona	CV exam (floor 2.61)		ML exam (floor 5.13)	
		MAE [95% CI]	bias	MAE [95% CI]	bias
claude-opus-5	*neutral*	3.54 [3.31, 3.77]	−3.24	4.37 [4.13, 4.63]	−2.64
	strict	5.17 [4.90, 5.45]	−5.02	—	—
gemini-3.1-pro-preview	*neutral*	1.86 [1.70, 2.02]	−0.82	3.40 [3.16, 3.65]	+1.18
	strict	2.75 [2.51, 2.99]	−2.09	3.67 [3.43, 3.94]	−1.00
	lenient	1.79 [1.62, 1.95]	+0.82	4.48 [4.22, 4.75]	+3.46
gemini-flash-lite	*neutral*	3.34 [3.15, 3.53]	−1.78	7.53 [7.20, 7.87]	+5.98
	strict	5.75 [5.46, 6.04]	−5.44	9.02 [8.61, 9.43]	−7.83
	lenient	4.49 [4.18, 4.82]	+4.01	**17.86** [17.30, 18.42]	+17.80
gpt-5.4	*neutral*	4.69 [4.46, 4.92]	−4.45	4.30 [4.07, 4.55]	−1.95
	strict	6.90 [6.63, 7.15]	−6.81	5.11 [4.85, 5.37]	−3.61
	lenient	2.69 [2.51, 2.89]	+1.00	8.30 [7.96, 8.64]	+7.94
gpt-5.5	*neutral*	2.43 [2.28, 2.59]	−1.65	3.54 [3.33, 3.77]	+1.23
	strict	4.43 [4.19, 4.68]	−4.25	3.70 [3.48, 3.93]	−1.38
	lenient	2.25 [2.08, 2.44]	+1.13	6.29 [6.02, 6.57]	+5.83

Table 10: The six full-cohort configurations, and five first-100 runs, whose point-estimate MAE meets or undercuts the human inter-grader floor, with 95% bootstrap CIs (2000 resamples), plus reference rows. The G-series and M02 rows cover only the first $\sim$ 100 students, an easier subset (the unmodified D01 recipe scores 1.23 there vs. 1.64 on the full cohort; Appendix C); they are not comparable to the full-cohort rows, and M02's CI upper bound crosses the floor.

Run	Model	n	MAE	95% CI	Bias
Human inter-grader floor (Section 4)					
—	G_1 vs. G_2	570	2.61	[2.37, 2.85]	—
Full-cohort configurations					
D01	gemini-3-flash-preview, neutral	570	**1.64**	[1.51, 1.79]	+0.01
F02	gemini-3.1-pro-preview + lenient	570	1.79	[1.62, 1.95]	+0.82
D02	gemini-3.1-pro-preview, neutral	570	1.86	[1.70, 2.02]	−0.82
B03	gemini-flash-lite + thinking	570	2.05	[1.89, 2.21]	−0.15
O03	gpt-5.5 + lenient (batch API)	570	2.25	[2.08, 2.44]	+1.13
O01	gpt-5.5, neutral (batch API)	570	2.43	[2.28, 2.59]	−1.65
First-100-students subset (easier subset; see caption)					
D01 (restricted)	gemini-3-flash-preview, baseline recipe on the same 100 students	100	1.23	[1.02, 1.49]	−0.18
G04	3-flash-preview, no rubric breakdown	100	1.25	[1.01, 1.51]	−0.07
G03	3-flash-preview + thinking	100	1.39	[1.18, 1.61]	−0.12
G02	3-flash-preview, no guidelines	99	1.60	[1.30, 1.93]	+0.79
G01	3-flash-preview, no reference solution	98	1.63	[1.32, 2.02]	+0.71
M02	3-flash-preview + lenient	100	2.18	[1.75, 2.65]	+1.90
Reference rows					
A01	gemini-flash-lite (baseline recipe)	570	3.34	[3.15, 3.53]	−1.78
L-R04	Qwen3-Coder-480B + rigorous (best robust open)	570	3.06	[2.88, 3.26]	−0.23

C CLOSED-MODEL DETAILS

The apparent G-series wins are a sample-size artifact ($n = 100$). The G-series replays the ablations on `gemini-3-flash-preview`, whose throughput limited these runs to the first 100 students. Against D01's full-cohort 1.64, G04 (no rubric breakdown, 1.25) appears to win; restricting

Table 11: The paired third-grader test (Eq. (1)) applied to the zero-shot configurations, on the full 570-student CV cohort: the same statistic and decision rule as Table 19, the standard the fine-tuned models face. *better* means the whole 95% CI of $\overline{d}$ sits below zero, "= grader" that it contains zero, *worse* that it lies above. CIs are the normal approximation used in Table 19; a 2000-resample percentile bootstrap gives the same verdict in every row. With the floor at 2.61, $\overline{d} = \frac{1}{2}(|\mathrm{AI} - \mathrm{G}_1| + |\mathrm{AI} - \mathrm{G}_2|) - 2.61$ exactly. The two single-grader distances differ by up to 0.63 here, so both are shown. Generated by `analysis/checks/paired_test_zero_shot_cv.py`.

Run	Configuration	$\lvert\mathrm{AI}-\mathrm{G}_{\mathrm{avg}}\rvert$	$\lvert\mathrm{AI}-\mathrm{G}_1\rvert$	$\lvert\mathrm{AI}-\mathrm{G}_2\rvert$	$\overline{d}$ [95% CI]	Verdict
D01	3-flash-preview, neutral	1.64	2.23	1.92	$-0.54\ [-0.71, -0.36]$	**better**
F02	3.1-pro-preview + lenient	1.79	2.32	2.15	$-0.37\ [-0.56, -0.18]$	**better**
D02	3.1-pro-preview, neutral	1.86	2.53	1.96	$-0.37\ [-0.55, -0.18]$	**better**
B03	flash-lite + thinking	2.05	2.55	2.29	$-0.19\ [-0.39, +0.01]$	= grader
O03	`gpt-5.5` + lenient (Section 6.1)	2.25	2.59	2.62	$-0.01\ [-0.22, +0.21]$	= grader
O01	`gpt-5.5`, neutral (Section 6.1)	2.43	3.10	2.51	$+0.19\ [-0.02, +0.41]$	= grader
Reference rows (do not clear the floor)						
A01	flash-lite, baseline recipe	3.34	3.83	3.46	$+1.03\ [+0.77, +1.29]$	worse
N01	`claude-opus-5`, neutral (Section 6.1)	3.54	4.05	3.42	$+1.13\ [+0.88, +1.37]$	worse

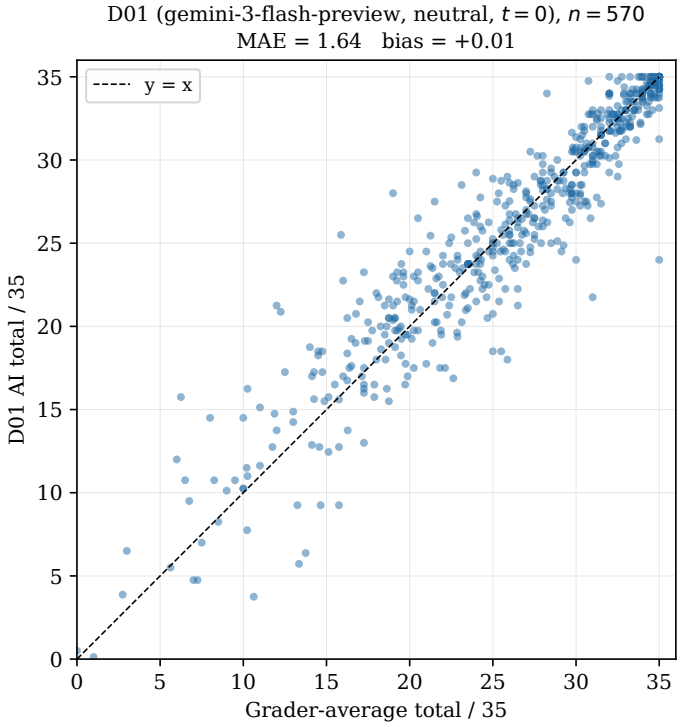

Figure 3: D01 (`gemini-3-flash-preview`, neutral, $t = 0$): AI total vs. grader-average total over the full $n = 570$ submissions, with $y = x$ as the perfect-agreement reference. Bias is $+0.01$, and residual error concentrates in the under-15 band, where graders also disagree more.

D01 to the same students gives 1.23, so the baseline ties G04 within noise and leads the other three (Table 13). Dropping the reference solution or the guidelines costs about 0.4 MAE, with CIs still grazing the baseline's.

Flash-Lite at $t = 0$ is deterministic here; sampling adds variance without gain. We measured sampling variance on `gemini-flash-lite` (E01/E02/E03 at $t = 0.0/0.5/0.7$, five reruns per cell). As everywhere in this paper the statistics cover the three graded questions only (Q4 is bonus-only), leaving 1,687/1,701/1,701 (student, question) cells per configuration. At $t = 0$ the model is fully deterministic: 100% of cells have zero standard deviation across reruns. At $t = 0.5$ the median

Table 12: Prompt-component ablations on `gemini-flash-lite`. B03 vs. A01 measures dynamic-budget thinking against the baseline's disabled thinking (`thinking_budget` $= 0$; Appendix A). $n = 570$ except B01 (567) and B02/B04 (569 each), which exclude students with a permanently failed grading call.

Run	Change vs. A01	MAE	Bias
A01	— (baseline)	3.34	-1.78
B01	drop reference solution	4.13	$+3.30$
B02	drop guidelines	3.34	-1.42
B03	thinking on (dynamic)	**2.05**	-0.15
B04	drop rubric breakdown	3.22	-1.45

Table 13: G-series on `gemini-3-flash-preview`, all on the first-100 subset ($n = 100$, except G01/G02 at 98/99 after excluding permanently failed grading calls). D01-restricted is the baseline recipe on the same students. Bootstrap CI on 2000 resamples.

Run	Change vs. baseline	MAE	95% CI
D01 (restricted)	— (baseline)	**1.23**	[1.02, 1.49]
G04	drop rubric breakdown	1.25	[1.01, 1.51]
G03	thinking on (dynamic)	1.39	[1.18, 1.61]
G02	drop guidelines	1.60	[1.30, 1.93]
G01	drop reference solution	1.63	[1.32, 2.02]

per-cell standard deviation is 0.65 points and 22.3% of cells have a standard deviation above one point; at $t = 0.7$ those figures rise to 0.76 and 31.8%. Qwen2.5-Coder-32B and Qwen3-Coder-Next at $t = 0.5$ ($\sim$ 50-student subset) show 0.76 and 0.82, comparable to Gemini at $t = 0.7$. Averaging the reruns does not beat a single $t = 0$ call: mean-of-reruns MAE is 3.38 at $t = 0$, 3.51 at $t = 0.5$, and 3.60 at $t = 0.7$, versus 3.34 for the single-call A01 baseline (the 0.04 gap at $t = 0$ comes from E01's smaller cohort, not from nondeterminism). We therefore recommend a single call at $t = 0$ as the production default.

D OPEN-WEIGHTS DETAILS

This appendix carries the exhibits behind Section 7: neutral-prompt baselines for all 17 variants and the few-shot comparison. L-A32 rationales often penalise valid stylistic divergence from the reference solution — different variable names or layer orderings — supporting the over-anchoring account of the 32B regression informally.

Few-shot prompting moves the open model substantially. Two worked examples per question (Table 15) improve Flash-Lite by 0.31 MAE and Qwen3-Coder-30B-A3B by ≈ 1.16; on the ML exam it nearly halves the same model's error ($11.12 \rightarrow 5.84$; Appendix H). Excluding the five demonstration students changes nothing measurable: on the 562 (Flash-Lite) and 565 (Qwen3-Coder-30B-A3B) students left, the uplift is $+0.30$ $[0.05, 0.54]$ and $+1.16$ $[1.04, 1.30]$ against $+0.31$ and $+1.16$ on the full cohort, and the ML exam's arms match ($+5.28$ against $+5.28$ for the 30B, $+3.56$ against $+3.55$ for Flash-Lite, less its three demonstration students).

Table 14: Open-weights baselines on the default prompt (neutral persona, $t = 0$), ordered by family then total parameters. MAE in points out of 35 with 95% bootstrap CI; bias is mean signed error (AI − grader-avg). Six vendors and both dense and mixture-of-experts architectures. All rows $n = 570$.

Run	Model	Params	MAE	95% CI	Bias
L-S01	DeepSeek-Coder-V2-Lite (MoE)	16B	5.98	[5.69, 6.29]	−3.44
L-W01	GLM-4-9B	9B	4.31	[4.02, 4.63]	+1.68
L-X01	GLM-4-32B	32B	**2.85**	[2.66, 3.05]	+0.50
L-J01	GLM-4.5-Air (MoE)	106B	5.66	[5.38, 5.95]	−5.37
L-F01	Gemma-3-12B	12B	5.20	[4.94, 5.48]	−2.75
L-Y01	Gemma-3-27B	27B	4.32	[4.12, 4.55]	−1.52
L-U01	Llama-3.1-8B	8B	7.31	[6.97, 7.68]	−5.15
L-V01	Llama-3.3-70B	70B	4.52	[4.31, 4.73]	−3.60
L-Z01	Mistral-Small-24B	24B	3.66	[3.46, 3.87]	−2.41
L-A07	Qwen2.5-Coder-7B	7B	5.68	[5.39, 5.98]	−4.84
L-A14	Qwen2.5-Coder-14B	14B	3.51	[3.30, 3.74]	−1.16
L-A30M	Qwen3-Coder-30B-A3B (MoE)	30B	4.50	[4.29, 4.72]	−2.70
L-A32	Qwen2.5-Coder-32B	32B	7.49	[7.19, 7.78]	−7.35
L-N01	Qwen2.5-72B (dense)	72B	3.14	[2.95, 3.35]	+0.19
L-ANxt	Qwen3-Coder-Next 80B (MoE)	80B	5.17	[4.92, 5.42]	−4.74
L-Q01	Qwen3-235B-A22B (MoE)	235B	4.09	[3.87, 4.32]	−3.56
L-R01	Qwen3-Coder-480B FP8 (MoE)	480B	3.14	[2.95, 3.35]	+0.51

Table 15: Two worked examples per question, drawn from the highest-agreement grader pair and labelled with D01's scores and rationales (in-context distillation, not human exemplars); the five demonstration students remain in the evaluated cohort (Appendix J), though excluding them moves the uplift by at most 0.01 (`analysis/checks/fewshot_demo_leakage.py`). Few-shot moves Gemini Flash-Lite by ≈ 0.3 MAE and Qwen3-Coder-30B-A3B by ≈ 1.2 MAE points.

Run	Model	MAE	Bias
A01	gemini-flash-lite (no few-shot)	3.34	−1.78
K01	gemini-flash-lite + few-shot	3.03	+1.09
L-A30M	Qwen3-Coder-30B-A3B (no few-shot)	4.50	−2.70
L-K01	Qwen3-Coder-30B-A3B + few-shot	3.34	−1.77

Table 16: Worst MAE drift over the three harsh personas (*strict*, *rigorous*, *exacting*): the largest signed difference persona − neutral; when all three are negative it is the least negative, which is why Qwen3-Coder-30B-A3B's ML entry of -1.26 is its *rigorous* drift while *strict* sits at -2.26 (marks recipe; full grid in Table 18). †: negative drift: the persona accidentally *improved* the over-grading base (Section 8.4). *lenient*, swept after tuning, is in Table 18: the pooled marks adapters drift $+0.05$ to $+0.39$ except Qwen3-Coder-30B-A3B ($+1.02$ CV, $+1.81$ ML); bd $+0.20$ to $+1.64$.

	CV exam		ML exam	
Model	base	pooled	base	pooled
Qwen2.5-Coder-7B	$+5.40$	$+0.02$	$+0.22$	-0.04
Qwen2.5-Coder-14B	$+19.75$	-0.02	$+1.13$	$+0.32$
Qwen3-Coder-30B-A3B	$+12.51$	$+0.07$	$-1.26^{\dagger}$	$+0.29$
Gemma-4-E4B	$+5.81$	$+0.05$	$+6.53$	$+0.30$
Llama-3.1-8B	$+17.28$	-0.05	$+18.62$	$+0.03$

Table 17: Held-out MAE vs. the grader average, mean over the five models, both recipes (the per-model marks grid is Table 1). The pooled adapter is best or tied in all eight columns.

Graded on	base	CV-trained	ML-trained	pooled
CV exam, marks	5.54	1.90	3.57	**1.87**
CV exam, bd	4.73	2.13	3.52	**2.01**
ML exam, marks	7.95	5.36	3.54	**3.41**
ML exam, bd	8.76	6.16	4.24	**4.15**

E Fine-Tuning Details

All cells use the pinned held-out students (114 CV / 208 ML exam). Evaluation follows each exam's ground truth: the CV exam records score and bonus separately, so agreement is score-only on both sides; the ML exam folds bonus into the grade, so agreement compares AI score + bonus to the grader total.

Hyperparameters and evaluation. We use LoRA with rank 16, $\alpha = 32$, dropout 0.05, and all-linear targets (attention-only for the 30B MoE), in bf16 for two epochs at learning rate 2×10^{-4} with cosine decay and effective batch size 16, held fixed across four data-parallel A100s; Gemma-4 trains on a single A100, as its unused parameters per step trip DDP's reduction check. All evaluation is served by vLLM, including Gemma-4's adapter, which reproduces the HuggingFace-generation numbers to within 0.1 MAE across its four re-run cells.

Q3 of the CV exam retains a residual ceiling. Fine-tuning lowers MAE on every graded CV-exam question, but a residual ceiling remains on Q3, semantic segmentation, the most open-ended: post-tuning MAE 1.1–1.3 across models under the marks recipe, against 0.6–1.0 on Q1/Q2.

The catastrophic tail disappears, on both exams. Under a neutral prompt the base models misgrade up to 43 of 114 CV-exam students (Llama) and 66 of 208 ML-exam students by more than 10 marks (>16 on the ML exam's larger scale); pooled, every model is at 0–1 (CV) and 2–4 (ML) under marks (0–1 and 4–6 under bd).

Errors become human-like, on both exams. We correlate each grader's per-student absolute error with the exam's intrinsic ambiguity, $|G_1 - G_2|$. On the CV exam the bases show no consistent relationship (-0.31 to $+0.37$) and four of five blunder on submissions the two graders *agreed* on; after pooled fine-tuning the correlation is positive for every model on both exams ($+0.35$ to $+0.48$), the residual errors concentrating on the genuinely ambiguous submissions. Part of that correlation is arithmetic: the grader-average target carries the graders' own noise, so any low-bias grader's residual co-varies with $|G_1 - G_2|$ (the coupling Appendix G notes for pairs) and we do not subtract it; the base-to-tuned change establishes the loss of blunders on submissions the graders agreed on, not human-like judgement.

Table 18: The full persona grid (MAE, marks recipe; per-persona bias and the bd grid are in the released workbooks). †: the sign-inconsistency cell of Section 8.4 — *strict* improves the over-grading base. *len.* is the inflating direction, swept after tuning: the pooled marks adapters drift $+0.05$ to $+0.39$ except Qwen3-30B ($+1.02$ CV, $+1.81$ ML, the latter above the 5.19 floor); every pooled lenient drift is positive (bias $+1.0$ to $+4.4$). bd worst drifts: base up to $+18.5$ on either exam; pooled $\leq +0.62$ under the three harsh wordings, with Llama-bd-*strict* at 2.67 (down from 4.08 under the CV-exam-only adapter); under *lenient* the bd adapters drift $+0.20$ to $+1.64$ (Qwen3-30B ML $4.07 \to 5.71$, Gemma-4 ML $4.22 \to 5.51$).

		CV exam					ML exam				
Model	Arm	neut.	strict	rigor.	exact.	len.	neut.	strict	rigor.	exact.	len.
Qwen2.5-7B	base	6.60	12.00	9.07	6.07	4.81	6.48	6.70	6.04	6.15	7.50
	pooled	1.84	1.84	1.85	1.81	2.00	3.40	3.36	3.30	3.31	3.52
Qwen2.5-14B	base	4.06	**23.81**	3.65	7.31	3.79	7.70	8.83	7.71	6.33	11.18
	pooled	1.89	1.85	1.87	1.81	2.00	3.29	3.61	3.23	3.28	3.50
Qwen3-30B	base	4.76	17.28	4.96	5.76	4.49	8.60	6.34†	7.34	6.30	13.00
	pooled	1.75	1.80	1.79	1.82	2.78	3.45	3.73	3.58	3.72	5.26
Gemma-4	base	4.29	10.10	4.69	6.03	3.89	4.66	11.19	5.11	6.22	7.39
	pooled	1.88	1.93	1.89	1.82	2.26	3.53	3.83	3.54	3.57	3.87
Llama-8B	base	7.99	**25.28**	12.35	11.95	6.86	12.30	**30.92**	7.16	7.97	29.23
	pooled	2.01	1.96	1.93	1.91	2.05	3.37	3.39	3.26	3.26	3.46

Table 19: Paired third-grader test (Eq. (1)) for the pooled adapters. Floors: 2.83 (CV), 5.19 (ML). *better*: the whole 95% CI of $\overline{d}$ below zero; "= grader": it contains zero. Both single-grader distances are shown: $|\mathrm{AI} - \mathrm{G}_1|$ is the larger in all ten ML rows and nine of ten CV rows, by up to 0.22 (CV) and 0.57 (ML). No paired-bootstrap difference excludes zero, and one ML student whose G_1 slot reads 0.0 against a real G_2 mark (Appendix H) alone accounts for 0.21–0.25 of that gap, so we read the asymmetry as the grader-slot artifact, not a systematic preference. Three CV upper bounds fall inside ± 0.05 and are printed to three decimals; all stay below zero under both the normal CI used here and a 2000-resample percentile bootstrap, though 7B bd, CV is marginal (Section 10). Recomputed by `analysis/checks/ftpaired_marginal_verdicts.py` and `ftpaired_symmetry_guard.py`.

Cell	n	$\lvert\mathrm{AI}-\mathrm{G}_{\mathrm{avg}}\rvert$	$\lvert\mathrm{AI}-\mathrm{G}_1\rvert$	$\lvert\mathrm{AI}-\mathrm{G}_2\rvert$	$\overline{d}$ [95% CI]	Verdict
7B marks, CV	114	1.84	2.35	2.35	$-0.48\ [-0.90, -0.05]$	**better**
14B marks, CV	114	1.89	2.38	2.35	$-0.46\ [-0.89, -0.044]$	**better**
30B marks, CV	114	1.75	2.38	2.17	$-0.56\ [-0.98, -0.13]$	**better**
Gemma marks, CV	114	1.88	2.41	2.34	$-0.45\ [-0.89, -0.020]$	**better**
Llama marks, CV	114	2.01	2.54	2.34	$-0.39\ [-0.85, +0.07]$	= grader
7B marks, ML	208	3.40	4.19	3.96	$-1.11\ [-1.73, -0.49]$	**better**
14B marks, ML	208	3.29	4.11	4.02	$-1.13\ [-1.75, -0.50]$	**better**
30B marks, ML	208	3.45	4.34	3.95	$-1.04\ [-1.65, -0.43]$	**better**
Gemma marks, ML	208	3.53	4.45	4.05	$-0.94\ [-1.57, -0.30]$	**better**
Llama marks, ML	208	3.37	4.33	4.03	$-1.01\ [-1.65, -0.36]$	**better**
7B bd, CV	114	1.94	2.41	2.35	$-0.45\ [-0.90, -0.001]$	**better**
14B bd, CV	114	1.84	2.34	2.29	$-0.51\ [-0.93, -0.10]$	**better**
30B bd, CV	114	1.99	2.57	2.35	$-0.37\ [-0.81, +0.08]$	= grader
Gemma bd, CV	114	2.22	2.53	2.58	$-0.27\ [-0.71, +0.17]$	= grader
Llama bd, CV	114	2.04	2.54	2.38	$-0.37\ [-0.80, +0.06]$	= grader
7B bd, ML	208	4.19	5.09	4.54	$-0.37\ [-1.06, +0.31]$	= grader
14B bd, ML	208	4.02	4.87	4.49	$-0.51\ [-1.16, +0.14]$	= grader
30B bd, ML	208	4.07	5.07	4.50	$-0.40\ [-1.08, +0.28]$	= grader
Gemma bd, ML	208	4.22	5.09	4.54	$-0.37\ [-1.05, +0.31]$	= grader
Llama bd, ML	208	4.25	5.05	4.71	$-0.31\ [-0.98, +0.37]$	= grader

Table 20: Per-question mean signed error for the headline runs of both exams, over exactly the students each run's published metric uses. b_q is the mean of $\mathrm{AI}_q - \bar{\mathrm{G}}_q$ on question q (CV exam: 35-point base scale, Q4 bonus excluded; ML exam: score plus bonus, ≈ 65-point scale). *Total* is the run's published bias. $C = \sum_q |b_q| - |\sum_q b_q|$ is the cancellation the total hides, zero exactly when every question is pushed the same way. At two decimals a row can miss $\sum_q b_q =$ *Total* by 0.01 through rounding. Generated by `analysis/checks/per_question_signed_error.py`.

Run	Configuration	n	b_1	b_2	b_3	Total	$\sum_q \lvert b_q \rvert$	C
CV exam — 35-point base scale (Q1 12, Q2 11, Q3 12)								
D01	3-flash-preview, neutral	570	−0.15	+0.08	+0.08	+0.01	0.30	0.29
D02	3.1-pro-preview, neutral	570	−0.44	−0.07	−0.31	−0.82	0.82	0.00
F02	3.1-pro-preview + lenient	570	−0.01	+0.28	+0.55	+0.82	0.84	0.02
B03	flash-lite + thinking	570	−0.50	+0.04	+0.30	−0.15	0.85	0.69
A01	flash-lite, neutral (baseline)	570	−0.28	−2.12	+0.63	−1.78	3.03	1.25
C01	flash-lite + strict	570	−1.80	−3.08	−0.55	−5.44	5.44	0.00
ML exam — ≈ 65-point scale (Q1 26, Q2 17, Q3 22)								
IG08	3.1-pro-preview, neutral	1038	+0.43	+0.20	+0.55	+1.18	1.18	0.00
IG07	3-flash-preview, neutral	1038	+1.55	+0.10	+0.59	+2.24	2.24	0.00
IG01	flash-lite, neutral	1013	+4.00	+0.79	+1.19	+5.98	5.98	0.00
IG05	flash-lite + strict	1008	−1.97	−2.82	−3.04	−7.83	7.83	0.00

F Per-Question Error Decomposition

Table 20 decomposes the mean signed error per question for the headline runs of both exams; both sides are sums of per-question marks, so the decomposition is exact, reproducing each run's published bias (Tables 10, 24 and 25) to within 4×10^{-15}.

Across the 44 runs decomposed — the ten in the table plus the 17 matched neutral/strict open-weights pairs of Table 3 — the largest cancellation is L-A14 (Qwen2.5-Coder-14B, neutral): per-question biases of $+0.89$, -1.30 and -0.75 sum to -1.16, hiding 1.78 of 2.94 points of directional error (61%); the largest in the table is A01, at 1.25 of 3.03. Cancellation is confined to the near-unbiased runs: all 17 matched *strict* runs and all four ML-exam Gemini runs have $C = 0$, so no persona-collapse claim rests on a cancelled total.

G GROUND-TRUTH STRUCTURE

Two checks probe the grader-average target, with D01 as the AI side and, on the ML exam, its best full-cohort run (IG08, `gemini-3.1-pro-preview`).

Whether the AI inherits per-pair human noise is exam-specific. If the AI inherited its target's noise structure, its per-pair error should track the human MAE, which spans 0.85 to 3.55. It does not ($r = -0.301$, $p = 0.40$; Table 21, Figure 4). Agreement with any one grader's judgements is a different metric, not measurable here.

And it does not generalise. On the ML exam's 18 stable pairs the test inverts — $r = +0.571$ (permutation $p = 0.014$), a difference itself significant (Fisher $z = 2.10$, $p = 0.036$), and the sign holds for every AI side we tried ($+0.57$ to $+0.72$). Some coupling must be arithmetic, since a noisier pair injects its variance into the target, but the correction is not clean enough to rest on; we report the divergence, not a mechanism; the ML exam's contiguous-block grading assignment is one candidate (Appendix H).

Per-grader harshness is real, modest, and best measured within pairs. The 20 graders span a 6.55-point range in mean awarded total. A one-way ANOVA over the $1{,}140$ grades puts that at $\eta^2 = 0.051$ ($p = 5 \times 10^{-6}$), but those grades are not independent: each submission is marked by both members of one fixed pair, and each pair marks its own batch of 55–60 students, confounding between-pair harshness with batch ability (Appendix H). The η^2 splits orthogonally into 0.034 between pairs and 0.017 within pairs, the only part identified from the *same* papers. A linear mixed model total $\sim$ grader with a random intercept per student (student variance 48.69, residual 5.73, ICC $= 0.89$) gives the ten within-pair contrasts standard errors of 0.44–0.46 against 1.36–1.40 for the between-pair ones, and the likelihood-ratio tests separate them: collapsing graders to pair means costs little ($\chi^2(9) = 21.1$, $p = 0.012$), dropping the within-pair contrasts costs a great deal ($\chi^2(10) = 168.9$, $p = 5 \times 10^{-31}$; blocked-ANOVA partial $\eta^2 = 0.26$ of the paired-difference variance). Within pairs the mean signed difference runs -2.75 to $+3.11$ points, six of ten pairs are non-zero after Holm correction, and a sign-flip omnibus over the 570 paired differences gives $p < 10^{-4}$. A student's expected score does depend on which pair graded them, and the AI's MAE incorporates that noise in its target; but only the within-pair third of the 5% is measurable as harshness rather than as batch ability. That η^2 is design-dependent and does not transfer; the quantity that does — signed bias between two graders on the *same* papers — agrees across exams at $0.61\times$ the floor (Appendix H).

Together the two results say the best closed-model grader acts as a MAE ≈ 1.6 stabiliser around a noisy human consensus on the CV exam, sitting below all but the tightest pair's disagreement bar.

The largest single-paper disagreement is 24.1 points. The one-way tests use all $1{,}140$ grades (570 students $\times$ two graders), and the non-parametric Kruskal–Wallis test agrees with the ANOVA ($H = 53.80$, $p = 3.5 \times 10^{-5}$); both share the independence assumption that the paired analysis above replaces. Neither that analysis nor the one-way tests turn on the one negative total (a TA_1 total of -1.8): removing it gives $H = 54.34$, a within-pair η^2 share of 0.017, and $\chi^2(10) = 171.6$.

Table 21: Per-pair human disagreement vs. D01's per-pair AI MAE, sorted by human MAE.

Pair	n	Human MAE	D01 AI MAE	D01 AI bias
G_6 / G_{16}	56	0.85	1.86	$+1.18$
G_9 / G_{19}	55	1.87	1.43	-0.46
G_5 / G_{15}	57	2.12	1.95	-0.90
G_4 / G_{14}	56	2.25	1.40	-0.88
G_7 / G_{17}	57	2.48	1.84	$+1.11$
G_8 / G_{18}	58	2.85	1.83	-0.34
G_{10} / G_{20}	60	3.05	1.85	-0.91
G_2 / G_{12}	58	3.45	1.02	$+0.12$
G_3 / G_{13}	56	3.54	1.92	$+1.63$
G_1 / G_{11}	57	3.55	1.32	-0.36

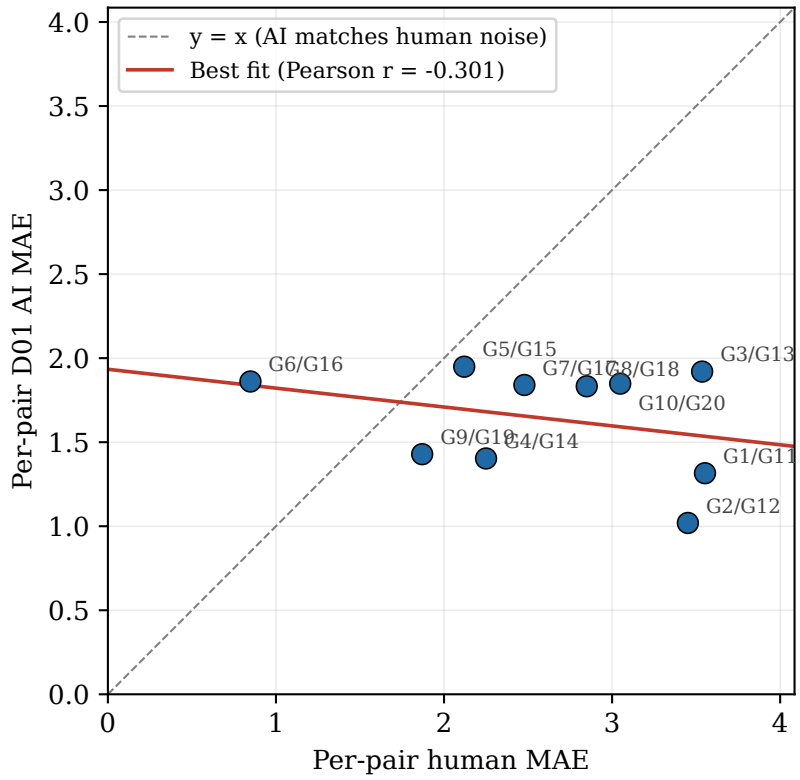

Figure 4: Per-pair human disagreement (x-axis) against D01's per-pair MAE (y-axis), one point per grader pair. The two are uncorrelated (Pearson $r = -0.30$, $p = 0.40$): the AI grader sits roughly equidistant from all pairs rather than inheriting pair-level noise.

H ML-EXAM REPLICATION DETAILS

The ML exam of Section 5.4 is an introductory AI practical from another course at the same institution: 1,038 students after excluding 24 whose uploads contained no gradable notebook, dual-graded by 49 graders, three questions (regression, PyTorch, classification) worth 23+3, 14+3 and 19+3 marks including bonuses (56 base points plus up to 9 bonus points, 65 in all). Question maxima come from the marking scheme's task tables, not its header totals, which contradict them in all three questions; two instructor corrections issued by email mid-grading are folded in, both affecting Q3. Per-question grader marks fold bonus into the score, so the comparable AI total is score plus bonus. A student who did not submit a question is scored 0 by graders and models alike.

The grading instrument is identical to Section 3 (same persona strings, JSON schema, output rules and grading scale, reference solution, rubric breakdown, no chain-of-thought, temperature 0, one sample per question), except that this exam has no student-facing guidelines document, so the `gd` component is fixed at 0 throughout.

The grid. The replication comprises 162 configurations — 30 closed-model (23 Gemini IG-, 6 OpenAI IO-, 1 Anthropic IN-) and 132 open-weights (IA-) — replaying the CV exam's programme: neutral and matched *strict* runs for all 17 open-weights models, five-persona sweeps on 13, the mechanism probes of Section 5.2, component removals, few-shot demonstrations on one closed and one open model, temperature-variance probes, and a closed arm covering all four Gemini models, `gpt-5.5`, `gpt-5.4` and `claude-opus-5`. Table 25 (Appendix K) lists every run.

Table 22: ML exam, 1,038 dual-graded students: matched neutral/strict pairs for all 17 open-weights models, sorted by damage ratio. MAE against the grader average on a $\approx$ 65-point scale, 95% bootstrap CI on the strict run, strict MAE as a multiple of the ML exam's human floor (5.13, defined as in Section 4). Behaviour classes are those of Section 5.1 and describe the strict run, at this exam's floor-matched band MAE $\geq$ 15.7 (3.07$\times$ its 5.13 floor, the multiple the CV exam's 8 fixes); four strict runs and one neutral baseline cross it. A fixed MAE $\geq$ 8 would be only 1.56$\times$ this floor and would label ten strict runs and eleven neutral baselines collapse. A refusal's MAE is a ceiling artifact, not a severity measurement. Ratios below 1 are real improvements, examined below. Generated by `analysis/introduction_to_ai_make_paper_tables.py`.

model	neutral	strict	strict CI	ratio	$\times$floor	zero-rate	bias	behaviour
GLM-4.5-Air	5.60	18.49	[17.96, 19.02]	$\times$3.30	3.6	37.8%	-18.31	collapse
Qwen2.5-Coder-32B	4.57	13.24	[12.79, 13.69]	$\times$2.90	2.6	10.5%	-13.07	graded
Llama-3.1-8B	13.16	29.58	[28.70, 30.45]	$\times$2.25	5.8	100.0%	-29.58	**refusal**
Llama-3.3-70B	4.70	10.17	[9.77, 10.58]	$\times$2.16	2.0	9.0%	-9.63	graded
Mistral-Small-24B	8.96	17.32	[16.77, 17.87]	$\times$1.93	3.4	32.2%	-17.20	collapse
Qwen3-Coder-Next 80B	4.47	8.63	[8.26, 9.00]	$\times$1.93	1.7	2.3%	-8.12	graded
GLM-4-32B	8.03	12.18	[11.68, 12.69]	$\times$1.52	2.4	14.1%	-10.68	graded
Qwen2.5-Coder-14B	8.61	11.06	[10.61, 11.50]	$\times$1.28	2.2	16.6%	-10.46	graded
GLM-4-9B	13.15	16.68	[16.15, 17.20]	$\times$1.27	3.2	35.7%	-16.52	collapse
Qwen2.5-Coder-7B	6.88	7.45	[7.11, 7.80]	$\times$1.08	1.5	7.4%	-5.67	graded
Qwen3-235B-A22B	5.40	5.18	[4.89, 5.47]	$\times$0.96	1.0	1.7%	-3.39	graded
Gemma-3-27B	8.33	7.61	[7.23, 7.99]	$\times$0.91	1.5	0.6%	-6.25	graded
Gemma-3-12B	10.36	6.79	[6.49, 7.10]	$\times$0.66	1.3	0.5%	$+0.09$	graded
Qwen3-Coder-480B	10.10	6.04	[5.75, 6.34]	$\times$0.60	1.2	1.0%	$+3.25$	graded
DeepSeek-Coder-V2-Lite	16.41	9.22	[8.83, 9.62]	$\times$0.56	1.8	0.0%	$+2.42$	graded
Qwen2.5-72B	10.64	5.42	[5.15, 5.70]	$\times$0.51	1.1	2.3%	-1.67	graded
Qwen3-Coder-30B-A3B	11.12	5.58	[5.31, 5.87]	$\times$0.50	1.1	2.2%	-1.55	graded

When severity helps, it is calibration, not robustness. Fifteen of the 17 neutral baselines over-mark, with biases from $+1.51$ to $+14.89$ (the exceptions are Qwen2.5-Coder-32B at -0.21 and GLM-4.5-Air at -1.34), and across the 16 non-refusal pairs a model's neutral bias predicts the strict effect at $r = -0.73$ (Spearman -0.73, permutation $p = 0.002$; $r = -0.50$ with the refusal's ceiling-artifact MAE included), whereas the CV exam's neutral biases sit at or below zero for all but two models and zero of 17 improve. The seven improvements in Table 22 are heavy over-markers pushed down — Qwen3-Coder-30B-A3B (bias $+10.44$), Qwen2.5-72B ($+10.31$), Qwen3-Coder-480B ($+9.52$) — two miscalibrations cancelling: the 480B is one such case (MAE $= 10.10$ at neutral,

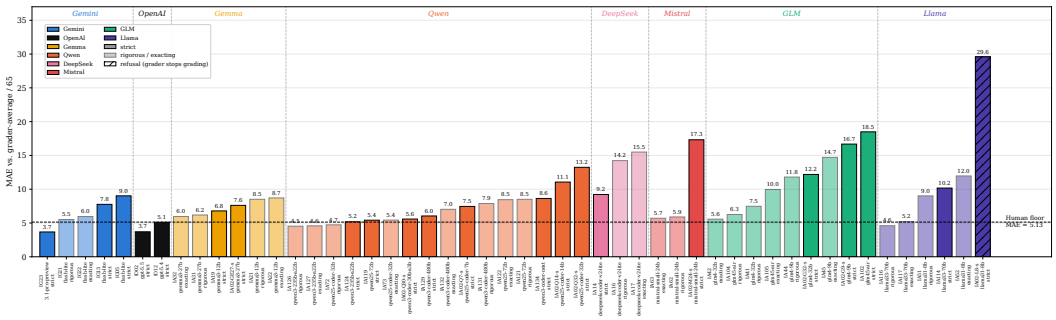

Figure 5: Every strict-flavoured persona run on the ML exam's full cohort (50 runs: *strict*/*rigorous*/*exacting* at the default prompt configuration, including IG13, which additionally enables thinking), coloured by family and by wording (*strict* darker; *rigorous*/*exacting* lighter). The dashed line is the ML exam's human inter-grader floor (MAE $= 5.13$). The refusal (Llama-3.1-8B + *strict*) is hatched: its bar height is the distance to the grader mean, not a severity measurement. The ML exam's counterpart to Figure 2.

twice its floor), whereas Qwen3-235B-A22B is genuinely indifferent to the sentence on both exams (here bias $+3.73$, strict effect -0.22).

Mechanism. The probe of Section 5.3 separates the same two shapes on the ML exam (Table 23). It is read only for runs that left the graded band and zeroed at least 5% of Q1, the volume gate the CV exam's classifier applies at 10%; two runs fall below it, DeepSeek-Coder-V2-Lite (6 of 1038, 0.6%) and Qwen3-Coder-Next 80B (39 of 1001, 3.9%).

Table 23: ML exam, *strict* runs: blanket versus selective zeroing, all 17 models. Q1 $= 0$ is the count of students awarded zero on Q1 out of those the run graded; the next column is how many of *those* still received a non-zero Q2. A run is read only if it left the graded band *and* zeroed Q1 on at least 5% of the students it graded (the CV exam applies the same volume gate at 10%, `MECH_ZERO_RATE`; on these runs any cut between 4.8% and 13.3% gives the same partition): below that, the zeros are ordinary non-attempts and the share is noise on a handful of students. Generated by `analysis/introduction_to_ai_make_paper_tables.py`.

model (strict)	Q1 $= 0$	of those, Q2 > 0	share	reading
Llama-3.1-8B	1038 / 1038 (100%)	0	0%	blanket
GLM-4-9B	669 / 1038 (64%)	255	38%	selective
GLM-4.5-Air	521 / 1038 (50%)	110	21%	blanket
GLM-4-32B	477 / 1038 (46%)	295	62%	n/a (in band)
Mistral-Small-24B	375 / 1038 (36%)	21	6%	blanket
Qwen2.5-Coder-14B	196 / 1038 (19%)	13	7%	n/a (in band)
Llama-3.3-70B	152 / 1038 (15%)	40	26%	n/a (in band)
Qwen2.5-Coder-32B	144 / 1038 (14%)	19	13%	n/a (in band)
Qwen2.5-Coder-7B	138 / 1038 (13%)	59	43%	n/a (in band)
Qwen3-Coder-30B-A3B	50 / 1038 (5%)	24	48%	n/a (in band)
Qwen2.5-72B	48 / 1038 (5%)	21	44%	n/a (in band)
Qwen3-Coder-Next 80B	39 / 1001 (4%)	13	33%	n/a (in band)
Qwen3-235B-A22B	30 / 1038 (3%)	10	33%	n/a (in band)
Gemma-3-12B	21 / 1038 (2%)	14	67%	n/a (in band)
Qwen3-Coder-480B	19 / 1038 (2%)	7	37%	n/a (in band)
Gemma-3-27B	15 / 1038 (1%)	7	47%	n/a (in band)
DeepSeek-Coder-V2-Lite	6 / 1038 (1%)	4	67%	n/a (in band)

The attribution series replicates. The minimal pairs of Section 5.2 reproduce both findings. S2 remains the only single component that stops graders grading: alone it takes Llama-3.1-8B to 29.39 (98% of students zeroed) and Mistral-Small-24B to 26.03 (76% zeroed), while the HARSH frame alone moves each of the three models probed by at most 2.4 points from neutral, twice in

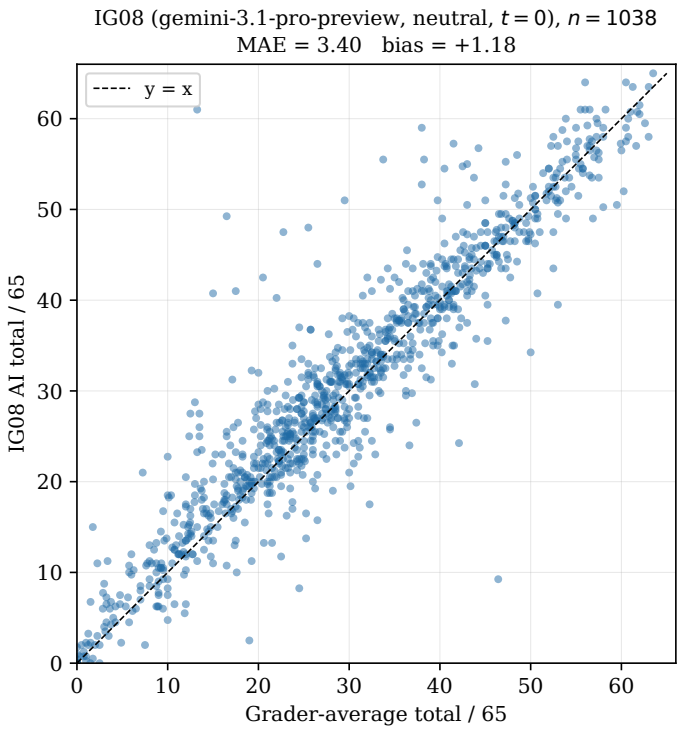

Figure 6: IG08 (`gemini-3.1-pro-preview`, neutral, $t = 0$): AI total vs. grader-average total over the ML exam's full $n = 1{,}038$ cohort, with $y = x$ as the perfect-agreement reference. The best full-cohort configuration on the ML exam ($0.66\times$ its human floor), the counterpart to Figure 3.

the *improving* direction. Nor does the adjective carry the damage: with both policy sentences held verbatim, swapping STRICT for RIGOROUS or FAIR moves MAE by 0.55–3.86 across the five models with no consistent direction — a FAIR teaching assistant carrying the harsh policy still more than doubles Qwen2.5-Coder-32B's error (10.40 against 4.57) — while the *nharsh* control (a harsh adjective with no policy) lands within 1.2 points of neutral on all five. Absolute damage is exam-dependent — S2 alone leaves Qwen2.5-Coder-32B in the graded band here (6.68, against 12.92 on the CV exam) — but which sentence dominates a model is not: S1 outweighs S2 for the 32B and S2 for the other two, on both exams.

The closed-model arm. The ML exam also carries all four Gemini models under the same instrument. The parity result replicates on the closed side — all three full-cohort neutral configurations sit wholly below the floor's CI: `gemini-3.1-pro-preview` at 3.40 $[3.16, 3.65]$ (IG08; Figure 6), `gemini-2.5-pro` at 3.47, and `gemini-3-flash-preview` at 3.92, against a floor of 5.13 $[4.80, 5.48]$; the 3.1-pro under *lenient* (4.48 $[4.22, 4.75]$) and under *strict* (3.67 $[3.43, 3.94]$, bias -1.00 against $+1.18$ at neutral; IG23) join them. The cross-vendor runs of Section 6.1 agree: no other closed run leaves this exam's band (Table 9). What does *not* replicate is Gemini's persona immunity at the cheaper tier. Flash-Lite, at 7.53 under neutral (bias $+5.98$), moves in *both* directions: *strict* takes it to 9.02 (bias -7.83) and *lenient* to 17.86 (bias $+17.80$), the one closed run past this exam's band, which only four open-weights strict runs cross — while *rigorous* and *exacting* improve it (5.52, 5.96), the same calibration correction the open side shows. The stronger `gemini-3-flash-preview` follows the same law on its first-100 subset: against a matched-subset baseline of 3.87 (bias $+3.04$), *strict* lands at 2.87 (bias $+0.09$) and *lenient* at 6.99 (bias $+6.60$). Component ablations echo the CV exam with larger magnitudes: dropping the reference solution costs Flash-Lite 4.93 points ($7.53 \rightarrow 12.46$) against 0.79 there, and thinking helps by 2.16 ($7.53 \rightarrow 5.37$) against 1.29.

Few-shot demonstrations transfer, with one measurement caveat. Two worked examples per question (from the two highest-agreement grader pairs, labelled by IG08; the three students remain

in the evaluated cohort, $3/1{,}038$) nearly halve the open model's error: Qwen3-Coder-30B-A3B goes from 11.12 to 5.84, bias $+10.44$ to $+4.56$, on the full cohort (IA201). On Flash-Lite (IG18) they lengthen the prompt enough that 334 students' calls fail permanently, with biased dropout: the dropped students average 32.78 points against the survivors' 28.02 (Welch $p = 3 \times 10^{-7}$). On the matched 679-student subsample the effect holds ($7.87 \to 4.32$ MAE, bias $+6.52 \to +0.37$), but its headline numbers are not comparable to full-cohort rows and Table 25 flags its n.

Determinism at $t = 0$ does not transfer. On the CV exam Flash-Lite reproduced 100% of (student, question) cells exactly across five $t = 0$ reruns (Appendix C). Here the same probe (IG10) reproduces only 59.3%; the median per-cell standard deviation is still 0.00 but a 40.7% tail varies, and at $t = 0.5/0.7$ the medians rise to $0.98/1.14$ (IG11/IG12) against $0.65/0.76$ there. Two open-weights probes on 41 students at $t = 0.5$ (IA92, Qwen2.5-Coder-32B; IA136, Qwen3-Coder-Next) show medians of 0.65 and 0.84.

Component removals do not transfer either. The CV exam's over-anchoring result ($7.49 \to 3.12$ on Qwen2.5-Coder-32B with the reference solution removed) does not reproduce: the same removal moves it $4.57 \to 4.73$ here. The rubric-breakdown test of Appendix B softens: under *strict*, removing the breakdown worsens the collapse on GLM-4-32B ($+6.26$) and Llama-3.3-70B ($+3.96$) but is flat on Qwen2.5-Coder-32B ($+0.19$), Gemma-3-27B (-0.26) and DeepSeek-Coder-V2-Lite (-0.10), neutral controls within ± 0.52 — against five-of-five worsening there.

Ground-truth structure of the ML exam. Its floor is $5.13/65$ (95% bootstrap CI $[4.80, 5.48]$ at $50{,}000$ resamples — the lower bound is not stable to two decimals at $2{,}000$; Pearson $r = 0.871$). Nine of the $1{,}038$ rows record 0.0 for exactly one grader while the partner awarded real marks: ungraded slots stored as zeros. They move the floor by 0.05 (inside the CI, so the headline retains them) but dominate the extremes, the maximum single-paper disagreement being 41.0 without them and 51.0 with. Excluding them everywhere changes no conclusion: the floor falls to 5.09 $[4.76, 5.42]$, the best full-cohort configurations move by at most 0.06 MAE (IG08 $3.40 \to 3.35$, IG09 $3.47 \to 3.42$, IG07 $3.92 \to 3.86$, Qwen3-Coder-Next neutral $4.47 \to 4.44$), no run of Table 22 or Table 25 changes behaviour class or damage-ratio order (largest move 0.21 MAE, the Llama-3.1-8B refusal), and on the 208 held-out students (three of the nine; the other six sit in the 830-student fine-tuning split) the pooled adapters grade 0.08–0.11 MAE better, every Table 19 verdict intact (`analysis/checks/ml_unrecorded_grader_rows.py`). A further nine rows carry 0.0 from both graders and are genuine non-submissions, retained.

The pairing is hybrid, unlike the CV exam's 10 fixed pairs: 49 assistants work in 48 grading slots (one slot is a pair who marked jointly) forming 51 pairings, of which 18 are stable pairs of 38–51 students covering 816 of $1{,}038$ (79%); the rest are ad hoc combinations around floating graders, and two of the 51 on the roster returned no marks. Every per-pair figure is scoped to that backbone, across which human MAE runs 1.90 to 10.36 — a $5.4\times$ spread, against $4.2\times$ on the CV exam — with eight of 18 pairs above the pooled floor, against five of ten. Mean within-pair bias magnitude is 3.14 ($0.61\times$ floor) against 1.58 ($0.61\times$ floor): systematic harshness is the same fraction of the noise floor on both exams, whereas a one-way η^2 over grader identity is not comparable here, because each pair marks a contiguous block of students of widely varying ability.

The per-pair test of Appendix G inverts here: per-pair human MAE against IG08's per-pair MAE gives Pearson $r = +0.571$ (permutation $p = 0.014$, $20{,}000$ shuffles; Spearman $+0.447$; Figure 7) against -0.301 on the CV exam, a difference significant at Fisher $z = 2.10$, $p = 0.036$, and the sign survives changing the AI side (Qwen3-Coder-Next neutral $+0.565$, Qwen2.5-Coder-32B $+0.723$). Some of the coupling is arithmetic, a noisier pair injecting its variance into the target. One candidate for the rest is the assignment itself: ML pairs mark contiguous blocks of students whose ability varies widely, so block difficulty raises both human disagreement and model error within a pair, whereas the CV exam's fixed pairs each drew a comparable slice. The AI is still the more uniform grader, its per-pair MAE spanning $3.6\times$.

What the replication does and does not establish. *Strict* worsens ten of 17 models and takes four above this exam's floor-matched band (MAE ≥ 15.7), none of them there at neutral. What transfers is the floor-relative attainability of parity (IG08 at $0.66\times$ this exam's floor; the best open-weights baseline, Qwen3-Coder-Next, 4.47 $[4.24, 4.72]$, also wholly below its CI), the S2 mechanism,

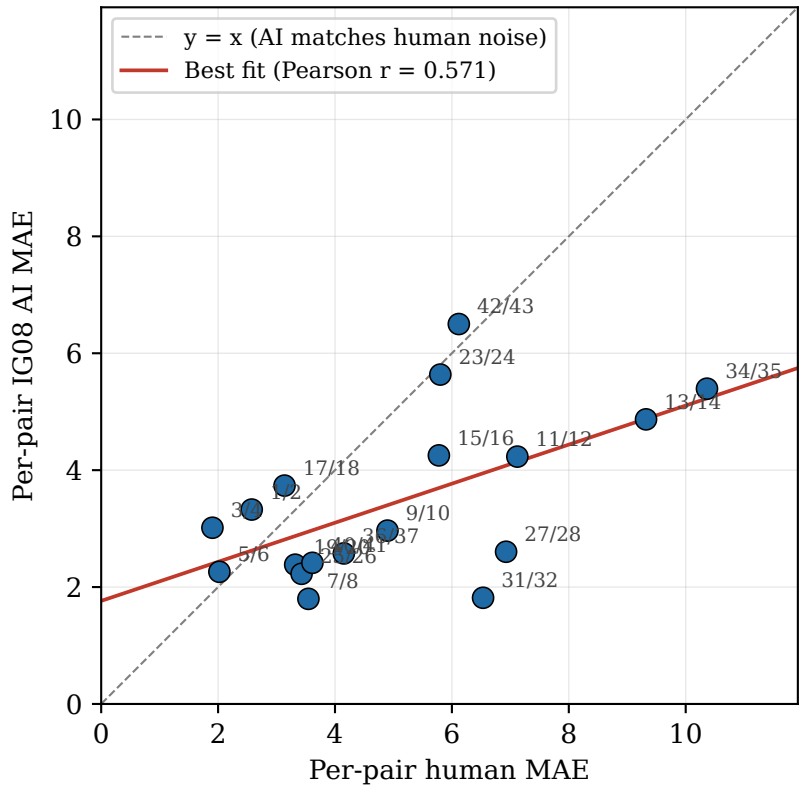

Figure 7: ML exam: per-pair human disagreement (x-axis) against IG08's per-pair MAE (y-axis) over the 18-pair stable backbone. Unlike the CV exam (Figure 4), the two correlate ($r = +0.571$): here the AI's error is largest exactly where the humans disagree most.

the blanket/selective split and the danger of an untested persona sentence; model choice, persona direction, component effects, failure shape (Qwen2.5-Coder-32B inverts from selective field collapse to blanket zeroing) and $t = 0$ determinism are exam-level properties.

I ALL GRADER CONFIGURATIONS

Table 24 lists every one of the 171 grader configurations with its headline metrics, one row per run. The machine-readable version, with per-question breakdowns and the full column set, ships as `analysis/computer_vision_master_comparison.csv` in the supplementary material.

Table 24: Every grader configuration in the study, one row per run: the 33 closed-model configurations (25 Gemini via Vertex, 6 OpenAI and 2 Anthropic via their batch APIs; Section 6.1) followed by the 138 open-weights (L-) configurations, each sorted by run id. *Prompt* lists the components present (S reference solution, G grading guidelines, B rubric breakdown, R thinking enabled, F few-shot demonstrations); personas prefixed `mp` are the mechanism probes of Section 5.2: `mpstrict`, `mprigorous` and `mpfair` keep the *strict* preset's two policy sentences and swap only its headword; `mpframe` is the HARSH frame alone, `mpnoclause` the frame with S1, `mps2only` the frame with S2, `mpharsh` the frame with both (the *strict* text, rerun on the current stack); `mpnharsh` is the neutral frame with "harsh" in place of "strict". L-MP04 and L-PS03 are one physical run (`qwen2.5-coder-32b`, *mpnoclause*), listed under both the headword series and the policy 2×2. n is the number of students scored against the grader average; MAE and bias are on the 35-point base scale, the 95% bootstrap CI on MAE; behaviour classes are defined in Section 5.1. A refusal's MAE is a ceiling artifact, not a severity measurement. D03 ran with thinking active regardless of its flag (`gemini-2.5-pro` cannot disable it; Appendix A). `claude-opus-5` exposes no sampling temperature (thinking disabled; shown as —). The final block lists the 8 prompt-only baselines of Appendix B (first 100 students; the persona column names the strategy), which are not configurations of the paper's grader and are excluded from every count; their CIs use 2000 resamples. Model keys match the released spreadsheets. Generated by `analysis/computer_vision_make_appendix_table.py`.

Run	Model	Persona	Prompt	t	n	MAE	95% CI	Bias	Behaviour
Closed-model configurations (Gemini: A–P series; OpenAI: O; Anthropic: N)									
A01	`flash-lite`	neutral	SGB	0	570	3.34	[3.15, 3.53]	−1.78	graded
B01	`flash-lite`	neutral	GB	0	567	4.13	[3.83, 4.42]	+3.30	graded

Table 24 continued

Run	Model	Persona	Prompt	t	n	MAE	95% CI	Bias	Behaviour
B02	flash-lite	neutral	SB	0	569	3.34	[3.16, 3.52]	−1.42	graded
B03	flash-lite	neutral	SGBR	0	570	2.05	[1.89, 2.21]	−0.15	graded
B04	flash-lite	neutral	SG	0	569	3.22	[3.03, 3.43]	−1.45	graded
C01	flash-lite	strict	SGB	0	570	5.75	[5.46, 6.04]	−5.44	graded
C02	flash-lite	lenient	SGB	0	569	4.49	[4.18, 4.82]	+4.01	graded
D01	3-flash-preview	neutral	SGB	0	570	1.64	[1.51, 1.79]	+0.01	graded
D02	3.1-pro-preview	neutral	SGB	0	570	1.86	[1.70, 2.02]	−0.82	graded
D03	2.5-pro	neutral	SGB	0	570	4.10	[3.83, 4.38]	−3.85	graded
E01	flash-lite	neutral	SGB	0	549	3.38	[3.19, 3.59]	−1.85	graded
E02	flash-lite	neutral	SGB	0.5	563	3.51	[3.31, 3.71]	−2.20	graded
E03	flash-lite	neutral	SGB	0.7	563	3.60	[3.40, 3.79]	−2.32	graded
F01	flash-lite	strict	SGBR	0	570	3.73	[3.50, 3.98]	−3.28	graded
F02	3.1-pro-preview	lenient	SGB	0	570	1.79	[1.62, 1.95]	+0.82	graded
F03	3.1-pro-preview	strict	SGB	0	570	2.75	[2.51, 2.99]	−2.09	graded
G01	3-flash-preview	neutral	GB	0	98	1.63	[1.32, 2.02]	+0.71	graded
G02	3-flash-preview	neutral	SB	0	99	1.60	[1.30, 1.93]	+0.79	graded
G03	3-flash-preview	neutral	SGBR	0	100	1.39	[1.18, 1.61]	−0.12	graded
G04	3-flash-preview	neutral	SG	0	100	1.25	[1.01, 1.51]	−0.07	graded
K01	flash-lite	neutral	SGBF	0	567	3.03	[2.83, 3.24]	+1.09	graded
M01	3-flash-preview	strict	SGB	0	99	4.23	[3.60, 4.88]	−4.16	graded
M02	3-flash-preview	lenient	SGB	0	100	2.18	[1.75, 2.65]	+1.90	graded
N01	claude-opus-5	neutral	SGB	—	570	3.54	[3.31, 3.77]	−3.24	graded
N02	claude-opus-5	strict	SGB	—	570	5.17	[4.90, 5.45]	−5.02	graded
O01	gpt-5.5	neutral	SGB	0	570	2.43	[2.28, 2.59]	−1.65	graded
O02	gpt-5.5	strict	SGB	0	570	4.43	[4.19, 4.68]	−4.25	graded
O03	gpt-5.5	lenient	SGB	0	570	2.25	[2.08, 2.44]	+1.13	graded
O11	gpt-5.4	neutral	SGB	0	570	4.69	[4.46, 4.92]	−4.45	graded
O12	gpt-5.4	strict	SGB	0	570	6.90	[6.63, 7.15]	−6.81	graded
O13	gpt-5.4	lenient	SGB	0	570	2.69	[2.51, 2.89]	+1.00	graded
P01	flash-lite	rigorous	SGB	0	570	3.92	[3.70, 4.16]	−3.01	graded
P02	flash-lite	exacting	SGB	0	557	4.58	[4.35, 4.85]	−4.05	graded
Open-weights (L-) configurations									
L-A07	qwen2.5-coder-7b	neutral	SGB	0	570	5.68	[5.39, 5.98]	−4.84	graded
L-A14	qwen2.5-coder-14b	neutral	SGB	0	570	3.51	[3.30, 3.74]	−1.16	graded
L-A30M	qwen3-coder-30b-a3b	neutral	SGB	0	570	4.50	[4.29, 4.72]	−2.70	graded
L-A32	qwen2.5-coder-32b	neutral	SGB	0	570	7.49	[7.19, 7.78]	−7.35	graded
L-ANxt	qwen3-coder-next	neutral	SGB	0	570	5.17	[4.92, 5.42]	−4.74	graded
L-B01	qwen2.5-coder-32b	neutral	GB	0	570	3.12	[2.92, 3.31]	−1.56	graded
L-B02	qwen2.5-coder-32b	neutral	SB	0	570	5.86	[5.60, 6.11]	−5.46	graded
L-B03	qwen2.5-coder-32b	neutral	SGBR	0	570	7.51	[7.22, 7.80]	−7.36	graded
L-B04	qwen2.5-coder-32b	neutral	SG	0	570	7.27	[6.98, 7.55]	−7.11	graded
L-BD01	qwen2.5-coder-32b	neutral	SG	0	570	7.36	[7.06, 7.64]	−7.23	graded
L-BD02	qwen2.5-coder-32b	strict	SG	0	570	21.69	[21.16, 22.23]	−21.68	collapse
L-BD03	llama-3.3-70b	neutral	SG	0	570	4.69	[4.46, 4.92]	−3.96	graded
L-BD04	llama-3.3-70b	strict	SG	0	570	19.94	[19.43, 20.51]	−19.94	collapse
L-BD05	deepseek-coder-v2-lite	neutral	SG	0	570	6.83	[6.51, 7.17]	−5.03	graded
L-BD06	deepseek-coder-v2-lite	strict	SG	0	570	16.02	[15.55, 16.49]	−15.99	collapse
L-BD07	glm-4-32b	neutral	SG	0	570	2.68	[2.49, 2.88]	+0.70	graded
L-BD08	glm-4-32b	strict	SG	0	570	15.58	[14.99, 16.20]	−15.46	collapse
L-BD09	gemma-3-27b	neutral	SG	0	570	4.32	[4.11, 4.56]	−1.64	graded
L-BD10	gemma-3-27b	strict	SG	0	569	11.69	[11.31, 12.11]	−11.64	collapse
L-C01	qwen2.5-coder-32b	strict	SGB	0	570	20.29	[19.83, 20.74]	−20.28	collapse
L-C02	qwen2.5-coder-32b	lenient	SGB	0	570	3.95	[3.75, 4.17]	−2.15	graded
L-D30M	qwen3-coder-30b-a3b	neutral	SGBR	0	570	4.49	[4.28, 4.71]	−2.70	graded
L-DNxt	qwen3-coder-next	neutral	SGBR	0	570	5.18	[4.93, 5.43]	−4.75	graded
L-E01	qwen2.5-coder-32b	neutral	SGB	0.5	50	7.15	[6.26, 8.04]	−7.02	graded
L-E02	qwen3-coder-next	neutral	SGB	0.5	48	5.38	[4.64, 6.16]	−5.20	graded
L-F01	gemma-3-12b	neutral	SGB	0	570	5.20	[4.94, 5.48]	−2.75	graded
L-F02	gemma-3-12b	strict	SGB	0	570	8.77	[8.42, 9.15]	−8.43	collapse
L-F03	gemma-3-12b	lenient	SGB	0	570	4.79	[4.54, 5.06]	−0.82	graded
L-F04	gemma-3-12b	rigorous	SGB	0	570	5.45	[5.19, 5.73]	−3.19	graded
L-F05	gemma-3-12b	exacting	SGB	0	570	5.10	[4.85, 5.38]	−2.76	graded
L-G01	qwen3-coder-30b-a3b	neutral	GB	0	570	3.60	[3.37, 3.86]	+1.29	graded
L-G02	qwen3-coder-30b-a3b	neutral	SB	0	570	3.63	[3.42, 3.85]	−0.18	graded
L-G04	qwen3-coder-30b-a3b	neutral	SG	0	570	4.80	[4.59, 5.03]	−3.49	graded
L-H01	qwen2.5-coder-7b	neutral	GB	0	570	4.29	[4.00, 4.62]	+2.16	graded
L-H02	qwen2.5-coder-7b	neutral	SB	0	570	5.25	[4.99, 5.53]	−4.02	graded
L-H04	qwen2.5-coder-7b	neutral	SG	0	570	5.83	[5.56, 6.14]	−4.92	graded
L-J01	glm-4.5-air	neutral	SGB	0	570	5.66	[5.38, 5.95]	−5.37	graded
L-J02	glm-4.5-air	strict	SGB	0	570	21.02	[20.49, 21.55]	−21.02	collapse
L-J03	glm-4.5-air	lenient	SGB	0	570	4.68	[4.39, 4.99]	+4.51	graded
L-J04	glm-4.5-air	rigorous	SGB	0	568	7.29	[6.99, 7.64]	−7.21	graded
L-J05	glm-4.5-air	exacting	SGB	0	570	11.29	[10.92, 11.65]	−11.27	collapse
L-K01	qwen3-coder-30b-a3b	neutral	SGBF	0	570	3.34	[3.16, 3.53]	−1.77	graded
L-M01	qwen3-coder-30b-a3b	strict	SGB	0	570	14.91	[14.52, 15.30]	−14.91	collapse
L-M02	qwen3-coder-30b-a3b	lenient	SGB	0	570	4.84	[4.50, 5.17]	+2.66	graded
L-M03	qwen3-coder-next	strict	SGB	0	570	11.70	[11.35, 12.04]	−11.70	collapse

Table 24 continued

Run	Model	Persona	Prompt	t	n	MAE	95% CI	Bias	Behaviour
L-M04	qwen3-coder-next	lenient	SGB	0	570	2.81	[2.64, 2.98]	−0.41	graded
L-M05	qwen2.5-coder-7b	strict	SGB	0	570	7.92	[7.54, 8.30]	−7.41	graded
L-M06	qwen2.5-coder-7b	lenient	SGB	0	570	3.88	[3.65, 4.11]	−0.86	graded
L-M07	qwen2.5-coder-14b	strict	SGB	0	570	24.52	[23.94, 25.13]	−24.52	collapse
L-MP01	qwen2.5-coder-32b	mpstrict	SGB	0	570	16.41	[16.01, 16.80]	−16.40	collapse
L-MP02	qwen2.5-coder-32b	mprigorous	SGB	0	570	17.74	[17.29, 18.16]	−17.73	collapse
L-MP03	qwen2.5-coder-32b	mpfair	SGB	0	570	18.55	[18.11, 19.00]	−18.54	collapse
L-MP04	qwen2.5-coder-32b	mpnoclause	SGB	0	570	15.55	[15.16, 15.95]	−15.55	collapse
L-MP05	qwen2.5-coder-32b	mpnharsh	SGB	0	570	8.25	[7.93, 8.55]	−8.13	collapse
L-MP06	mistral-small-24b	mpstrict	SGB	0	570	25.99	[25.42, 26.57]	−25.99	near-refusal
L-MP07	mistral-small-24b	mprigorous	SGB	0	570	25.87	[25.29, 26.45]	−25.87	near-refusal
L-MP08	mistral-small-24b	mpfair	SGB	0	570	25.05	[24.48, 25.64]	−25.05	near-refusal
L-MP09	mistral-small-24b	mpnoclause	SGB	0	570	10.40	[10.06, 10.73]	−10.34	collapse
L-MP10	mistral-small-24b	mpnharsh	SGB	0	570	4.23	[4.02, 4.46]	−3.31	graded
L-MP11	llama-3.3-70b	mpstrict	SGB	0	570	12.67	[12.28, 13.07]	−12.66	collapse
L-MP12	llama-3.3-70b	mprigorous	SGB	0	570	11.37	[11.00, 11.74]	−11.35	collapse
L-MP13	llama-3.3-70b	mpfair	SGB	0	570	11.41	[11.04, 11.78]	−11.39	collapse
L-MP14	llama-3.3-70b	mpnoclause	SGB	0	570	8.14	[7.83, 8.43]	−8.03	collapse
L-MP15	llama-3.3-70b	mpnharsh	SGB	0	570	4.60	[4.38, 4.81]	−3.71	graded
L-MP16	glm-4-32b	mpstrict	SGB	0	570	7.02	[6.58, 7.49]	−6.61	graded
L-MP17	glm-4-32b	mprigorous	SGB	0	570	7.43	[7.00, 7.86]	−7.12	graded
L-MP18	glm-4-32b	mpfair	SGB	0	570	6.41	[6.00, 6.82]	−5.85	graded
L-MP19	glm-4-32b	mpnoclause	SGB	0	570	5.51	[5.22, 5.83]	−5.10	graded
L-MP20	glm-4-32b	mpnharsh	SGB	0	570	2.88	[2.71, 3.08]	−0.49	graded
L-MP21	gemma-3-27b	mpstrict	SGB	0	570	10.49	[10.12, 10.85]	−10.42	collapse
L-MP22	gemma-3-27b	mprigorous	SGB	0	570	9.23	[8.87, 9.57]	−9.09	collapse
L-MP23	gemma-3-27b	mpfair	SGB	0	570	8.60	[8.27, 8.93]	−8.42	collapse
L-MP24	gemma-3-27b	mpnoclause	SGB	0	570	9.90	[9.52, 10.26]	−9.73	collapse
L-MP25	gemma-3-27b	mpnharsh	SGB	0	570	4.56	[4.34, 4.79]	−2.36	graded
L-N01	qwen2.5-72b	neutral	SGB	0	570	3.14	[2.95, 3.35]	+0.19	graded
L-N02	qwen2.5-72b	strict	SGB	0	570	9.82	[9.46, 10.18]	−9.78	collapse
L-N03	qwen2.5-72b	lenient	SGB	0	570	4.47	[4.17, 4.81]	+3.69	graded
L-N04	qwen2.5-72b	rigorous	SGB	0	570	3.09	[2.91, 3.28]	−0.29	graded
L-N05	qwen2.5-72b	exacting	SGB	0	570	3.43	[3.25, 3.62]	−1.57	graded
L-P01	qwen2.5-coder-32b	rigorous	SGB	0	570	8.70	[8.38, 9.01]	−8.61	collapse
L-P02	qwen2.5-coder-32b	exacting	SGB	0	570	10.23	[9.91, 10.56]	−10.20	collapse
L-PS01	qwen2.5-coder-32b	mpframe	SGB	0	570	9.37	[9.04, 9.68]	−9.31	collapse
L-PS02	qwen2.5-coder-32b	mps2only	SGB	0	570	12.92	[12.56, 13.27]	−12.91	collapse
L-PS03	qwen2.5-coder-32b	mpnoclause	SGB	0	570	15.55	[15.16, 15.95]	−15.55	collapse
L-PS04	qwen2.5-coder-32b	mpharsh	SGB	0	570	20.26	[19.79, 20.73]	−20.26	collapse
L-PS05	mistral-small-24b	mpframe	SGB	0	570	5.62	[5.36, 5.89]	−5.33	graded
L-PS06	mistral-small-24b	mps2only	SGB	0	570	26.04	[25.45, 26.61]	−26.04	refusal
L-PS07	mistral-small-24b	mpnoclause	SGB	0	570	10.33	[9.99, 10.66]	−10.28	collapse
L-PS08	mistral-small-24b	mpharsh	SGB	0	570	26.03	[25.44, 26.61]	−26.03	refusal
L-PS09	llama-3.1-8b	mpframe	SGB	0	570	8.46	[8.10, 8.86]	−6.94	collapse
L-PS10	llama-3.1-8b	mps2only	SGB	0	570	26.04	[25.45, 26.61]	−26.04	refusal
L-PS11	llama-3.1-8b	mpnoclause	SGB	0	570	23.45	[22.89, 23.99]	−23.45	collapse
L-PS12	llama-3.1-8b	mpharsh	SGB	0	570	26.04	[25.45, 26.61]	−26.04	refusal
L-Q01	qwen3-235b-a22b	neutral	SGB	0	570	4.09	[3.87, 4.32]	−3.56	graded
L-Q02	qwen3-235b-a22b	strict	SGB	0	570	7.32	[7.01, 7.61]	−7.24	graded
L-Q03	qwen3-235b-a22b	lenient	SGB	0	570	6.01	[5.65, 6.37]	+5.94	graded
L-Q04	qwen3-235b-a22b	rigorous	SGB	0	570	5.18	[4.92, 5.45]	−4.90	graded
L-Q05	qwen3-235b-a22b	exacting	SGB	0	570	6.85	[6.56, 7.13]	−6.75	graded
L-R01	qwen3-coder-480b	neutral	SGB	0	570	3.14	[2.95, 3.35]	+0.51	graded
L-R02	qwen3-coder-480b	strict	SGB	0	570	3.78	[3.58, 3.98]	−2.52	graded
L-R03	qwen3-coder-480b	lenient	SGB	0	570	4.13	[3.83, 4.44]	+3.38	graded
L-R04	qwen3-coder-480b	rigorous	SGB	0	570	3.06	[2.88, 3.26]	−0.23	graded
L-R05	qwen3-coder-480b	exacting	SGB	0	570	3.33	[3.14, 3.52]	−1.56	graded
L-S01	deepseek-coder-v2-lite	neutral	SGB	0	570	5.98	[5.69, 6.29]	−3.44	graded
L-S02	deepseek-coder-v2-lite	strict	SGB	0	570	12.09	[11.66, 12.52]	−11.87	collapse
L-S03	deepseek-coder-v2-lite	lenient	SGB	0	570	5.92	[5.62, 6.22]	−3.31	graded
L-S04	deepseek-coder-v2-lite	rigorous	SGB	0	570	7.01	[6.70, 7.34]	−5.31	graded
L-S05	deepseek-coder-v2-lite	exacting	SGB	0	570	6.69	[6.39, 7.01]	−4.83	graded
L-U01	llama-3.1-8b	neutral	SGB	0	570	7.31	[6.97, 7.68]	−5.15	graded
L-U02	llama-3.1-8b	strict	SGB	0	570	26.04	[25.45, 26.61]	−26.04	refusal
L-U03	llama-3.1-8b	lenient	SGB	0	570	6.49	[6.06, 6.96]	+5.36	graded
L-U04	llama-3.1-8b	rigorous	SGB	0	570	10.31	[9.88, 10.74]	−9.43	collapse
L-U05	llama-3.1-8b	exacting	SGB	0	570	6.89	[6.57, 7.23]	−4.50	graded
L-V01	llama-3.3-70b	neutral	SGB	0	570	4.52	[4.31, 4.73]	−3.60	graded
L-V02	llama-3.3-70b	strict	SGB	0	570	13.52	[13.11, 13.93]	−13.52	collapse
L-V03	llama-3.3-70b	lenient	SGB	0	570	3.99	[3.72, 4.28]	+1.82	graded
L-V04	llama-3.3-70b	rigorous	SGB	0	570	4.69	[4.47, 4.91]	−3.89	graded
L-V05	llama-3.3-70b	exacting	SGB	0	570	6.74	[6.46, 7.01]	−6.56	graded
L-W01	glm-4-9b	neutral	SGB	0	570	4.31	[4.02, 4.63]	+1.68	graded
L-W02	glm-4-9b	strict	SGB	0	570	25.48	[24.88, 26.04]	−25.48	near-refusal
L-W03	glm-4-9b	lenient	SGB	0	570	6.94	[6.43, 7.46]	+6.40	graded
L-W04	glm-4-9b	rigorous	SGB	0	570	4.53	[4.21, 4.87]	+1.70	graded
L-W05	glm-4-9b	exacting	SGB	0	570	5.18	[4.80, 5.57]	+3.80	graded

Table 24 continued

Run	Model	Persona	Prompt	t	n	MAE	95% CI	Bias	Behaviour
L-X01	glm-4-32b	neutral	SGB	0	570	2.85	[2.66, 3.05]	+0.50	graded
L-X02	glm-4-32b	strict	SGB	0	570	9.66	[9.15, 10.19]	−9.43	collapse
L-X03	glm-4-32b	lenient	SGB	0	570	5.33	[4.99, 5.69]	+5.00	graded
L-X04	glm-4-32b	rigorous	SGB	0	570	2.74	[2.56, 2.93]	+0.03	graded
L-X05	glm-4-32b	exacting	SGB	0	570	3.72	[3.49, 3.96]	−2.34	graded
L-Y01	gemma-3-27b	neutral	SGB	0	570	4.32	[4.12, 4.55]	−1.52	graded
L-Y02	gemma-3-27b	strict	SGB	0	570	11.55	[11.16, 11.92]	−11.50	collapse
L-Y03	gemma-3-27b	lenient	SGB	0	570	4.55	[4.25, 4.86]	+2.29	graded
L-Y04	gemma-3-27b	rigorous	SGB	0	570	4.88	[4.65, 5.13]	−3.49	graded
L-Y05	gemma-3-27b	exacting	SGB	0	570	5.47	[5.21, 5.73]	−4.63	graded
L-Z01	mistral-small-24b	neutral	SGB	0	570	3.66	[3.46, 3.87]	−2.41	graded
L-Z02	mistral-small-24b	strict	SGB	0	570	26.03	[25.44, 26.61]	−26.03	refusal
L-Z03	mistral-small-24b	lenient	SGB	0	570	4.76	[4.44, 5.12]	+3.57	graded
L-Z04	mistral-small-24b	rigorous	SGB	0	570	6.15	[5.87, 6.41]	−5.89	graded
L-Z05	mistral-small-24b	exacting	SGB	0	569	7.04	[6.75, 7.33]	−6.93	graded
Prompt-only baselines (Appendix B; not grader configurations)									
PB-A01	qwen2.5-coder-32b	strict (arbitrated)	SGB	0	100	9.29	[8.58, 9.98]	−9.27	collapse
PB-A02	glm-4-32b	strict (arbitrated)	SGB	0	100	2.64	[2.23, 3.10]	−1.53	graded
PB-A03	mistral-small-24b	strict (arbitrated)	SGB	0	100	3.13	[2.74, 3.53]	−2.31	graded
PB-A04	llama-3.1-8b	strict (arbitrated)	SGB	0	99	7.46	[6.60, 8.28]	−5.77	graded
PB-D01	qwen2.5-coder-32b	strict (decomposed)	SGB	0	100	20.00	[19.10, 20.93]	−20.00	collapse
PB-D02	glm-4-32b	strict (decomposed)	SGB	0	100	5.01	[4.50, 5.51]	−4.95	graded
PB-D03	mistral-small-24b	strict (decomposed)	SGB	0	100	26.27	[24.98, 27.55]	−26.27	collapse
PB-D04	llama-3.1-8b	strict (decomposed)	SGB	0	100	27.26	[25.89, 28.65]	−27.26	refusal

J Full Limitations Inventory

Persona coverage. Qwen2.5-Coder-14B received no persona in the full-cohort grid and Qwen3-Coder-Next only *strict*; the held-out sweep adds all three harsh wordings and *lenient* on the 7B, 14B and 30B (base and pooled), leaving the 80B evaluated on only one wording. On the ML exam the 14B received *neutral* and *strict* only, and Qwen3-Coder-Next no *rigorous* or *exacting*. The headword series covers five models and the policy 2×2 three, with only Qwen2.5-Coder-32B and Mistral-Small-24B in both, and we did not cross them, so an adjective-only interaction is untested. On the closed side only Gemini received all five wordings; Section 10 itemises the cross-vendor coverage, whose decoding differs where the APIs force it (Anthropic's current models expose no temperature; `gpt-5.5` is not bit-reproducible at temperature 0).

Model and serving confounds. `gemini-2.5-pro` rejects a zero thinking budget, so D03 against D01/D02 compares an older thinking-on model with newer thinking-off ones, if anything understating the newer models' advantage. The Qwen3-Coder *Instruct* variants implement no thinking mode, so their thinking-on configurations are inert replicates.

Fine-tuning. Held-out MAE CIs are roughly ± 0.45, so the paired test carries the verdicts. The bd arm trains on targets produced by D01 (CV exam) and IG07 (ML exam), so it measures closed-model distillation; only the marks recipe trains on human labels alone. The 30B is not a clean size point (MoE, attention-only LoRA); base rows use the fine-tuning harness, not identical reruns of Section 5. vLLM batching reorders reductions (~ 0.01 MAE); all evaluation, Gemma-4's adapter included, is vLLM-served, with Gemma-4's path cross-validated against HuggingFace generation to within ~ 0.1 MAE (Appendix E).

K ML-Exam Run Table

Table 25 lists every one of the 162 ML-exam configurations individually — 30 closed-model and 132 open-weights runs — with each run's own confidence interval, bias, and behaviour class. Table 22 pairs the matched neutral/strict runs for the headline comparison; this one is the flat record, and is the ML exam's counterpart to Table 24. The machine-readable version, with per-question breakdowns and the full column set, ships as `analysis/introduction_to_ai_master_comparison.csv` in the supplementary material.

Table 25: Every second-exam grader configuration, one row per run: the 30 closed-model configurations (23 Gemini IG-, 6 OpenAI IO- and 1 Anthropic IN-; Section 6.1) followed by the 132 open-weights (IA-) configurations, each in run-id order. *Prompt* lists the components present (S reference solution, B rubric breakdown, R thinking enabled, F few-shot demonstrations); this exam has no student-facing guidelines document, so the G component of Table 24 never appears. Personas prefixed `mp` are the mechanism probes of Section 5.2, replayed on this exam (keys decoded in Table 24's caption). n is the number of students scored against the grader average ($*$: IG18's 693 are the survivors of prompt-length failures, an easier subsample whose MAE is not comparable to full-cohort rows; Appendix H); MAE and bias are on the ≈ 65-point scale (score plus bonus, matching how its graders record marks), the 95% bootstrap CI (50,000 resamples) on MAE. Behaviour classes are those of Section 5.1, at this exam's floor-matched collapse band (MAE ≥ 15.7), and describe a run's error level rather than persona damage: DeepSeek-Coder-V2-Lite over-marks badly enough to leave it unprompted (IA11, IA13). A refusal's MAE is a ceiling artifact, not a severity measurement. IG15–IG20 ran on the first 100 students; IG10–IG12, IA92 and IA136 are repeated-sampling probes whose scores are the per-student mean across five reruns (Appendix H). `claude-opus-5` exposes no sampling temperature (thinking disabled; shown as —). Run ids match the released tracker and spreadsheets. Generated by `analysis/introduction_to_ai_make_appendix_table.py`.

Run	Model	Persona	Prompt	t	n	MAE	95% CI	Bias	Behaviour
Closed-model configurations (Gemini IG-, OpenAI IO-, Anthropic IN-)									
IG01	`flash-lite`	neutral	SB	0	1013	7.53	[7.20, 7.87]	+5.98	graded
IG02	`flash-lite`	neutral	B	0	1032	12.46	[12.03, 12.89]	+12.16	graded
IG03	`flash-lite`	neutral	SBR	0	1037	5.37	[5.09, 5.66]	−4.04	graded
IG04	`flash-lite`	neutral	S	0	1016	6.88	[6.57, 7.20]	+5.27	graded
IG05	`flash-lite`	strict	SB	0	1008	9.02	[8.61, 9.43]	−7.83	graded
IG06	`flash-lite`	lenient	SB	0	1020	17.86	[17.30, 18.42]	+17.80	collapse
IG07	`3-flash-preview`	neutral	SB	0	1038	3.92	[3.68, 4.16]	+2.24	graded
IG08	`3.1-pro-preview`	neutral	SB	0	1038	3.40	[3.16, 3.65]	+1.18	graded
IG09	`2.5-pro`	neutral	SB	0	1035	3.47	[3.24, 3.72]	−0.83	graded
IG10	`flash-lite`	neutral	SB	0	955	7.78	[7.44, 8.13]	+6.14	graded
IG11	`flash-lite`	neutral	SB	0.5	963	7.65	[7.32, 8.00]	+5.95	graded
IG12	`flash-lite`	neutral	SB	0.7	963	7.63	[7.29, 7.98]	+5.85	graded
IG13	`flash-lite`	strict	SBR	0	1038	7.76	[7.43, 8.09]	−7.06	graded
IG14	`3.1-pro-preview`	lenient	SB	0	1033	4.48	[4.22, 4.75]	+3.46	graded
IG15	`3-flash-preview`	neutral	B	0	100	4.35	[3.32, 5.64]	+3.51	graded
IG16	`3-flash-preview`	neutral	SBR	0	100	3.76	[2.78, 5.01]	+2.78	graded
IG17	`3-flash-preview`	neutral	S	0	96	3.37	[2.42, 4.70]	+2.38	graded
IG18	`flash-lite`	neutral	SBF	0	693*	4.30	[4.01, 4.60]	+0.36	graded
IG19	`3-flash-preview`	strict	SB	0	100	2.87	[1.95, 4.10]	+0.09	graded
IG20	`3-flash-preview`	lenient	SB	0	100	6.99	[5.76, 8.43]	+6.60	graded
IG21	`flash-lite`	rigorous	SB	0	1012	5.52	[5.25, 5.79]	+2.79	graded
IG22	`flash-lite`	exacting	SB	0	1021	5.96	[5.68, 6.25]	+3.18	graded
IG23	`3.1-pro-preview`	strict	SB	0	1038	3.67	[3.43, 3.94]	−1.00	graded
IN01	`claude-opus-5`	neutral	SB	—	1038	4.37	[4.13, 4.63]	−2.64	graded
IO01	`gpt-5.5`	neutral	SB	0	1038	3.54	[3.33, 3.77]	+1.23	graded
IO02	`gpt-5.5`	strict	SB	0	1038	3.70	[3.48, 3.93]	−1.38	graded
IO03	`gpt-5.5`	lenient	SB	0	1038	6.29	[6.02, 6.57]	+5.83	graded
IO11	`gpt-5.4`	neutral	SB	0	1038	4.30	[4.07, 4.55]	−1.95	graded
IO12	`gpt-5.4`	strict	SB	0	1038	5.11	[4.85, 5.37]	−3.61	graded
IO13	`gpt-5.4`	lenient	SB	0	1038	8.30	[7.96, 8.64]	+7.94	graded
Open-weights (IA-) configurations									
IA01-G32-n	`glm4-32b`	neutral	SB	0	1038	8.03	[7.69, 8.37]	+7.29	graded
IA01-G9-n	`glm4-9b`	neutral	SB	0	1038	13.15	[12.60, 13.71]	+12.10	graded
IA01-GE27-n	`gemma3-27b`	neutral	SB	0	1038	8.33	[8.00, 8.67]	+6.83	graded
IA01-L8-n	`llama31-8b`	neutral	SB	0	1037	13.16	[12.61, 13.71]	+10.27	graded
IA01-M24-n	`mistral-small-24b`	neutral	SB	0	1038	8.96	[8.59, 9.34]	+8.16	graded
IA01-Q14-n	`qwen25-coder-14b`	neutral	SB	0	1038	8.61	[8.25, 8.98]	+7.40	graded
IA01-Q30-n	`qwen3-coder-30b-a3b`	neutral	SB	0	1038	11.12	[10.64, 11.59]	+10.44	graded
IA01-Q32-n	`qwen25-coder-32b`	neutral	SB	0	1038	4.57	[4.33, 4.83]	−0.21	graded
IA01-Q7-n	`qwen25-coder-7b`	neutral	SB	0	1038	6.88	[6.56, 7.20]	+3.77	graded
IA02-G32-s	`glm4-32b`	strict	SB	0	1038	12.18	[11.68, 12.69]	−10.68	graded
IA02-G9-s	`glm4-9b`	strict	SB	0	1038	16.68	[16.15, 17.20]	−16.52	collapse
IA02-GE27-s	`gemma3-27b`	strict	SB	0	1038	7.61	[7.23, 7.99]	−6.25	graded
IA02-L8-s	`llama31-8b`	strict	SB	0	1038	29.58	[28.70, 30.45]	−29.58	refusal
IA02-M24-s	`mistral-small-24b`	strict	SB	0	1038	17.32	[16.77, 17.87]	−17.20	collapse
IA02-Q14-s	`qwen25-coder-14b`	strict	SB	0	1038	11.06	[10.61, 11.50]	−10.46	graded
IA02-Q30-s	`qwen3-coder-30b-a3b`	strict	SB	0	1038	5.58	[5.31, 5.87]	−1.55	graded
IA02-Q32-s	`qwen25-coder-32b`	strict	SB	0	1038	13.24	[12.79, 13.69]	−13.07	graded
IA02-Q7-s	`qwen25-coder-7b`	strict	SB	0	1038	7.45	[7.11, 7.80]	−5.67	graded
IA11	`deepseek-coder-v2-lite`	neutral	S	0	1038	16.48	[15.82, 17.15]	+14.96	collapse
IA12	`deepseek-coder-v2-lite`	strict	S	0	1038	9.12	[8.70, 9.54]	−5.19	graded
IA13	`deepseek-coder-v2-lite`	neutral	SB	0	1038	16.41	[15.76, 17.07]	+14.89	collapse
IA14	`deepseek-coder-v2-lite`	strict	SB	0	1038	9.22	[8.83, 9.62]	+2.42	graded
IA15	`deepseek-coder-v2-lite`	lenient	SB	0	1038	16.29	[15.64, 16.95]	+14.96	collapse
IA16	`deepseek-coder-v2-lite`	rigorous	SB	0	1038	14.23	[13.64, 14.83]	+12.29	graded
IA17	`deepseek-coder-v2-lite`	exacting	SB	0	1038	15.50	[14.87, 16.13]	+13.90	graded
IA18	`gemma3-12b`	neutral	SB	0	1038	10.36	[9.96, 10.76]	+9.13	graded
IA19	`gemma3-12b`	strict	SB	0	1038	6.79	[6.49, 7.10]	+0.09	graded
IA20	`gemma3-12b`	lenient	SB	0	1038	13.70	[13.22, 14.18]	+13.17	graded
IA21	`gemma3-12b`	rigorous	SB	0	1038	8.51	[8.16, 8.86]	+6.02	graded
IA22	`gemma3-12b`	exacting	SB	0	1038	8.71	[8.36, 9.06]	+6.48	graded
IA23	`gemma3-27b`	neutral	S	0	1038	8.09	[7.77, 8.42]	+6.55	graded
IA24	`gemma3-27b`	strict	S	0	1038	7.35	[6.98, 7.72]	−5.88	graded
IA25	`gemma3-27b`	mpstrict	SB	0	1038	6.77	[6.43, 7.12]	−4.90	graded

Table 25 continued

Run	Model	Persona	Prompt	t	n	MAE	95% CI	Bias	Behaviour
IA26	gemma3-27b	mprigorous	SB	0	1038	6.31	[5.98, 6.64]	−4.05	graded
IA27	gemma3-27b	mpfair	SB	0	1038	5.93	[5.62, 6.25]	−3.35	graded
IA28	gemma3-27b	mpnoclause	SB	0	1038	6.48	[6.16, 6.79]	−2.35	graded
IA29	gemma3-27b	mpnharsh	SB	0	1038	7.78	[7.47, 8.10]	+5.74	graded
IA30	gemma3-27b	lenient	SB	0	1038	18.54	[18.02, 19.06]	+18.47	collapse
IA31	gemma3-27b	rigorous	SB	0	1038	6.18	[5.91, 6.46]	+3.18	graded
IA32	gemma3-27b	exacting	SB	0	1038	5.95	[5.69, 6.23]	+1.83	graded
IA33	glm4-32b	neutral	S	0	1038	7.51	[7.19, 7.84]	+6.75	graded
IA34	glm4-32b	strict	S	0	1038	18.44	[17.86, 19.03]	−17.63	collapse
IA35	glm4-32b	mpstrict	SB	0	1038	7.29	[6.92, 7.67]	−4.76	graded
IA36	glm4-32b	mprigorous	SB	0	1038	7.11	[6.74, 7.50]	−4.74	graded
IA37	glm4-32b	mpfair	SB	0	1038	6.74	[6.38, 7.11]	−3.85	graded
IA38	glm4-32b	mpnoclause	SB	0	1038	5.60	[5.29, 5.92]	−3.40	graded
IA39	glm4-32b	mpnharsh	SB	0	1038	6.83	[6.52, 7.15]	+5.82	graded
IA40	glm4-32b	lenient	SB	0	1038	15.35	[14.90, 15.80]	+15.23	graded
IA41	glm4-32b	rigorous	SB	0	1038	7.47	[7.15, 7.80]	+6.64	graded
IA42	glm4-32b	exacting	SB	0	1038	5.57	[5.28, 5.86]	+2.97	graded
IA43	glm4-9b	lenient	SB	0	1038	25.98	[25.22, 26.74]	+25.96	collapse
IA44	glm4-9b	rigorous	SB	0	1038	11.80	[11.29, 12.31]	+10.49	graded
IA45	glm4-9b	exacting	SB	0	1038	14.71	[14.11, 15.33]	+14.05	graded
IA46	llama31-8b	mpframe	SB	0	1038	11.54	[11.07, 12.02]	+7.78	graded
IA47	llama31-8b	mps2only	SB	0	1038	29.39	[28.52, 30.25]	−29.37	near-refusal
IA48	llama31-8b	mpnoclause	SB	0	1038	25.87	[25.08, 26.65]	−25.84	collapse
IA49	llama31-8b	mpharsh	SB	0	1038	29.58	[28.70, 30.45]	−29.58	refusal
IA50	llama31-8b	lenient	SB	0	1037	27.73	[26.92, 28.54]	+27.69	collapse
IA51	llama31-8b	rigorous	SB	0	1038	9.01	[8.61, 9.42]	+3.07	graded
IA52	llama31-8b	exacting	SB	0	1038	11.96	[11.45, 12.48]	+9.15	graded
IA53	mistral-small-24b	mpstrict	SB	0	1038	12.49	[12.03, 12.96]	−12.14	graded
IA54	mistral-small-24b	mprigorous	SB	0	1038	8.81	[8.43, 9.21]	−7.67	graded
IA55	mistral-small-24b	mpfair	SB	0	1038	8.63	[8.24, 9.02]	−7.41	graded
IA56	mistral-small-24b	mpnoclause	SB	0	1038	6.37	[6.08, 6.67]	−0.67	graded
IA57	mistral-small-24b	mpnharsh	SB	0	1038	8.09	[7.74, 8.45]	+6.94	graded
IA58	mistral-small-24b	mpframe	SB	0	1038	6.58	[6.27, 6.90]	+4.10	graded
IA59	mistral-small-24b	mps2only	SB	0	1038	26.03	[25.29, 26.77]	−26.02	collapse
IA60	mistral-small-24b	mpharsh	SB	0	1037	17.29	[16.73, 17.84]	−17.17	collapse
IA61	mistral-small-24b	lenient	SB	0	1038	15.56	[15.08, 16.05]	+15.44	graded
IA62	mistral-small-24b	rigorous	SB	0	1038	5.89	[5.61, 6.19]	+2.96	graded
IA63	mistral-small-24b	exacting	SB	0	1038	5.70	[5.43, 5.98]	+2.41	graded
IA64	qwen25-coder-32b	neutral	S	0	1038	4.45	[4.21, 4.71]	+0.13	graded
IA65	qwen25-coder-32b	strict	S	0	1038	13.43	[12.97, 13.89]	−13.25	graded
IA66	qwen25-coder-32b	lenient	SB	0	1038	6.96	[6.66, 7.26]	+5.56	graded
IA67	qwen25-coder-32b	mpstrict	SB	0	1038	9.65	[9.26, 10.04]	−9.30	graded
IA68	qwen25-coder-32b	mprigorous	SB	0	1038	10.19	[9.80, 10.59]	−9.87	graded
IA69	qwen25-coder-32b	mpfair	SB	0	1038	10.40	[10.00, 10.81]	−10.09	graded
IA70	qwen25-coder-32b	mpnoclause	SB	0	1038	9.45	[9.05, 9.85]	−9.09	graded
IA71	qwen25-coder-32b	mpnharsh	SB	0	1038	4.67	[4.42, 4.93]	−0.64	graded
IA72	qwen25-coder-32b	rigorous	SB	0	1038	4.73	[4.47, 5.00]	−1.47	graded
IA73	qwen25-coder-32b	exacting	SB	0	1038	5.43	[5.14, 5.73]	−3.73	graded
IA74	qwen25-coder-32b	mpframe	SB	0	1038	4.78	[4.52, 5.05]	−1.71	graded
IA75	qwen25-coder-32b	mps2only	SB	0	1038	6.68	[6.35, 7.02]	−5.67	graded
IA76	qwen25-coder-32b	mpharsh	SB	0	1038	13.24	[12.80, 13.70]	−13.07	graded
IA77	qwen25-coder-7b	neutral	S	0	1038	7.55	[7.20, 7.89]	+4.92	graded
IA78	qwen25-coder-7b	lenient	SB	0	1038	8.14	[7.79, 8.50]	+6.50	graded
IA79	qwen3-coder-30b-a3b	neutral	S	0	1038	9.13	[8.73, 9.54]	+7.80	graded
IA80	qwen3-coder-30b-a3b	lenient	SB	0	1038	17.16	[16.58, 17.74]	+17.02	collapse
IA91	qwen25-coder-32b	neutral	B	0	1033	4.73	[4.49, 4.98]	+1.20	graded
IA92	qwen25-coder-32b	neutral	SB	0.5	41	4.65	[2.78, 7.37]	+4.51	graded
IA93	qwen25-coder-32b	neutral	SBR	0	1033	4.59	[4.34, 4.85]	−0.24	graded
IA94	qwen25-coder-7b	neutral	B	0	1038	7.51	[7.16, 7.88]	+5.57	graded
IA95	qwen3-coder-30b-a3b	neutral	B	0	1013	11.96	[11.50, 12.40]	+11.53	graded
IA96	qwen3-coder-30b-a3b	neutral	SBR	0	1017	11.09	[10.61, 11.57]	+10.40	graded
IA101	glm45-air	neutral	SB	0	1037	5.60	[5.31, 5.89]	−1.34	graded
IA102	glm45-air	strict	SB	0	1038	18.49	[17.96, 19.02]	−18.31	collapse
IA103	glm45-air	lenient	SB	0	1038	15.30	[14.74, 15.87]	+15.17	graded
IA104	glm45-air	rigorous	SB	0	1037	6.26	[5.95, 6.58]	−3.97	graded
IA105	glm45-air	exacting	SB	0	1022	9.95	[9.55, 10.36]	−9.15	graded
IA106	llama33-70b	neutral	S	0	1016	4.29	[4.05, 4.55]	+0.01	graded
IA107	llama33-70b	strict	S	0	1013	14.13	[13.62, 14.64]	−13.64	graded
IA108	llama33-70b	mpstrict	SB	0	1038	9.56	[9.16, 9.96]	−9.05	graded
IA109	llama33-70b	mprigorous	SB	0	1038	8.55	[8.18, 8.94]	−7.86	graded
IA110	llama33-70b	mpfair	SB	0	1038	8.13	[7.76, 8.50]	−7.36	graded
IA111	llama33-70b	mpnoclause	SB	0	1038	6.12	[5.82, 6.43]	−4.87	graded
IA112	llama33-70b	mpnharsh	SB	0	1038	4.60	[4.36, 4.85]	+1.61	graded
IA113	llama33-70b	neutral	SB	0	1038	4.70	[4.45, 4.96]	+1.90	graded
IA114	llama33-70b	strict	SB	0	1038	10.17	[9.77, 10.58]	−9.63	graded
IA115	llama33-70b	lenient	SB	0	1038	11.74	[11.34, 12.15]	+11.55	graded
IA116	llama33-70b	rigorous	SB	0	1038	4.60	[4.36, 4.85]	+1.71	graded
IA117	llama33-70b	exacting	SB	0	1038	5.23	[4.95, 5.51]	−3.31	graded

Table 25 continued

Run	Model	Persona	Prompt	t	n	MAE	95% CI	Bias	Behaviour
IA118	qwen25-72b	neutral	SB	0	1038	10.64	[10.25, 11.03]	+10.31	graded
IA119	qwen25-72b	strict	SB	0	1038	5.42	[5.15, 5.70]	−1.67	graded
IA120	qwen25-72b	lenient	SB	0	1038	16.93	[16.44, 17.42]	+16.89	collapse
IA121	qwen25-72b	rigorous	SB	0	1038	8.50	[8.16, 8.85]	+7.94	graded
IA122	qwen25-72b	exacting	SB	0	1038	8.45	[8.09, 8.81]	+7.80	graded
IA123	qwen3-235b-a22b	neutral	SB	0	1037	5.40	[5.13, 5.67]	+3.73	graded
IA124	qwen3-235b-a22b	strict	SB	0	1038	5.18	[4.89, 5.47]	−3.39	graded
IA125	qwen3-235b-a22b	lenient	SB	0	1038	17.46	[16.99, 17.93]	+17.41	collapse
IA126	qwen3-235b-a22b	rigorous	SB	0	1038	4.53	[4.29, 4.78]	+2.11	graded
IA127	qwen3-235b-a22b	exacting	SB	0	1038	4.58	[4.32, 4.84]	−1.58	graded
IA128	qwen3-coder-480b	neutral	SB	0	1038	10.10	[9.66, 10.54]	+9.52	graded
IA129	qwen3-coder-480b	strict	SB	0	1038	6.04	[5.75, 6.34]	+3.25	graded
IA130	qwen3-coder-480b	lenient	SB	0	1038	15.37	[14.87, 15.88]	+15.29	graded
IA131	qwen3-coder-480b	rigorous	SB	0	1038	7.89	[7.54, 8.25]	+6.99	graded
IA132	qwen3-coder-480b	exacting	SB	0	1038	7.03	[6.69, 7.37]	+5.04	graded
IA133	qwen3-coder-next	neutral	SB	0	1021	4.47	[4.24, 4.72]	+1.51	graded
IA134	qwen3-coder-next	strict	SB	0	1001	8.63	[8.26, 9.00]	−8.12	graded
IA135	qwen3-coder-next	lenient	SB	0	1021	7.64	[7.32, 7.96]	+7.10	graded
IA136	qwen3-coder-next	neutral	SB	0.5	41	5.04	[3.26, 7.76]	+5.04	graded
IA137	qwen3-coder-next	neutral	SBR	0	1021	4.47	[4.24, 4.72]	+1.51	graded
IA201	qwen3-coder-30b-a3b	neutral	SBF	0	1038	5.84	[5.56, 6.13]	+4.56	graded