ACM IUI 2026 · 31st International Conference on Intelligent User Interfaces · Paphos, Cyprus

Smarter Together:

Enhancing Human-AI Collaborative Grading with Teacher-Cognition Multi-Agent LLM Framework (TC-MAG)

Sanskriti Uma · Geniebook, Singapore

Surjya Ghosh · BITS Pilani Goa

Dio Dzaky · Geniebook

THE PROBLEM

Open-ended grading doesn't scale, and students pay for it

Open-ended answers surface far more misconceptions than multiple choice

Partial credit shows where a student's reasoning breaks down

But grading them by hand doesn't keep up

Feedback lags ~a week; cutting latency to a day halves conceptual-error correction time

So classrooms retreat to MCQs

The cost is highest in resource-constrained schools: many students, few teachers

Our goal: automate short-answer, partial-credit grading - without removing teacher oversight

THE GAP

Why can't we simply prompt an LLM to grade?

Reliability

QWK ≈ 0.68

best GPT-4 prompting on standard short-answer benchmarks - below human agreement (0.83-0.92 in K-12 programs)

Calibration

22%

of mis-scores carry high confidence - so selective teacher review breaks down

Transparency

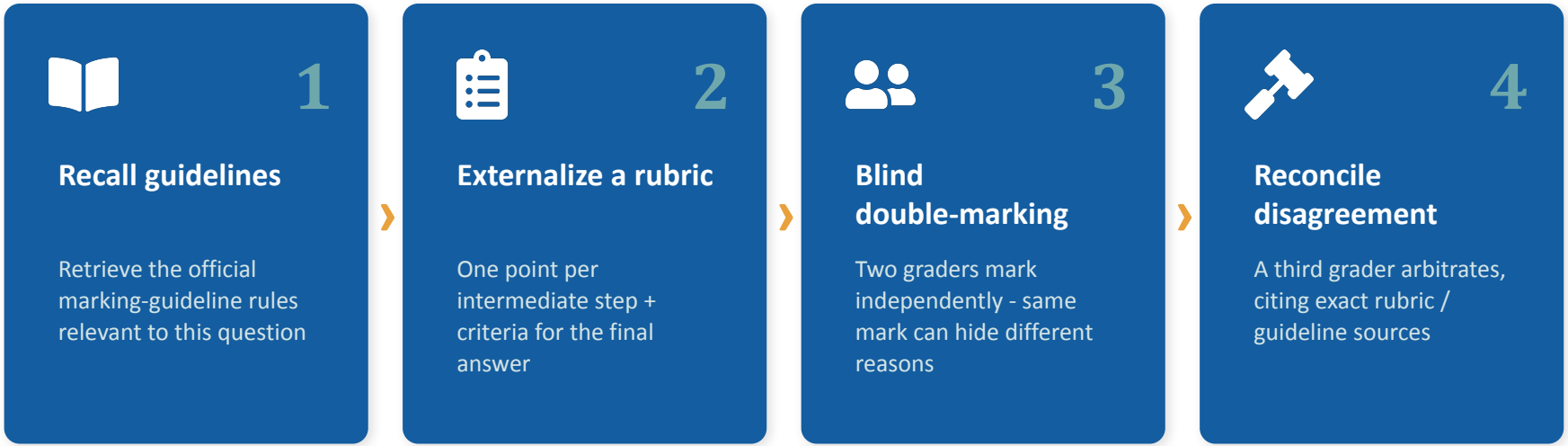
little time saved

post-hoc “explain your score” is inconsistent & pedagogically thin - teachers re-mark anyway

The bar for deployment: automated scores must stay within ≈ 0.10 κ of expert human agreement (Williamson et al. 2012) - and teachers must be willing to delegate.

How do teachers actually mark?

Cognitive Task Analysis - think-aloud marking with 5 Singaporean primary school math teachers (mean 7 yrs experience)



A consistent 4-step behavioral chain - teachers externalize these micro-decisions to keep marks accurate, fair, and defensible. This chain becomes our prompting blueprint.

MOTIVATIONAL STUDY: RELIABILITY

One agent, all four steps? It drifts.

Pilot: single GPT-o3 agent, multi-step prompt, 400 question-answer pairs.

One agent, chain-of-thought prompt (GPT-o3)

91.2% accurate - but the errors are systematic.

- Ignored its own extracted guidelines - **30.4% of errors**
- Drifted from the marking instructions - **32.8%**
- Loose tolerance on answer formatting - **15.7%**

+ 21.1% traced to gaps in the source guideline document itself

Drift = the LLM's output deviating from the instructions in its own prompt.

MOTIVATIONAL STUDY: RELIABILITY

One agent, all four steps? It drifts.

Pilot: single GPT-o3 agent, multi-step prompt, 400 question-answer pairs.

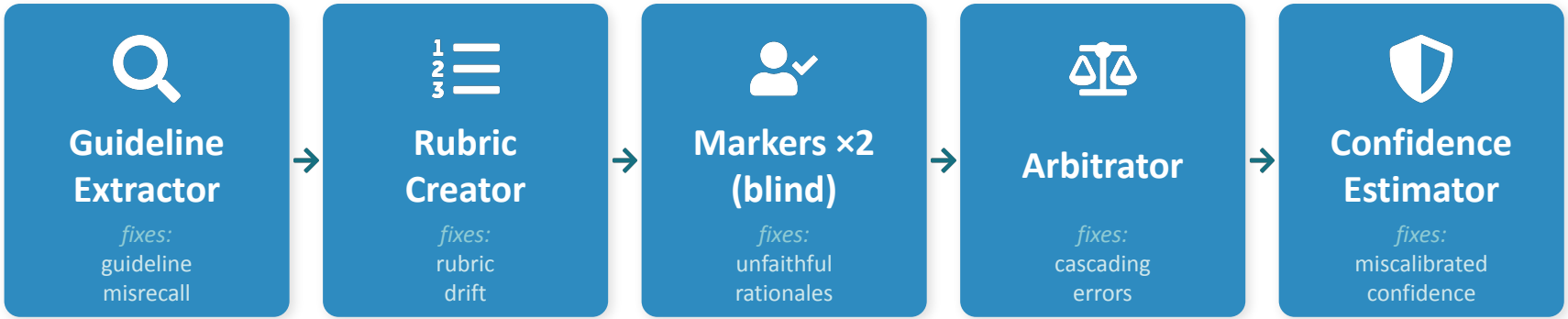
Design takeaway

Root cause: “lost-in-the-middle” - long chained prompts lose earlier instructions.

Fix: one agent per step; each output is committed & logged - a constraint the next agent cannot silently override.

MOTIVATIONAL STUDY: REDEFINING THE ARCHITECTURE

TC-MAG: six agents, one per teacher micro-decision



Any rubric criterion-level disagreement between marker agents, even with equal totals, triggers arbitrator agent.
Every agent emits a short, verifiable micro-explanation.

Why exactly two markers?
1→2 markers: +2.71% · 2→3: +1.21%
beyond 3, accuracy degrades (arbitration overhead)

Every agent earns its place (ablation)
Remove Rubric Creator: −4.08% Δ accuracy
Guideline Extractor: −3.95% · Arbitrator: −2.85%

TWO RESEARCH QUESTIONS

RQ1 · Reliability

Can a teacher-cognition-inspired multi-agent LLM grader surpass the deployment bar on primary-school short-answer math questions - **outperforming human teachers and LLM baselines?**

Deployment bar: automated scores must stay within ≈ 0.10 κ of expert human agreement (Williamson et al. 2012).

2,000 real student answers vs. teacher gold labels

Dataset

- 2,000 Singapore primary-math answers (P1-P6)
- Balanced: 500 each at 1-, 2-, 3-, 4-mark questions strata
- Gold labels: double-blind teacher marking + senior-teacher adjudication
- Arithmetic & word problems in English · 20% include images

Strata balanced to avoid the prevalence-induced "kappa paradox" [23].

RQ1 · RELIABILITY: METHODOLOGY

Same answers, graded by TC-MAG, human teachers, and LLM baselines

Baselines

- **Human teachers** - marking as part of regular duties, unaware of the study (real workload)
- **GPT-4o & GPT-o3** - vanilla and chain-of-thought prompting
- **TC-MAG on GPT-4o** - isolates architecture from model quality

All LLM variants share one I/O contract (question, solution, guidelines → JSON criterion marks) and greedy decoding (temperature = 0).

RQ1 · RELIABILITY: SUCCESS METRICS

Agreement with gold labels, judged against a deployment bar

κ (1-mark)

Cohen’s kappa - chance-corrected agreement on binary right/wrong questions

QWK (2-4-mark)

Quadratic-weighted kappa - larger mark deviations penalized more (partial credit)

Significance

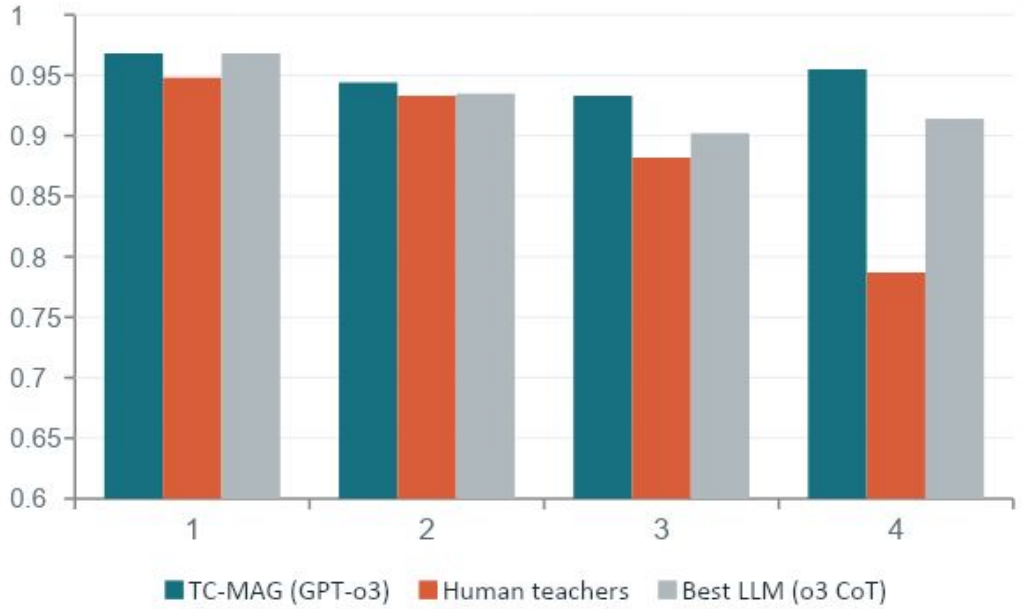
95% CIs, paired bootstrap, Holm-Bonferroni

One-sided tests vs. every baseline

Deployment bar: automated scores must stay within ≈ 0.10 κ of expert human agreement (Williamson et al. 2012).

RQ1 · RELIABILITY: RESULTS

TC-MAG stays reliable where humans slip



Agreement with gold labels (κ / QWK) by question complexity

vs. humans (4-mark) **ΔQWK +0.168**
human reliability collapses to 0.79 as partial-credit complexity rises · $p < .001$

vs. best LLM baseline **ΔQWK +0.012**
and exact accuracy 0.878 vs 0.813 on 2-4-mark · $p < .001$

architecture transfers **+4-7% acc.**
TC-MAG lifts GPT-4o too (exact accuracy) - gains aren't just model quality

Deployment bar: automated scores must stay within ≈ 0.10 κ of expert human agreement (Williamson et al. 2012).

TWO RESEARCH QUESTIONS

RQ2 · Trust & delegation

Do micro-step explanations (anchored on teacher micro-decisions) and calibrated confidence change **teachers' willingness to delegate grading to AI?**

Prior work: explanations raise perceived transparency and trust (Ehsan et al., IUI '19; Poursabzi-Sangdeh et al., CHI '21), and confidence tags support trust calibration (Zhang et al., FAT '20).*

RQ2 · TRUST & DELEGATION: PARTICIPANTS & TASK

Will teachers actually delegate?

Mixed-methods, within-subjects study · N = 14 credentialed teachers (mean 12.1 yrs) · 32 time-pressured marking decisions each

Participants

14 credentialed primary-math teachers · mean 12.1 yrs marking experience (SD 4.6) · generative-AI exposure from rare to weekly

Marking task

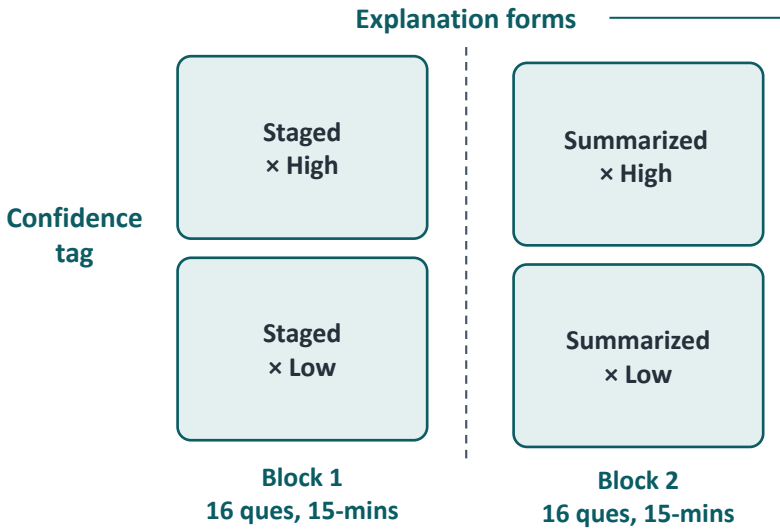
32 short-answer questions per teacher - previously unseen. Balanced across mark stratum, confidence, and AI accuracy. two 15-minute blocks induce real time pressure

Per question, teachers chose: **Mark myself** vs **Use AI** (then accept or override grades). Delegation = accepting the AI mark unchanged.

RQ2 · TRUST & DELEGATION: STUDY DESIGN

Two explanation forms × two confidence tags

2 × 2 within-subjects design



Staged (micro-step) explanation

Extracted guideline + rubric criteria + preferred marker's rationale - mirrors the teacher's marking chain

Summarized explanation

One short paragraph focused on why marks were deducted - a low-effort triage cue

Prior work: explanations raise perceived transparency and trust (Ehsan et al., IUI '19; Poursabzi-Sangdeh et al., CHI '21), and confidence tags support trust calibration (Zhang et al., FAT '20).*

Success metric: does the teacher actually delegate?

Per question, teachers chose: **Mark myself** vs **Use AI** (then accept or override grades). Delegation = accepting the AI mark unchanged.

Delegation rate (primary)

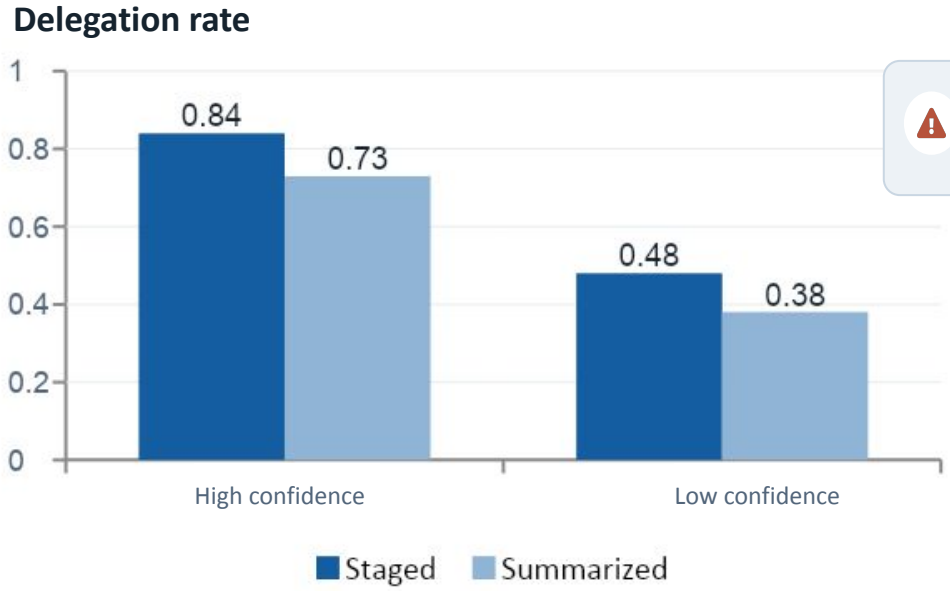
Share of decisions where teachers accepted the AI mark unchanged - compared across explanation form × confidence

Diagnosticity

LR+ / LR− of overriding AI's mark: how much an override shifts the odds that the AI mark is wrong.

Also measured: triage time, workload (NASA-TLX), and post-task self-reported trust.

Teachers delegate more with staged explanations



Attitude-behavior gap: teachers rated summaries more preferable in post-task survey, but delegated more with staged explanations (§6.5.5).

Delegation rate (Table 7): staged > summarized at both levels; the confidence badge throttles reliance.

Staged explanations make teacher oversight more diagnostic

How diagnostic is a teacher's override of TC-MAG's grade?

LR+ 11.5

Staged - an override is a strong signal the AI mark is wrong

LR+ 4.6

Summarized - an override is only a moderate error signal

⚠ Staged: low sensitivity → over-reliance risk (missed AI errors)

⚠ Summarized: low specificity → over-correction of right marks

Confidence-adaptive progressive disclosure of AI explanations

Transferable pattern for AI-assisted judgment: calibrated confidence sets the depth of explanation.

HIGH → staged micro-steps. Preserves correct marks: harmful overrides 0.14 vs 0.24 with summaries; serves accountability.

LOW → summary first, expand on demand. Surfaces risk fast since teachers need to re-mark low conf cases anyway; one click opens full staged detail.

Scope: typed answers · Singapore primary math · experienced teachers.

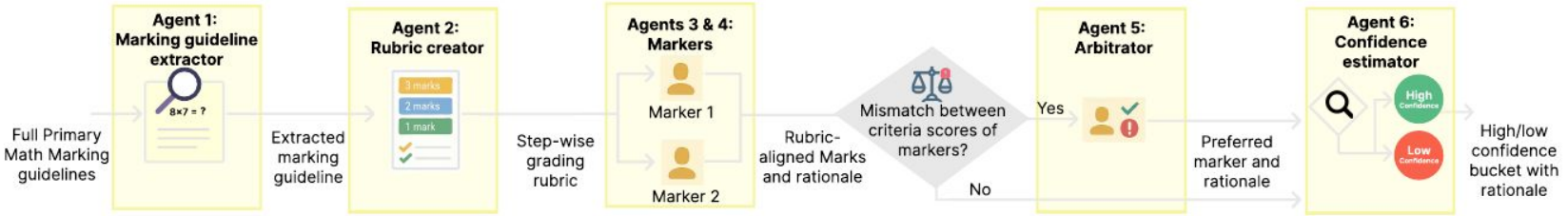
Three guidelines for delegable AI assistance

- **Lead with the decisive criterion -** open with the evidence that moved the grade (e.g., “unit missing: −1”)
- **Capture override reasons -** high LR+ makes overrides a rich signal for prompt refinement
- **Guard against over-reliance -** calibration exercises; “are you sure?” nudges on high-confidence overrides

Scope: typed answers · Singapore primary math · experienced teachers.

SUMMARY

Teacher-cognition multi-agent grading: deployment-level reliability, delegable oversight



QWK 0.936

reliability on 2-4-mark partial-credit grading

> human markers

Δ = +0.168 QWK on the hardest questions

LR+ 11.5

diagnostic teacher oversight with staged explanations

Sanskriti Uma

Graduate Student at Harvard University

Formerly: Senior AI/ML Product Manager at Geniebook, Singapore

Paper (DOI)

Contact

ANNEX SLIDES

Ablation · Error analysis · Full results · Confusion matrices · Marker sweep · Cost & latency · Triage time · Ethics & COI

Ablation: every agent earns its place (Table 2)

Configuration	1-mark	2-mark	3-mark	4-mark	Avg. accuracy drop
Original - six agents	0.984	0.930	0.864	0.840	-
Guideline extractor removed	0.985	0.900	0.840	0.735	**−3.95%**
Rubric creator removed	0.985	0.885	0.845	0.740	**−4.08%**
Arbitrator removed	0.984	0.906	0.834	0.780	**−2.85%**

Held constant: LLM prompt context. Removed extractor/creator prompts were folded into the Marker agents; removed Arbitrator replaced by random marker selection. **Largest losses concentrate on 4-mark questions (−10% without the Rubric Creator) - exactly where partial-credit structure matters.**

ANNEX · RQ1

Where TC-MAG still errs (§5.6)

37% — **Partial-credit rubric mismatch**

Multi-answer / mixed-sequential questions break the one-point-per-step template. Fix: add a Rubric-Type Classifier agent conditioning the rubric template.

26% — **Marking guideline gaps**

Topic-specific rules absent or ambiguous in the source document → markers improvise and diverge. Fix: make source documents pedantic for LLM use.

21% — **Marker prompt drift**

Lenient credit for unlabeled intermediate values; position-sensitivity in long prompts + training-data strictness bias.

4% + 3% — **Rubric / Arbitrator drift**

Keyword errors (nouns treated as units); occasional judge biases (position/verbosity). Fix: unit whitelist.

Share of cases where TC-MAG's mark differed from gold; two researchers coded independently, reconciled by discussion.

ANNEX · RQ1

Full comparison vs. LLM baselines (Tables 4-5)

1-mark (binary) - Cohen's κ

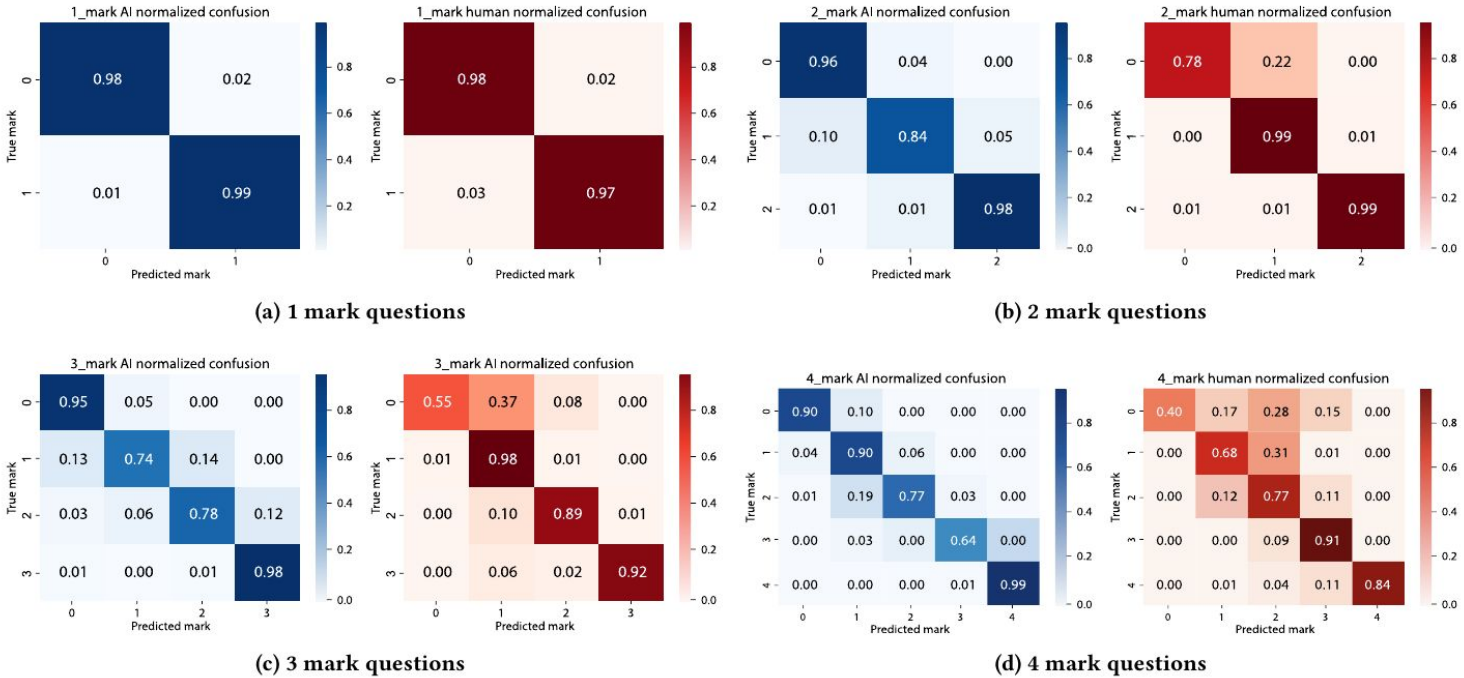
Model	κ	Exact	p (adj.)
GPT-4o vanilla	0.848	0.924	<.001
GPT-4o CoT	0.900	0.950	<.001
GPT-4o TC-MAG	0.920	0.960	<.001
GPT-o3 vanilla	0.960	0.980	.500
GPT-o3 CoT	0.968	0.984	.425
TC-MAG (GPT-o3)	**0.968**	**0.984**	-

2-4-mark (ordinal) - QWK

Model	QWK	Exact	MAE
GPT-4o vanilla	0.788	0.643	0.449
GPT-4o CoT	0.799	0.655	0.423
GPT-4o TC-MAG	0.811	0.685	0.385
GPT-o3 vanilla	0.918	0.809	0.211
GPT-o3 CoT	0.924	0.813	0.199
TC-MAG (GPT-o3)	**0.936**	**0.878**	**0.129**

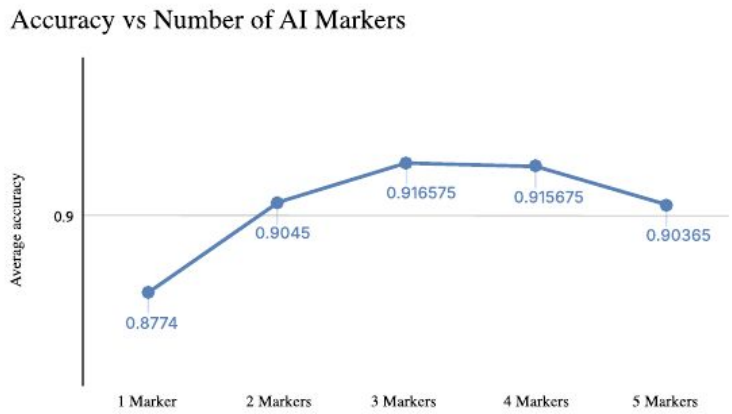
All ordinal baselines significantly inferior (adj. $p < .001$). On binary 1-mark, GPT-o3 CoT ties TC-MAG - the architecture's gains concentrate on partial-credit grading. GPT-o3 > GPT-4o reflects model quality; TC-MAG improves both.

Confusion matrices, all strata (Fig. 4)



TC-MAG (blue) near-diagonal across strata; humans (red) lenient on zero-credit answers, larger off-diagonal errors at higher marks.

Why two markers (Fig. 2)



Accuracy vs. number of marker agents (motivational study, 400 pairs)

- **+2.71% from 1⟶2 markers**
- +1.21% from 2⟶3 markers
- Degrades beyond three - token overhead burdens the arbitrator
- Two markers balance accuracy vs. cost; arbitration fires even when totals agree but criterion-level marks diverge

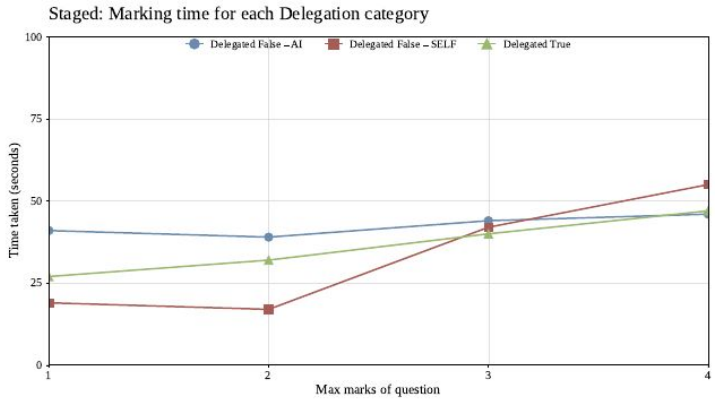
Cost and latency per question (Appendix C)

Configuration	Mean cost / question	Cost / 10,000 q	Runtime / question
TC-MAG · GPT-o3 · standard	$0.0348	$347.66	27-42 s
TC-MAG · GPT-o3 · batch	$0.0174	$173.83	(batch API)
GPT-o3 CoT · standard	$0.0169	$169.20	18-35 s
GPT-o3 vanilla · standard	$0.0130	$130.15	21-29 s

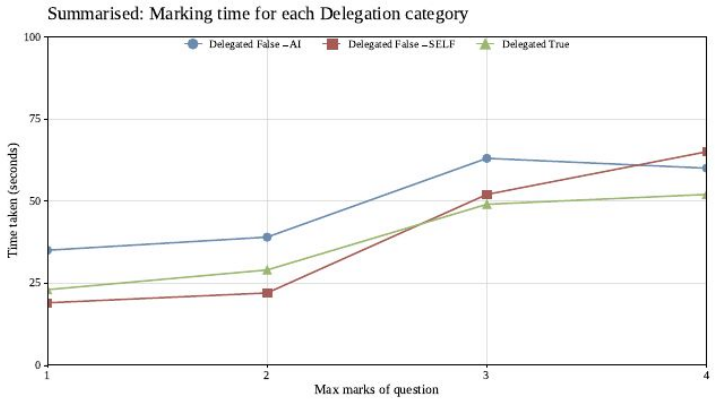
TC-MAG costs ≈ 2× a CoT baseline (~15,500-17,100 tokens/question on 3-4-mark items) and buys the ordinal-reliability gap plus the audit trail. **Compute, connectivity, and cost remain real barriers in low-resource contexts (§7.4.3); cheaper models work with the framework but the accuracy drop (0.936 → 0.811 macro) can prohibit delegation.**

OpenAI API prices as of Oct 4, 2025 (Table 19).

Triage time by explanation form (Fig. 7)



(a) Staged variant



(b) Summarised variant

- Self-marking is fastest on 1-2-mark, slowest on 4-mark questions
- Staged keeps 3-4-mark review time moderate and below self-marking - better default for high-mark items
- Summaries are quickest on 1-2-mark, but cost +15-20 s vs. staged when teachers open AI and then overturn on 3-4-mark

Ethics, data, and conflict management (§9)

Approval & consent

Written approval from the partner organization's IRB-equivalent; written participant consent; right to withdraw; responses could not affect employment.

Data protection

All student responses de-identified before analysis; OpenAI API used with enterprise data opt-out (no retention or training on student data).

Conflict of interest

Two authors are employees of the data-providing edtech partner; one independent academic collaborator oversaw design and analysis of de-identified data.

Mitigations

Pre-existing marking protocols with blind senior-teacher adjudication; pre-planned analyses; all tested conditions reported; partner had no veto over publication.