Smarter Together: Enhancing Human-AI Collaborative Grading With Teacher-Cognition Multi-Agent LLM Framework

Sanskriti Uma
Geniebook
Singapore, Central Singapore
Singapore
sanskriti.uma@geniebook.com

Surjya Ghosh
Department of Computer Science and Information Systems
BITS Pilani Goa
Zuarinagar, Goa, India
surjyag@goa.bits-pilani.ac.in

Dio Dzaky Achmad Mustaqim
Geniebook
Surabaya, Indonesia
dio.dzaky@geniebook.com

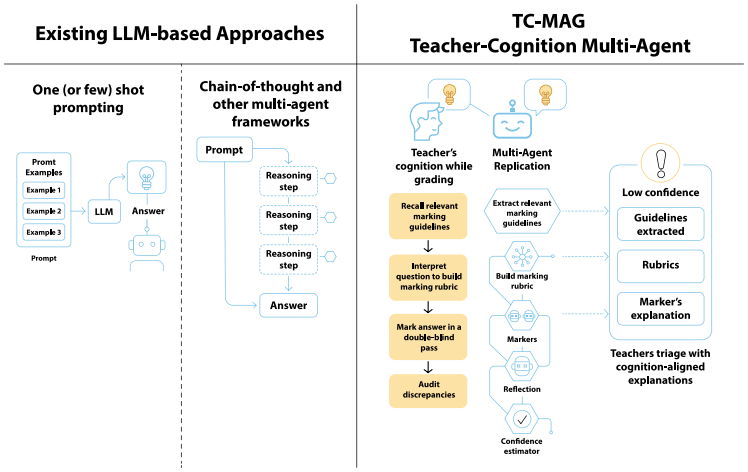


Figure 1: Overview of the TC-MAG framework in contrast to the existing single-pass LLM-based automated grading. The figure highlights TC-MAG 's teacher-cognition pipeline (guideline extraction → rubric creation → parallel markers → arbitration → calibrated confidence) with short, verifiable micro-explanations and a High/Low confidence tag for selective teacher review.

Abstract

Automated grading in rubric-based, short-answer open-ended questions often mishandles partial credit, calibration, and actionable transparency, ultimately requiring teachers to reevaluate. This challenge is amplified in resource-constrained settings (e.g., with limited teachers and a large student population), resulting in weaker learning outcomes. To address this challenge, we present the Teacher-Cognition Multi-Agent Grading framework (TC-MAG), which mirrors teachers' micro-steps via anchored LLM agents for rubric creation, guideline checks, blind double marking, arbitration, and cross-checking to calibrate confidence. Each step produces a concise explanation for targeted review. We first conducted a motivational study to inform the design of TC-MAG. Next, we validated the effectiveness of the TC-MAG framework on a dataset of 2,000 Singapore primary school students' responses across 1–4-mark mathematics questions with teacher-adjudicated gold labels. TC-MAG attained deployment-level reliability (κ=0.968 on 1-mark; quadratic-weighted κ=0.936 on 2–4 marks) by outperforming human teachers ($\Delta\kappa$=+0.063, p<.001) and state-of-the-art LLM baselines (min $\Delta\kappa$=+0.012, p<.001). In a mixed-methods teacher study (N=14; 12.1 years' experience), explanation format and TC-MAG 's confidence score influenced whether teachers delegated grading to TC-MAG. Staged explanations yielded greater diagnosticity (LR+ 11.5 vs. 4.60 for summarized explanations), informing a progressive disclosure strategy of explanations based on confidence. Overall, TC-MAG offers replicable multi-agent framework and triage methods for classroom deployment while preserving teacher oversight.

CCS Concepts

• **Applied computing** → *Computer-assisted instruction*; **Education**; • **Human-centered computing** → *Systems and tools for interaction design*; • **Computing methodologies** → *Artificial intelligence.*

Keywords

Human–AI collaboration, LLM Application, Automatic Grading, K-12 education

IUI '26, Paphos, Cyprus

ACM ISBN 979-8-4007-1984-4/26/03
https://doi.org/10.1145/3742413.3789130

ACM Reference Format:
Sanskriti Uma, Surjya Ghosh, and Dio Dzaky Achmad Mustaqim. 2026. Smarter Together: Enhancing Human-AI Collaborative Grading With Teacher-Cognition Multi-Agent LLM Framework. In *31st International Conference on Intelligent User Interfaces (IUI '26), March 23–26, 2026, Paphos, Cyprus.* ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3742413.3789130

1 Introduction

Marking open-ended student work at classroom scale is a binding constraint on equity: in resource-constrained settings, students receive fewer constructed-response questions, wait longer for feedback, and accumulate less durable knowledge than peers who have access to individualized instruction [90]. Primary-school teachers spend a disproportionate share of their time grading open-ended answers that cannot be machine-scored [35]. To keep pace with syllabus, classrooms often shift toward multiple-choice items; yet corpus studies and classroom analyses show recognition-based questions surface far fewer misconceptions than generative responses, narrowing what assessment can diagnose [62, 76]. Timely feedback matters too - reducing latency from a week to a day halves correction time for conceptual bugs and aligns with spaced-practice cognitive models [35, 82]. Therefore, we aim to automate grading for open-ended questions to reduce teachers' effort and improve learning outcomes.

Prior works span feature-engineered short-answer scoring pipelines [62], advanced transformer-based grading mechanisms [87], LLM-based automated scoring approaches using prompt engineering like few-shot learning, etc. [17, 26, 103]. However, existing approaches suffer from various shortcomings. **First**, systems often fail to achieve human-level inter-rater reliability in high-stakes grading. For example, on ASAP-SAS, the best GPT-4 prompting reaches QWK of 0.677 [40]. Moreover, LLMs match humans only on subsets of questions and can be gamed by prompt-hacking [15]. **Second**, zero-/few-shot prompting performs poorly, and without fine-tuning, GPT-4 falls to 'not feasible for practical use' in three benchmark corpora [11]. **Third**, uncertainty is untrustworthy - confidence drifts on longer/partial answers, rubric non-compliance is common, and 22% of the mis-scores being high-confidence undermine selective review. [15, 37, 38]). **Finally**, post-hoc 'explain your score' prompts do not solve this; empirical evaluations rate such explanations as inconsistent and pedagogically thin, yielding little time savings because teachers still re-mark the original answers [40]. Prior HCI work similarly finds that teachers delegate grading only when systems surface clear uncertainty cues and support low-friction overrides [2, 78, 102]. These shortcomings highlight the necessity of a comprehensive framework for better AI grading reliability and high degree of trustworthiness while aiming to reduce the teachers effort and obtaining better learning outcomes.

Research Challenges: However, developing such a trustworthy and reliable AI grading framework poses multiple research challenges as discussed next. *High reliability requirement* - AI graders must exhibit a high reliability agreement while handling partially correct, multi-step reasoning. Eliciting structured reasoning helps as LLMs are more accurate when they externalize intermediate steps or explore multiple branches, yet chain of thought alone does not ensure verifiable grading [27, 95]. *Calibrated uncertainty and auditability*, selective teacher review is viable only if confidence tracks correctness. However, LLMs are miscalibrated on open-ended answers, and unverified explanations can increase over-reliance [2, 37]. *Usable transparency for human–AI orchestration*, teachers delegate when systems surface clear uncertainty cues and enable low-friction overrides [2, 78, 102]

Research questions: In this paper, we ask the following research questions while aiming to develop a trustworthy AI grader addressing the aforementioned challenges. **RQ1 (Reliability)**, can a teacher-cognition–inspired multi-agent[1] LLM grader surpass the reliability requirement for deployment on primary school short answer mathematics questions, by outperforming human teachers and existing LLM baselines **RQ2 (Teacher trust and delegation)**, does the presentation of micro-step explanations (anchored on teacher micro-decisions) and calibrated confidence buckets change teachers' willingness to delegate grading to AI? To address these research questions, we develop TC-MAG (Teacher-Cognition Multi-Agent Grader), a six-agent framework mirroring teachers' micro-decisions while marking: identify the relevant rubric step, extract evidence from the student answer, check alignment with marking guidelines, assign provisional marks, solicit a second opinion, reconcile disagreements, and calibrate confidence. The framework operationalizes advances in LLM reasoning by forcing each micro-decision to be committed and logged before the next, reducing hidden leaps that make single-pass grading brittle [27, 95]. We stabilize decisions by aggregating multiple reasoning paths and reconciling them via critique [59, 92], and by interleaving verification prompts that explicitly check outputs against guideline text - a strategy shown to improve factual and mathematical reliability [92]. To support selective review, every agent emits a short, verifiable explanation tied to specific guidelines; a cross-checker flags mismatches and down-ranks confidence, producing discrete confidence buckets that teachers can act on [2, 58]. We demonstrate the framework of TC-MAG in Fig. 1, highlighting differences from the traditional LLM prompting frameworks.

We evaluated TC-MAG on short-answer primary mathematics against LLM baselines and conducted a mixed-method classroom study with practicing teachers to measure trust, delegation, and audit time under different explanation and uncertainty configurations [37, 40, 102]. In a study involving 2,000 Singapore primary-school answers, TC-MAG exceeds human and LLM baselines in QWK across all mark strata, e.g., increase from 0.79 (human baseline) to 0.95 (TC-MAG), while preserving teacher oversight. In summary, we make the following contributions in this paper,

- A teacher-cognition–inspired [2] multi-agent LLM framework for grading short answer, rubric-based mathematics questions. We introduce TC-MAG, a six-agent framework that decomposes grading into teacher micro-decisions, committing to and logging

[1]We use 'multi-agent' to denote a system of LLMs with specialized functional roles with distinct prompts and information boundaries, following conventions in MetaGPT [34], CAMEL [54], and AutoGen [100]. TC-MAG comprises five such roles. We acknowledge that stricter definitions requiring model heterogeneity or persistent inter-invocation state have been proposed; our usage reflects the current predominant convention in LLM applications literature.

[2]Here, we use 'teacher cognition' in a procedural sense to refer to the the cognitive-task-analysis elicited sequence of micro-decisions teachers externalize during grading. We do not claim a validated cognitive theory or computational model of teachers' judgment.

each bounded step before moving on. The design builds on evidence that structured, multi-step reasoning improves reliability over single-pass outputs [27, 95].
- Human–AI orchestration insights for trustworthy grading. Mixed-methods analyses show how micro-step explanations (anchored on teacher micro-decisions) and confidence buckets shift teachers' willingness to delegate grading and concentrate attention on likely failure points, addressing documented barriers to adoption in classrooms [2, 78, 102].

These findings underscore the effectiveness of a multi-agent LLM framework, demonstrating the reliability and trustworthiness of automatic grading required for real-world deployment.

2 Related Works

In this section, we present the related literature on LLM-based automated grading and the challenges associated with wide-scale deployment of such approaches. In addition, we highlight how the proposed framework aims to address some of these challenges.

2.1 LLM-based Automated Grading

In the existing literature, LLM-based automated grading can be broadly divided into the following categories - single-pass LLMs, rubric-based, AI-assisted assessment (including human-in-the-loop and few-shot grading) and prior multi-agent prompting.

Single-pass LLM based automated grading uses a single prompt that combines the assignment description, grading rubric, and student response, allowing the model to generate a score and feedback in one inference step [28, 29]. It relies on the LLM's ability to interpret instructions and apply rubric-based reasoning directly, without multi-step evaluation or external code execution. This approach enables fast, scalable, and consistent assessment across many submissions. However, the major limitations of these approaches are that they suffer from hallucinations, prompt sensitivity, and reliability on edge cases [28].

Rubric-based LLM for automated grading is used to overcome some of the limitations observed in single-pass LLM. This grading method uses structured evaluation criteria - typically expressed as a rubric with specific dimensions like correctness, clarity, and completeness - as explicit guidance for the model during assessment [8, 73]. The LLM processes the rubric and student submission in a single or multi-turn prompt, assigning per-criterion scores and generating targeted feedback aligned with each rubric item. This approach improves transparency and consistency by guiding grading decisions in well-defined standards. However, this approach often suffers from the quality of the rubric and may not generalize across different types of grading tasks [8, 81].

AI-assisted LLM-based grading leverages large language models (LLMs) to support automated grading while keeping humans involved in the decision loop to ensure accuracy and fairness [18, 55, 103]. In this framework, the LLM first generates draft evaluations - such as preliminary grades, rubric-aligned comments, or qualitative feedback, based on the student submission and grading criteria. Human graders then review, validate, or adjust these AI-generated outputs, which reduces grading workload while maintaining accountability [107]. The few-shot grading variant improves the LLM's calibration by showing it a small number of graded exemplars, helping the model better align its scoring with human standards [107]. This combination enables scalable and explainable grading that still preserves human oversight, especially in open-ended tasks like essays or programming assignments. Overall, AI-assisted assessment seeks to combine LLM efficiency with human judgment to enhance both grading consistency and teacher productivity. However, it remains limited by bias propagation from exemplars, and the continued need for human verification to prevent hallucinated or inconsistent assessments [6, 107].

In *LLM-based multi-agent* approaches, the grading task is split across specialized agents - each handling a sub-task (e.g., extracting rubric components, generating feedback, checking rubric compliance, or refining guidelines) rather than having a single model do everything at once. For example, the AutoSCORE [93] framework uses one agent to extract rubric-relevant components from student responses and another to assign the final score based on those components. Similarly, GradeOpt [18], uses a trio of agents (Grader, Reflector, Refiner) where the Reflector identifies errors or misalignments compared to grading guidelines, the Refiner updates them, and the Grader applies them. However, these systems suffer from inconsistent behavior across tasks and heavily dependent on expert validated grading rubric and error detection mechanisms to avoid propagating mistakes.

2.2 Adopting LLM-based Auto-graders in Real-world Classroom Setting

Despite the availability of various LLM-based auto-graders, their adoption is limited in a real-world classroom setting anchored in three converging lines of evidence. **First**, *Normative psychometric guidance*, current psychometric standards hold that when 'subjective judgment enters into test scoring, evidence should be provided on inter-rater consistency commensurate with the stakes of the decision' [99], and leading automated-scoring frameworks therefore foreground Quadratic Weighted Kappa (QWK) [23] as the most interpretable reliability index for operational approval. **Second**, *Interpretive frameworks for operational practice*, empirically, gold-standard human-human agreement in large U.S. K-12 programmes cluster in the 0.83 - 0.92 range, Texas STAAR 2024 writing tasks report human QWKs up to 0.92 [88]. Landis & Koch proposed the seminal scale [53] $\kappa > 0.81$ as 'almost perfect' agreement. Later, Williamson et al. 2012 proposed that in high-stakes automated scoring, AI grades must match expert human agreement and remain within ≈ 0.10 of κ of the human baseline [97]. Due to a lack of fixed deployment kappa criteria, we therefore evaluate reliability **relative to the human baseline** and report kappa(for binary grading) and QWK(for ordinal grading).

Key Takeaways: In summary, while evaluating open-ended questions, LLM-based autograders rely on techniques such as rubric-guided prompting, few-shot exemplars, and chain-of-thought or self-consistency reasoning to enhance scoring reliability and alignment with human judgment. Some systems incorporate human-in-the-loop mechanisms where LLM predictions are calibrated or verified by educators to reduce bias and improve transparency. Others employ multi-agent or debate-based setups where multiple LLMs critique and justify each other's grading decisions to improve

fairness and explainability. However, still these approaches face limitations in subjectivity, limited interpretability of explanations for grading decisions, and demand a higher degree of reliability for high-stake assessments so that teachers willingly delegate grading to AI, which we aim to address in this work.

3 Motivational Study

The objective of this study is to design LLM prompts and to understand the limitations of an off-the-shelf single LLM agent framework for grading tasks. We conducted a two-part pre-study. First, a cognitive task analysis (CTA) with primary school teachers revealed their grading workflow. Second, we implemented this workflow using a single LLM agent with multi-step prompts and identified its specific failure points. We adopt a decomposed stepwise prompt design because prior work shows that structured intermediate reasoning aggregation reliably increases accuracy [96, 110]. We describe them next.

3.1 Cognitive Task Analysis for Prompt Design

The goal of this workshop was to map the micro-decisions that expert teachers make while grading. We conducted a Cognitive Task Analysis (CTA) workshop with primary math teachers (N = 5, mean 7 years' experience). We stopped at five participants because the behavioral patterns and insights reached thematic saturation, likely due to similar training and experience [63]. We asked participants to mark a previously unseen[3] question (to mitigate learned shortcuts and to surface the original decision process) by thinking aloud. Two authors of the paper independently coded the think-aloud sessions and resolved disagreements by discussion with a senior teacher (15 years' experience). This qualitative analysis revealed a consistent four-stage behavioral chain to ensure the accuracy, fairness, and defensibility of the marks awarded. We note these findings below, which guide our LLM prompt design,

- **Internalizing the official marking guidelines:** Teachers first recall all relevant marking guidelines for the question at hand. Our partner edtech maintains a 'Marking guidelines document' that outlines common errors in student answer representations or units, specifying which cases deserve credit and which don't, based on the topic and level of the student. For example, for a 'Money' topic question with fraction calculations, teachers recalled relevant guidelines on the presentation of Money and fraction values, based on their memory of the 'Marking guidelines document'.
- **Externalizing guidelines into a rubric:** It is crucial to generate a grading rubric of one point per intermediate step and full points for the final answer to ensure consistent grading of a question, in accordance with the Singapore curriculum. Based on the 'Marking guidelines document' and the question text, the grading rubric should specify which steps require students to write 'units' to receive full credit.
- **Performing independent double-marking:** We saw cases where two teachers gave the same marks but for different reasons. For example, one teacher deducted 1 mark for missing units; another teacher deducted 1 mark for both missing units and incorrect answer representation. When another student's answer was shown in which the units were corrected but the representation remained incorrect, the first teacher gave full marks, which was ruled wrong by a senior teacher. They clarified that 1 mark is deducted for missing/incorrect units and/or presentation errors. We hypothesized that, for multiple AI grading passes of the same question-answer pair, LLMs could also produce identical marks with different explanations, risking unfaithful explanations. We therefore created a **criterion level** for each rubric step. For example, a 1-mark question that evaluates only the final answer includes criteria: (1) correct numerical or textual value, (2) correct unit, (3) correct answer representation (e.g., mixed fractions if specified; improper fractions not accepted).
- **Reconciling any disagreement:** If there is discrepancy in marks by the two human graders (in double blind grading), a third grader arbitrates disagreements. Any arbitrator must cite exact sources from rubrics and marking guidelines to provide **evidence-grounded** explanations [89] when analyzing upstream errors to ensure transparency among all graders.

3.2 Reliability Study: Multi-step Prompts

In this section, we discuss the shortcomings of single-agent multi-step prompting and how multi-agent system with multiple passes for key actions should be designed for better performance. We describe them next.

3.2.1 Question-Answer Dataset. We curated a dataset that contains 400 question-student answer pairs from the Singapore primary mathematics curriculum (Primary 1–6). It contains correct solution and maximum marks for each question. The corpus is divided equally into 1 and 2 mark questions (200 per stratum), with a balanced distribution of gold marks (e.g., 50/50 correct/incorrect for 1-mark; 33% per category for 0–2 marks for 2-mark questions). Further details about the questions dataset are mentioned in Section 5.1.

3.2.2 Study Procedure. We replicated the teacher marking process in a multi-step prompt [110] for higher faithfulness to instructions for each step of the marking process (using a single LLM agent), and used the finer observations from our Cognitive Task Analysis to design the prompts. We designed a three step prompt, consisting the steps for extracting relevant marking guidelines, creating marking rubrics and marking student's answer. We did not include the double blind evaluation and disagreement reconciliation steps, as we wanted to test the efficacy of a single agent first. Inputs included: marking guidelines, question, correct solution, student answer, and maximum marks. We summarize the steps with the prompt snippet in Table 1. We tested the prompt described above using OpenAI's GPT-o3 model [68].

3.2.3 Key Findings. We used the prompt strategies described in the Study procedure section, and obtained an accuracy of 91.23% (98% for 1 mark questions, 84.46% for 2 mark questions). Based on the single agent's step-wise output (extracted marking guidelines,

[3]When teachers know a question and typical student responses, they often skip explicit guideline lookup or rubric drafting because they recall common answer–mark patterns. They return to guidelines when a student's answer is novel in presentation or procedure.

Step number	Prompt snippet
Step 1: Internalizing marking guideline	Given the full marking guidelines and one question, for this step, produce a copy of only the rules needed to grade this question ($\leq$ 250 words). Keep section / rule numbers unchanged.
Step 2: Externalizing guidelines into a rubric	Create marking rubrics for a {max_marks} mark question. Parse the solution into chunks of intermediate computational steps, and label their step_code as $S_1, S_2, \ldots, S_k$, and Final answer. Each intermediate step has a maximum `mark_value` of 1 mark, and the Final answer has a maximum `mark_value` of {max_marks} mark.
Step 3: Mark based on rubric and criteria	Mark a student's answer using the provided question, solution, and maximum marks. Marking criteria for the Final answer steps are as follows: *1. Criteria 1 – Answer Equivalence:* If the student's final answer is mathematically equivalent to the `evidence_required` mentioned in the Marking Rubrics, assign 1 for this criterion. *2. Criteria 2 – Unit Accuracy:* If `units_required` is mentioned, verify the unit correctness. If `units_required` is `none`, assign 1 for this criterion. *3. Criteria 3 – Answer Format and Clarity:* Confirm that the student's answer format follows the marking guidelines.

Table 1: Prompt snippets used in the multi-step pipeline (key instructions for each agent stage).

marking rubric, and marking process breakdown and explanation), we could isolate the main error causes for incorrect marks. Two authors of this paper independently coded all incorrect AI marks (different from gold labels) and resolved disagreements with a senior teacher (15 years' experience). Notably, we observed several reasons causing drift (i.e., the deviation in the LLM output from the instructions in the prompt input). We elaborate these next.

- **Prompt drift in marking (32.8% errors)**: . Broadly, there are two reasons for this. *Lack of precision*, in this case, mark awarded if the correct value appeared somewhere rather than as a precise final answer (e.g., awarding full marks if the student listed all factors of two numbers when the question asked for the highest common factor). *Inconsistent partial marking*, in this case, an intermediate result without a label received marks, while other times, an intermediate result with a label did not get any marks.
- **Prompt drift in rubric adherence (30.4% errors):** This stems from several reasons such as awarding full marks in case the solution is correct but it does not include the required unit (e.g., student wrote 5027 instead of 5027 gm).
- **Prompt drift in answer presentation and formatting (15.7% errors):** We noted instances like *lexical or presentation tolerance*, i.e., near-miss spellings or presentation variants treated as correct and *inability to recognize equivalent forms*, i.e., fraction or decimal answers not accepted as the same value for specific question types.
- **Gaps in the original marking guideline document (21.1% errors):** All allowable answer formats were not pedantically listed in source document (e.g., all variations of time formats: correct - 09 00, 9:00 AM; wrong - 9 00 AM, 09 00 AM, 09:00 AM).

In summary, using a single LLM agent to perform a sequence of cognitively distinct tasks - rule extraction, rubric synthesis, and rubric application - resulted in a cascade of errors. This vulnerability to drift, even in a chained prompt setup, motivated our multi-agent approach. Research shows that decomposing complex tasks into focused sub-problems mitigates 'lost-in-the-middle' effects where LLMs lose track of initial instructions [57]. By designing agents with orthogonal scopes - each mirroring a specific stage of teacher cognition - we can ensure that the output of one stage is a constraint for the next. Our design is grounded in both observed human expertise and established principles for creating reliable and verifiable AI systems [5, 84].

Additionally, we also investigated the need for independent double-marking and reconciliation (as seen in the CTA). Multiple passes can raise accuracy [91, 108], so we tested the accuracy of multiple marker agents (the key actors in TC-MAG) and a single adjudicator. Moving from one to **two** marker agents yielded a **2.71%** accuracy gain; a third marker gave **1.21%** additional gain; beyond three, performance degraded (Fig. 2), likely due to token overhead affecting the adjudicator agent. We therefore selected a **2-marker** assembly to balance token cost and accuracy. Notably, we trigger the arbitrator to adjudicate even when markers agree on the total mark but diverge at the criterion level (as noted in Section 3.1). Based on all these findings, we now describe the end-to-end framework in the next section.

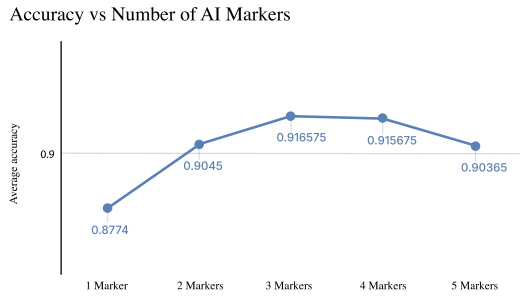


Figure 2: Accuracy vs. number of marker agents. Gains plateau after two markers (+2.71% from 1→2; +1.21% from 2→3), with degradation beyond three due to arbitration overhead.

4 TC-MAG Framework

In this section, we describe the TC-MAG framework (Fig. 3). It comprises of six agents (guideline extractor, rubric creator, markers (two), arbitrator, and confidence estimator) and demonstrates how the grading for the questions are done based on the interaction among these agents. Next, we discuss the role of each agent.

4.1 Guideline Extractor

The goal of this agent is to extract relevant marking guidelines for the question to be marked. The inputs provided are the full version of the "primary math marking guideline" (created by our partner edtech organization - as noted in Section 3.1), question and solution. We prompt the agent to extract the relevant marking guideline, and it outputs the extracted marking guidelines for a particular question and solution.

4.2 Rubric Creator

This agent works to create grading rubrics. It takes the following as inputs - question, solution, the extracted marking guidelines (from the previous agent) and the maximum marks for the question. We prompt the agent to create a step-wise grading rubric by identifying essential intermediate steps and final answer from the solution. We also prompt the agent to identify essential numerical (or textual) evidence required for each step to be correct and identify the unit requirements based on the extracted marking guidelines. The outputs from this agent are step-wise grading rubric with essential intermediate steps and the final answers. It also includes the units required (if any).

4.3 Markers

We employ two passes of the marking agent (called agents 3 and 4 for distinctive labeling in the framework) to score the student's answer. The dual-pass design exploits documented non-determinism in OpenAI's API (even with temperature=0) to highlight meaningful disagreement [64, 69]. We provide each agent the following inputs - marking guidelines, rubric, question, solution, maximum marks and student's answer. We prompt each agent to grade the student's answer for each step in the rubric based on three criteria (numerical equivalence, unit accuracy, and answer format). We also instruct to use the extracted marking guideline to assess alternative answer representations for assessing answer format. The agents output the following: final marks and criteria-level marks, along with explanation for marks awarded for each step of the rubric.

4.4 Arbitrator

The role of this agent is to resolve any discrepancy between the two marking agents (in criteria-level marks for any step). We provide this agent with all the inputs that were provided to the marking agents, along with the outputs from the two marking agents. We prompt the agent to decide which agent is correct. The agent provides the output of the preferred marker along with the explanation for its preference, citing the exact rubric step where marking erred or the exact line number of the extracted marking guideline missed.

4.5 Confidence Estimator

This is the final agent of the framework (Fig. 3). It takes the following as input - marking guidelines, grading rubrics, questions, solutions, maximum marks, student's answer, marker 1 output, marker 2 output (if there is discrepancy between marker 1 and marker 2), and arbitrator's output (if any). We prompt the agent to estimate confidence by assessing faithfulness of previous agents to their prompts, and flag inconsistencies . We obtain the confidence tags (high or low) and the explanation behind it as output. We introduced a separate confidence agent (rather than embedding confidence in earlier prompts to avoid lost-in-the-middle effects [83]; [105].

In summary, TC-MAG operates as a *single causal chain* - guidelines → rubric → twin grounded reasoners → arbitrator →anchor-audited confidence. By embedding *grounded-attribution CoT* in every link and stress testing it via disagreement and ablation, we make the framework faithful and verifiable.

4.6 Motivation for Multiple Agents: Empirical Findings

In this section, we empirically back the motivation of having multiple agents (in the TC-MAG framework). We ablated one agent at a time while holding the LLM prompt context constant, and tested on 800 question-student answer pairs from the Singapore primary mathematics curriculum (Primary 1–6). The corpus is divided equally into **1-**, **2-**, **3-** and **4-mark** questions (200 per stratum), with a balanced distribution of gold marks (50/50 correct/incorrect for 1-mark; 20% per category for 0–4 marks for 4-mark questions). Further details about the questions dataset are mentioned in Section 5.1. When the Marking guideline extractor and Rubric creator were removed, their prompts were inserted into the Marker agents. When the arbitrator agent was removed, we adjudicated discrepancies between the two markers by selecting one marker at random. We present the ablation study findings in Table 2.

Results show that each agent contributes meaningfully to accuracy. Removing the **Rubric-Creator** produced the largest average drop (**4.08%**), with nearly **10%** loss on 4-mark **questions**. Removing the **Guideline-Extractor** had a similarly large impact (**3.95%**), indicating that relying on Marker agents to extract guidelines leads to significant errors. Removing the **Arbitrator** reduces accuracy by **2.85%**, confirming that reconciliation is crucial with two markers.

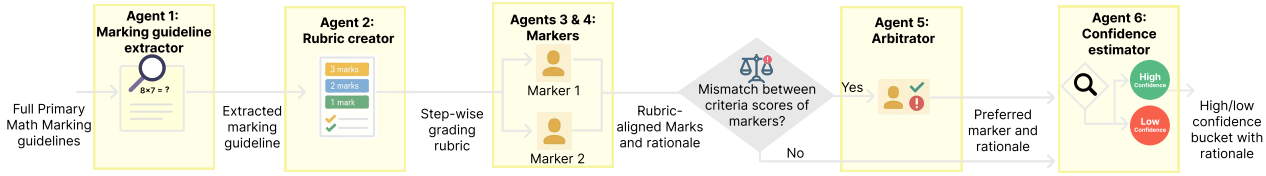


Figure 3: TC-MAG framework overview: the Guideline Extractor outputs extracted marking guidelines; the Rubric Creator generates a step-wise grading rubric; two Markers perform independent double-marking and output final marks plus criteria-level marks and explanations for numerical equivalence, unit accuracy, and answer format; the Arbitrator resolves any criteria-level discrepancy by citing the exact rubric step or extracted marking guideline line number; and the Confidence Estimator flags inconsistencies to output confidence tags (high or low) with an explanation.

Agent	1 mark	2 marks	3 marks	4 marks	Avg accuracy drop
Original (six agents)	0.984	0.930	0.864	0.840	–
Marking 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%

Table 2: Ablation study of TC-MAG with one agent removed at a time while holding the LLM prompt context constant. When the Marking guideline extractor or Rubric creator is removed, its prompt is inserted into the Marker agents. When the Arbitrator is removed, discrepancies are adjudicated by selecting one marker at random. The table reports per-stratum accuracy and average accuracy drop relative to the original six agents.

These findings further highlight the necessity of having different agents in the TC-MAG framework.

4.7 Disclosure on TC-MAG LLM Prompts

In the deployed LLM pipeline, each agent prompt contains additional scaffolding for reliability (e.g., schema enforcement, safety checks, edge-case handling). These additions do not change the task definition or grading policy, but they can change output stability and error rates in corner cases. The actual LLM prompts (excluding the variable question text, solution, extracted guidelines, rubric JSON, and marker outputs) are typically 10× longer than the excerpts for corresponding stages as seen in Table 1. TC-MAG prompts are the Intellectual Property of our partner edtech; however, we provide a redacted prompt template for each agent to aid replicability in A.5.

5 RQ1: Reliability of TC-MAG as an Autograder

In this section, we evaluate **RQ1**, i.e., reliability of the TC-MAG framework as an auto-grader. We report first the dataset, followed by the different baselines and experiment setup. We present the comparative analysis with the baselines and human evaluators and eror analysis subsequently.

5.1 Dataset

The dataset contains **2,000** question–student answer pairs from the Singapore primary-math curriculum (Primary 1–6); containing correct solution and maximum marks. The question set was divided into **1-**, **2-**, **3-** and **4-mark** questions (500 per stratum), with a balanced distribution of gold marks (50/50 correct/incorrect for 1-mark; 20% per category for 0–4 marks for 4-mark questions) to avoid prevalence-induced 'kappa-paradox' [23]. The questions contain arithmetic and word problems in English; 20% questions contain images. Singapore national curriculum emphasizes structured problem solving and multi-step constructions [44, 45, 60], which makes partial-credit accuracy central to reliable grading. The full questions dataset is Intellectual Property of our partner edtech organization, and with their approval, we present a set of 12 representative questions in Appendix A.4.

Gold-label Creation: For the creation of gold labels, each student's answer was double-blind-marked by primary-math teachers (mean 7 years' experience), and a **senior-teacher verified all marks** (15 years' experience) [23]. Further details regarding the questions dataset are provided in Appendix A.1. The question and student answer data are owned by a partner regional edtech provider and are protected intellectual property.

5.2 Baselines

In this section, we introduce the baselines along with reasoning for selecting them as baselines. We compare the performance of TC-MAG framework using GPT-o3 against (a) human teachers and (b) five LLM prompt variants (using GPT-4o and GPT-o3). We discuss these baselines next.

5.2.1 Human-based Baselines. For the creation of human marks, each student answer was independently double-blind marked by

two primary-math teachers (mean 7 years' experience) and a senior-teacher (15 years' experience) adjudicated the discrepancies. Teachers marked these questions as part of their regular duties, unaware that their grading accuracy would be assessed in our research study, thereby inducing effects of real-world tradeoffs, such as time pressure, in this dataset [52]. Teachers marked using our partner edtech's official marking guidelines and the same rubric conventions as the inputted in TC-MAG's prompts.

5.2.2 LLM-based Baselines. We used following five LLM variants as baselines. All variants have a a common I/O contract (Inputs: question, solution, maximum marks, student answer, marking guidelines; Outputs: JSON of rubric-criterion marks and the final mark). We enforce greedy decoding for all variants (temperature= 0, nucleus top-p = 1) to reduce output diversity versus stochastic sampling (higher temperatures or $p < 1$ increase diversity at the cost of stability [33, 39]).

(1) **GPT-4o vanilla:** Prompting strategies outlined in the Section 3.2 using OpenAI's GPT-4o model.
(2) **GPT-o3 vanilla:** Prompt identical to (1); using OpenAI's o3 reasoning model (but it may have internal CoT (Chain of Thought) systems that cannot be altered via prompting [68]).
(3) **GPT-4o with CoT:** Model identical to (1) but prompted to generate free-form natural-language reasoning before emitting the final mark.
(4) **GPT-o3 with CoT:** Identical to (3) but using OpenAI's GPT-o3 reasoning model.
(5) **GPT-4o with TC-MAG:** TC-MAG using OpenAI's GPT-4o model (to isolate architectural effects from model effects).

Reasoning for these LLM-based baselines: CoT (Chain of Thought) prompting is the most established single-model reasoning baselines for math word-problem reasoning [92, 95]. We include both (i) *vanilla* (no explicit CoT) and (ii) CoT variants to span the typical practice space. Program-of-Thought [14] motivates separating reasoning from computation but, since we do not calculate model answers or evaluate intermediate step calculations in student answers, we retain CoT as the closest prompting baseline. Other judge-style prompting schemes (e.g., LLM-as-a-judge [109] or majority-vote self-consistency [91]) are in effect, analyzed in our ablation study in section 4; when we remove other TC-MAG agents while keeping the arbitrator agent (equivalent of a judge), the model underperformed. We compare across two recent OpenAI families - GPT-4o and GPT-o3 reasoning line to isolate TC-MAG framework effects from model effects.

5.3 Evaluation Metrics

We use the following metrics to evaluate the performance of TC-MAG.

- **1-mark** (binary) questions: **Cohen's kappa (κ)** [24], **Matthews correlation coefficient (MCC)** [16], **Mean absolute error (MAE)** [98], **exact accuracy** and **F1-mark** [75] for 1-mark questions.
- **2–4 marks** (ordinal) questions: **Quadratic Weighted Kappa (QWK)** [24], **exact** and **within-1 accuracy** and **MAE** [98] for 2-4 marks questions.
- **Normalized confusion matrices (1-4 mark questions):** 2×2 matrices for 1-mark and full K×K tables for 2–4 marks to highlight over/under-scoring patterns or near-diagonal (near-miss) vs. far-off errors.
- We compute **95% CIs** via **stratified paired bootstrap** (1-4 mark questions; 200 resamples per stratum; paired by item), and assess **one-sided paired bootstrap** p-values for the null hypothesis (H_0 :*baseline κ/QWK $\geq$ TC-MAG κ/QWK*, adjusting for each baseline via **Holm–Bonferroni**.

We present formulae used for each metric in Appendix A.2.1. We assess TC-MAG and the baselines solely on marks because our dataset includes gold labels only for the awarded marks, and not for the question-level extracted marking guidelines or rubrics - these elements are non-unique and lie outside the main scope of this paper.

5.4 Findings: Comparison with Human Baselines

We present the comparative performance analysis between TC-MAG and human-based baselines in Table 3. Across 1-4 mark strata, TC-MAG achieves a higher macro-average κ/QWK (simple mean of κ/QWK) compared to human markers (0.950 vs 0.888).

5.4.1 Performance for 1-mark (binary) questions. The TC-MAG 's κ, MCC, balanced accuracy and F1-mark all exceed human consensus. However, a paired bootstrap test shows the difference in κ is not statistically significant.

5.4.2 Performance for 2-, 3- and 4-mark (ordinal) questions. TC-MAG achieves QWK values between 0.94 and 0.96 on 2- and 4-mark questions. Human QWK decreases markedly as the number of marks increases, particularly in the 4-mark questions where it drops to 0.79. TC-MAG significantly outperforms the human baseline on the 3- and 4-mark questions, while the 2-mark difference is not significant.

5.4.3 Confusion patterns. In this section, we present the systematic differences between TC-MAG and human markers using confusion matrices,

- In 1-mark questions, TC-MAG 's false positive and false negative counts are symmetric (2% and 1%, respectively). Human reviewers show a slight bias towards false negatives (3% vs 2%). The normalized matrices in Fig. 4a show that both systems rarely mis-label.
- In 2-mark questions, the TC-MAG slightly over-predicts full marks for mid-quality answers (10% instances of a true mark 1 predicted as 0 vs 5% predicted as 2, See Fig. 4b). Humans, however, leniently over-predict marks for many zero-credit responses (22% instances of a true 0 marked as 1).
- In 3-mark questions, the TC-MAG tends to err by at most one mark (predicted mark within ±1 in 96% of cases). Human marks exhibit both under-scoring and some over-scoring of mid-level answers (37% true 0 answers marked 1, and 10% true 2 marked 1, See Fig. 4c).
- In 4-mark questions, the human confusion matrix is markedly diagonal but with heavy off-diagonal elements: 16% answers

Task	Metric	TC-MAG (95% CI)	Human (95% CI)	Δ	Significance
1-mark	Cohen's κ	0.968 (0.944–0.988)	0.948 (0.916–0.976)	0.020	n.s.
	MCC	0.968 (0.944–0.988)	0.948 (0.916–0.976)	0.020	n.s.
	Exact acc.	0.984 (0.972–0.994)	0.974 (0.960–0.988)	0.010	n.s.
	F1	0.984 (0.972–0.994)	0.974 (0.959–0.988)	0.010	n.s.
2-marks	QWK	0.944 (0.924–0.962)	0.933 (0.909–0.954)	0.011	n.s.
	MAE	0.072 (0.050–0.096)	0.080 (0.054–0.108)	-0.008	n.s.
	Exact acc.	0.930 (0.908–0.950)	0.922 (0.902–0.940)	0.008	n.s.
	Within-1 acc.	0.998 (0.994–1.000)	0.998 (0.994–1.000)	0.000	n.s.
3-marks	QWK	0.933 (0.913–0.951)	0.882 (0.851–0.911)	0.051	$p < 0.006$
	MAE	0.148 (0.114–0.182)	0.198 (0.160–0.234)	-0.050	$p < 0.006$
	Exact acc.	0.864 (0.832–0.890)	0.838 (0.806–0.864)	0.026	$p < 0.006$
	Within-1 acc.	0.990 (0.980–0.998)	0.966 (0.936–0.982)	0.024	$p < 0.006$
4-marks	QWK	0.955 (0.945–0.965)	0.787 (0.738–0.828)	0.168	$p < 0.001$
	MAE	0.168 (0.136–0.198)	0.412 (0.356–0.466)	-0.244	$p < 0.001$
	Exact acc.	0.840 (0.806–0.868)	0.722 (0.684–0.756)	0.118	$p < 0.001$
	Within-1 acc.	0.992 (0.982–0.998)	0.902 (0.880–0.922)	0.090	$p < 0.001$

Table 3: TC-MAG vs human evaluators (baseline). TC-MAG outperforms human evaluators in all types of questions. Δ indicates the difference between TC-MAG and the baseline for a given metric; for MAE, a more negative value is better, for the rest, a higher value is better. n.s. indicates no significance difference is observed.

deserving full credit were marked lower and 60% true 0-mark answers were marked higher, indicating a potential grading rubric misalignment. The TC-MAG 's confusion matrix remains nearly tri-diagonal; most errors are one-step deviations and only 4% answers were mis-marked by more than one mark. Normalized confusion matrices for each task are shown in Figs. 4a, 4b, 4c, and 4d.

5.5 Findings: Comparison with LLM-based Baselines

We present the comparison with the LLM-based baselines in Table 4 and 5. Our results suggest that TC-MAG yields Δ QWK =+0.012 across ordinal grading tasks against GPTo3 using Chain of Thought (CoT) (see Table 5). TC-MAG framework with GPT-o3 has superior performance for ordinal grading, while the GPT-o3 CoT approach could suffice for binary grading (see Table 4). The apparent superiority of GPTo3 over GPT4o in our study reflects model quality rather than architectural effects [65, 68].

1-mark (binary) task. The CoT-prompt GPT-o3 baseline achieved the same κ as TC-MAG on this binary task (Table 4). All other baselines delivered κ marks 0.04–0.12 points lower than TC-MAG and were significantly inferior (adj. p < 0.05). **Applying the TC-MAG framework helped increase the accuracy of vanilla GPT-4o by 3.6%.**

Multi-mark tasks (2–4 marks aggregated). All baselines were significantly inferior to our system on the multi-mark tasks (Table 5). GPT-4o baselines performed considerably worse; both vanilla and CoT prompts produced QWK around 0.79–0.80 and high MAE (0̃.43–0.45), indicating frequent under-/over-scoring. **Interestingly, applying the TC-MAG framework increased exact accuracy by 4% for vanilla GPT-4o, and by 7% for vanilla GPT-o3.** Evaluation of 2-4 mark strata individually is in Appendix A.3.

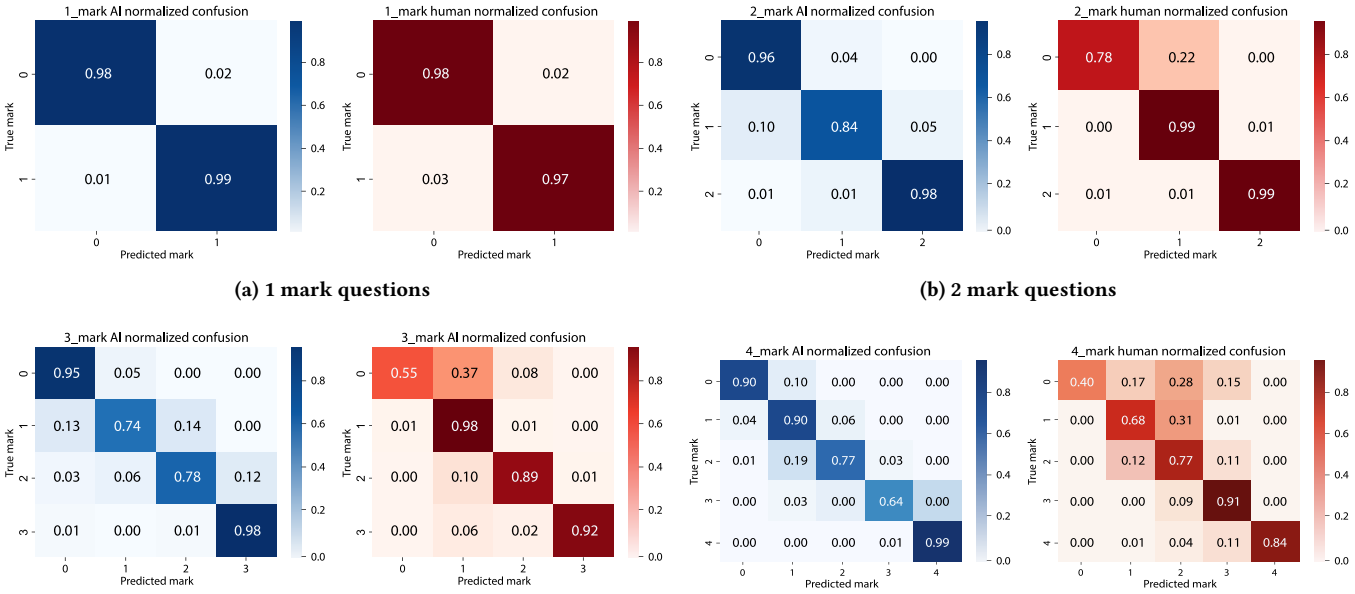
5.6 Error analysis of TC-MAG (TC-MAG GPT-o3)

In this section, we perform the error analysis. Two researchers independently coded every case where the TC-MAG ' mark differed from the gold, inspecting the **Marking Guideline Extractor**, both **Marker** agents, and the **Arbitrator** agent to locate the fault. The codes were reconciled with the discussion. We summarize the key findings from error analysis below,

(1) **Partial-credit rubric mismatch (multi-answer questions): 37%**
   (a) TC-MAG's Rubric creator is prompted to create a one-point-per-intermediate-step rubric, with the total number of rubric steps ≤ maximum marks. However, some questions have more intermediate steps than the maximum marks (especially in 3-4 mark questions); some demand a *single* final answer, and some mix sequential steps with independent sub-answers. Therefore, the Rubric creator fails to identify the essential, mark-carrying steps. A potential solution could be adding a Rubric-Type Classifier agent (single-answer / multi-answers / mixed-sequential) that conditions the Rubric creator agent to choose a matching rubric template based on the question's solution structure.

(2) **Marking guideline gaps: 26%**
   (a) Some topic-specific guidelines (e.g., allowable answer formats) are either absent or ambiguous in the document, so Marker agents improvise and diverge. This calls for ensuring that the source marking guideline documents (or any source documents) are pedantic enough to reduce ambiguity for LLMs.

(3) **Prompt drift (Markers): 21%**
   (a) Marker agents frequently ignore the 'no marks for unlabeled intermediate steps' prompt, and leniently award marks if



(a) 1 mark questions

(b) 2 mark questions

(c) 3 mark questions

(d) 4 mark questions

Figure 4: Normalized confusion matrices (1–4 marks). TC-MAG is near-diagonal across strata; humans show leniency on zero-credit answers and larger off-diagonal errors at higher marks.

Model type	Model	κ (95% CI)	Exact Acc.	MAE	MCC	p (adj.)	$\Delta\kappa$
Baseline	GPT-4o vanilla	0.848 (0.800 – 0.888)	0.924	0.076	0.854	$p < 0.001$	−0.120
	GPT-4o CoT	0.900 (0.860 – 0.932)	0.950	0.050	0.901	$p < 0.001$	−0.068
	GPT-4o TC-MAG	0.920 (0.884 – 0.952)	0.960	0.040	0.922	$p < 0.001$	−0.048
	GPT-o3 vanilla	0.960 (0.932 – 0.984)	0.980	0.020	0.960	0.500	−0.008
	GPT-o3 CoT	0.968 (0.944 – 0.988)	0.984	0.016	0.968	0.425	0.000
Proposed	TC-MAG (GPT-o3)	0.968 (0.944 – 0.988)	0.984	0.016	0.968	–	0.000

Table 4: 1-mark (binary) results. CoT-prompt GPT-o3 ties TC-MAG on κ; other baselines are significantly lower.

Model type	Model	QWK (95% CI)	Exact Acc.	Within-1 Acc.	MAE	p (adj.)	Δ QWK
Baseline	GPT-4o vanilla	0.788 (0.762–0.805)	0.643	0.926	0.449	$p < 0.001$	−0.148
	GPT-4o CoT	0.799 (0.780–0.819)	0.655	0.935	0.423	$p < 0.001$	−0.137
	GPT-4o TC-MAG	0.811 (0.786–0.832)	0.685	0.944	0.385	$p < 0.001$	−0.125
	GPT-o3 vanilla	0.918 (0.906–0.928)	0.809	0.981	0.211	$p < 0.001$	−0.018
	GPT-o3 CoT	0.924 (0.913–0.933)	0.813	0.989	0.199	$p < 0.001$	−0.012
Proposed	TC-MAG (GPT-o3)	0.936 (0.923–0.951)	0.878	0.993	0.129	–	0.000

Table 5: 2–4 marks (ordinal) results. All baselines are inferior.

the student wrote an intermediate step value without context/label. Based on the agent's produced explanations, we think this could be due to two reasons: (i) position-sensitivity in long prompts [56], and (ii) bias from appropriate marking strictness learned in the training dataset. arbitrator agent helps catch some errors, but does not guarantee detection, especially when both Markers share the same mistaken premise.

(4) **Prompt drift (Arbitrator): 3%**

(a) We prompted the arbitrator agent to cite **evidence-grounded** explanations [89] while analysing errors in preceding agents.

However, it occasionally misses rubric/marker prompt context and sides with the wrong explanation; consistent with known **LLM-as-judge biases** (position/verbosity) [108].

(5) **Prompt drift (Rubric generator): 4%**
   (a) We prompted the rubric agent to specify if units carry marks or not (essential/non-essential) for each step. Rubric generator made keyword errors, treating nouns as **units** (e.g., 'oranges', 'apples'). A solution could be to restrict the rubric generator to a list of keywords that can be classified as units (standard SI, unit keywords like "years old", etc).

In summary, we show that **for a given model, applying the TC-MAG framework produced substantial improvements over vanilla and CoT prompts** for ordinal grading (largest competing model: $\Delta k = +0.012$, $p < 0.001$). Teachers show the known accuracy decline as partial-credit complexity rises, and TC-MAG significantly outperforms human baseline on harder (3–4 mark) questions ($\Delta\kappa$=+0.046, p<.001). The next section (RQ2) will address how teachers perceive and trust these AI-generated marks, which is another critical aspect of deployment feasibility.

6 RQ2: Effectiveness of TC-MAG as a Trustworthy Autograder

In this section, we evaluate **RQ2,** i.e., teacher trust and willingness to delegate marking to TC-MAG. We study the baseline trust of teachers in AI-generated marks relative to their own, and how trust changes when micro-step-based explanations and confidence scores are displayed.

6.1 Variants

In this section, we examine how explanation granularity (Staged micro-steps vs Summarized) and confidence (High or Low) shape delegation and perceived trust.

- Staged (micro-step) explanation: Key outputs from agents in the TC-MAG framework: the marking guideline extracted, rubric criteria applied, the preferred marker's explanation, and a high/low confidence badge. The staged micro-steps are designed to align with teachers' mental models of the marking process and to support a sense of control [5, 22, 49].
- Summarized explanation: Summary with a focus on why TC-MAG deducted marks, with a high/low confidence badge to offer teachers a low-effort, quicker way to triage AI marking accuracy. Summarised versions of explanations were generated using GPT-o3 by inputting the staged explanation and prompting it to create a summarised version with all key marking decisions. All summarised explanations were verified by a senior teacher (15 years' experience).

We evaluate two directional hypotheses:

- H1 (*Delegation*): **Staged** (micro-step) explanations yield **higher** delegation than **Summarized**, because staged explanations increase perceived transparency/control (teachers see how grading criteria were applied) [50].
- H2 (*Calibration*): **Low** confidence reduces delegation relative to **High** confidence in both UIs; and the reduction is **larger** under Staged than Summarised, reflecting stronger reliance on the confidence cue [106].

6.1.1 Reasoning for these explanation-type variants. Prior work shows that higher predictive accuracy increases reported trust, and explanations can raise perceived transparency; therefore, we include explanations in all conditions and model accuracy as a covariate to isolate the effect of explanation design on delegation [25, 52, 74, 104], while communicating uncertainty with confidence tags [106].

6.2 Participants

We recruited **N=14** credentialed primary-school teachers from the exhaustive pool of our partner edtech organization's teaching staff employed as primary mathematics teachers via volunteer sampling (sample mean 12.1 years of teaching/marking experience, SD=4.6) across experience bands and qualifications. Highest qualifications (normalized): Bachelors n=11; Masters+ n=3. Prior exposure to generative AI ranged from never/rarely (n=4) to several times per week (n=4). This population represents the practitioners who would adopt an AI grader in our deployment context. We conducted a pre-task survey to understand our participants' baseline trust in AI and automation by adapting questions from the work of Papenmeier et al. [72]. Individual survey questions and responses can be found in Appendix B.1.1.

6.3 Study Design

We used a within-subjects design to compare two explanation forms crossed with confidence tags (as seen in Table 6). Descriptions of our 3-part user study (60 minutes) are provided next.

6.3.1 Part 1: Familiarization block (8 questions; 5–10 minutes). **Objective:** Let participants view both Staged and Summarized explanations to become familiar with the types of explanations and the user interface (UI) of the study. **The setup and user flow** can be seen in Fig. 5.

Questions: 8 primary-math short-answer questions (balanced across mark stratum, confidence, and AI accuracy; previously unseen questions; from the same bank used in RQ1), so that participants develop fair expectations of the system. The distribution of questions shown in part 1 can be found in Appendix B.2.1. To mitigate anchoring, we first elicited a human mark before revealing AI outputs [9, 52]. The gold label marks were revealed at the end of Part 1.

6.3.2 Part 2: Assessment of delegation under crossed factors (32 questions; 30 minutes). **Objective:** Measure **behavioral delegation** and **triage time** under **Explanation form × Confidence** with realized accuracy as a covariate.

Questions: 32 primary-math short-answer questions (balanced across mark stratum, confidence, and AI accuracy; previously unseen questions; from the same bank used in RQ1; questions different from Part 1). The distribution of questions shown in Part 2 can be found in Appendix B.3.1. **The setup and user flow** can be seen in Fig. 6. **Questions split between Staged and Summarized (with A/B sets)** are mentioned below,

- From the fixed 32 questions, we created two blocks of 16 randomised questions.
- Each block was time-bounded to 15 minutes, designed to induce real-world trade-offs [52].

Factor type	Factor	Description
Predictive (IV)	Explanation form	Staged (micro-step); summarized.
Predictive (IV)	Confidence tag	High; low.
Control (IV)	Accuracy info	Participants informed of model accuracy by mark stratum (pre-task)
Control (IV)	Realized accuracy of TC-MAG	Treated as a covariate (TC-MAG vs. gold label) and shown before the study.
Design	Structure	Within-subjects; 2×2 factorial design (Explanation $\times$ Confidence).

Table 6: Study factors and design summary.

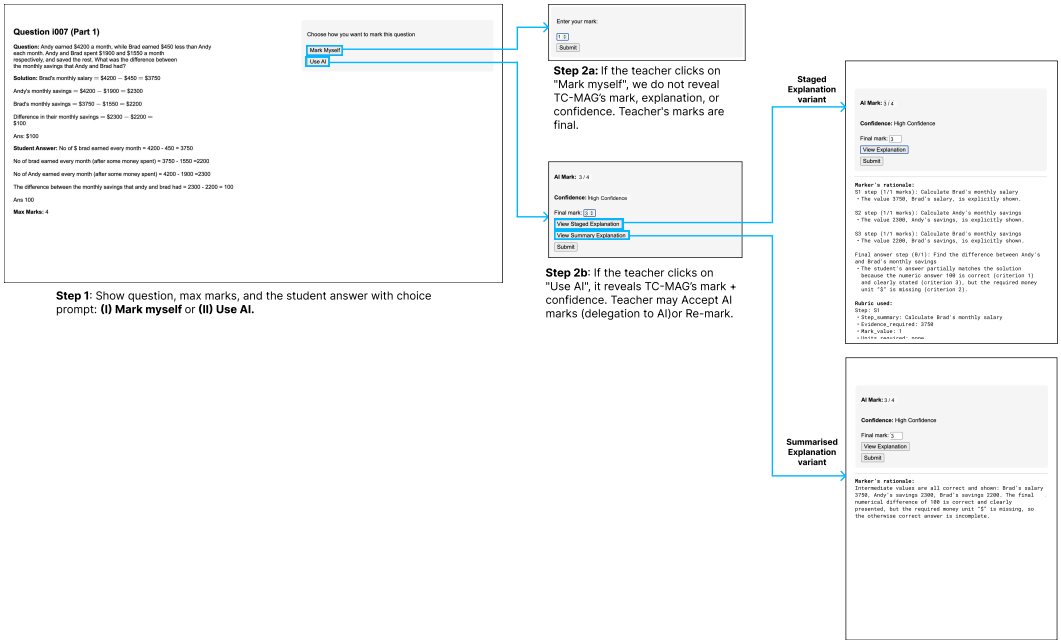


Figure 5: UI flow for Part 1 (familiarization). Participants can view both Staged (micro-step) and Summarized explanations for each question before proceeding.

- We used two variants, L1 and L2, to counterbalance the order of showing Staged vs Summarized explanations:
  - L1: Block 1→Staged, Block 2→Summarized
  - L2: Block 1→Summarized, Block 2→Staged
- Gold labels were revealed after both Block A and Block B ended, so that the AI mark accuracy doesn't affect behavior in the subsequent block [9, 52].
- Participants were asked to think aloud while marking each question. Think aloud protocol can be found in Appendix B.3.1.

6.3.3 *Part 3: Post-task surveys and interview (15-20 mins):*

- We collected self-reported trust, perceived predictability/understanding, willingness to delegate, and workload (NASA-TLX [31]). Questions can be found in Appendix B.4.2.
- We conducted a semi-structured interview about what made AI grades trustworthy/untrustworthy, how explanation granularity affected control, and the role of confidence tags [5, 72].

6.4 Performance Metrics

We use the following metrics to evaluate the effectiveness of staged explanations (aligned with teachers' mental models of the marking process) vs summarised explanations.

- **Delegation (binary)**: We define delegation as accepting AI mark without re-marking [72]. Non-delegation is defined as teachers marking themselves or re-marking, either after viewing AI's marks or without viewing them. (Behavioral trust proxy)
  - **Delegation rate** for different explanation types and confidence tags [9, 74].
  - **Selective delegation rate** to capture whether staged explanations improved the discrimination between correct/wrong AI marks (needed for safe autonomy) in case the participant delegated marking to AI [9].
    - **Sensitivity** (when teacher changed AI marks when AI was wrong), **Specificity** (when teacher kept AI marks when AI was right).

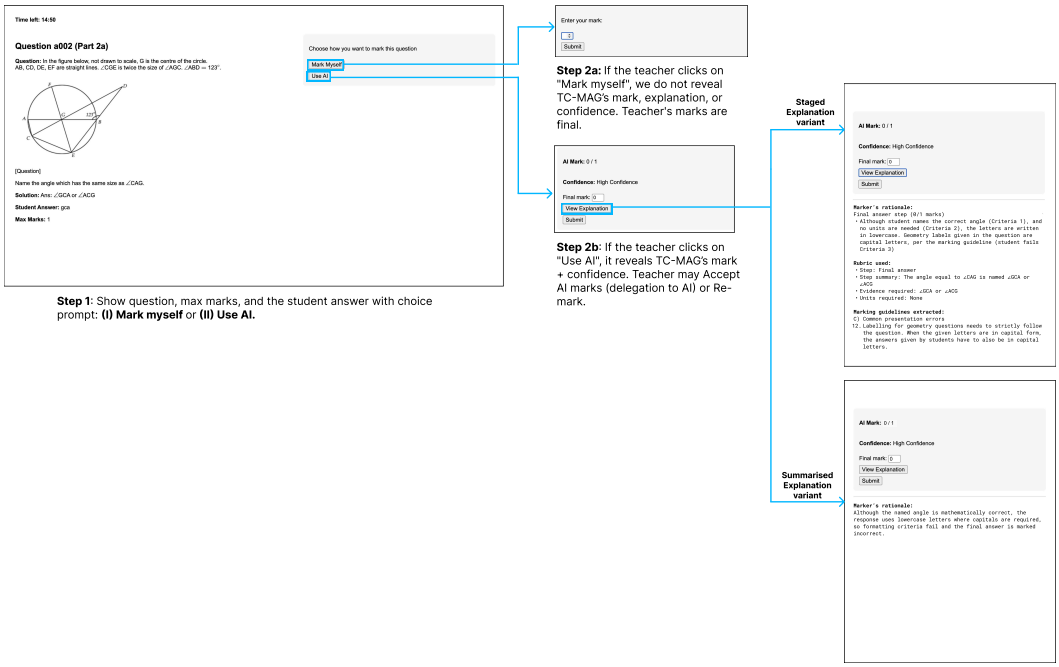


Figure 6: UI flows for Part 2 (teacher study). Staged vs. Summarised explanations crossed with Low vs. High confidence, showing the marking and explanation review sequence teachers followed.

  - **LR+ and LR-** to quantify how much a teacher's 'flip' or 'no-flip' of AI marks changes the odds that the AI mark is wrong or right respectively [3, 19].
- **Triage time (seconds)**: time from seeing a question to submitting marks (by accepting AI's or submitting own mark) as an efficiency proxy [9].
- **Workload (adapted NASA-TLX)**(from post-task survey): minimally adapted version of the original NASA-TLX questions to fit our grading context; survey done on a scale of 0-10. [30, 31, 47].
- **Post-task self-reported trust** (from post-task survey): perceived fairness, accuracy, understanding, predictability, overall trust, and willingness to delegate [72].

We present formulae used for each metric in Appendix B.5.

6.5 Results

6.5.1 Delegation rate. Teachers delegated more with **Staged** than **Summarized** at both confidence levels. The confidence badge provided an effective risk cue that teachers used to throttle reliance [104]. We present the delegation rate details in Table 7. Based on these findings, we evaluate our hypotheses:

- H1 (Delegation: Staged > Summarized) is **supported.** Staged elicited **higher delegation** at both confidence levels (0.84 vs 0.73 at High; 0.48 vs 0.38 at Low).
- H2 (Calibration: Low confidence reduces delegation; larger drop for Staged is **supported**, but by a small margin.

6.5.2 Selective delegation rate. We treat a teacher *flip* (overturning the AI's mark) as a test for the condition **AI mark is wrong**. We only look at the samples where the teacher decided to view AI marks.

Explanation type	Confidence	Delegation rate
Staged	High	**0.84**
	Low	0.48
Summarized	High	0.73
	Low	0.38

Table 7: Delegation rates by explanation type and confidence (teacher kept AI mark vs. changed or self-marked)

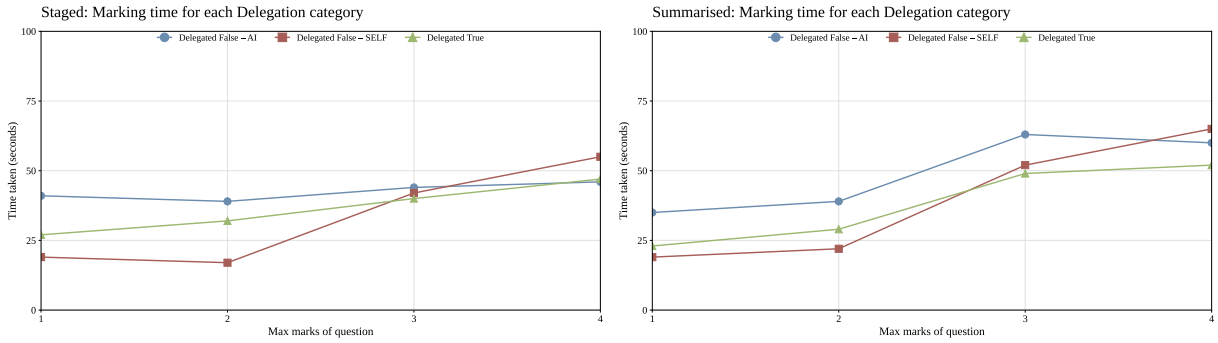
- **Condition positive** = AI mark **Incorrect** (relative to gold).
- **Test positive** = teacher **changed** the AI's mark ('flip').
- **Test negative** = teacher **did not change** the AI's mark.

Flip diagnostics (overturning TC-MAG's marks). We present the statistics of selective delegation from flip in Table 8. Teachers' flips are most selective under **Staged–High confidence** (Spec=95%, LR+≈11), indicating these flips are highly diagnostic, but the low Sensitivity under **Staged-High confidence** (57%) indicates potential overreliance. **Summarised-Low confidence** increases sensitivity (69%), but both **Summarised-Low confidence** and **Summarised-High confidence** substantially reduce specificity (87% and 64% , respectively), indicating over-correction.

LR+:how much a flip raises the odds the AI is wrong. Merged across confidence tags :

Explanation type	Confidence	Sensitivity (flip|wrong)	Specificity (keep|right)	LR+ of a flip	LR- of a flip
Staged	High	0.57	**0.95**	**12.14**	0.45
	Low	0.62	0.86	4.58	**0.40**
Summarized	High	0.64	0.87	4.97	0.41
	Low	**0.69**	0.64	1.94	0.48

Table 8: Selective-delegation diagnostics from flip-vs-correctness confusion matrices when teacher choose to see AI marks.



(a) Staged variant **(b) Summarised variant**

Figure 7: Triage time for staged and summarised explanation variants. 1. Delegated True: Participant submitted AI's mark without changing it; 3. Delegated False - Self: Participant marked themselves; 2. Delegated False - AI: Participant used AI to mark but, later changed marks.

- **Staged:** ≈ 11.50 →a flip is a **strong** error signal.
- **Summarized:** ≈ 4.60 →a flip is a **moderate** error signal.

6.5.3 Triage time (sec) by UI × Confidence. Now, we analyze the time taken by teachers to mark under different delegation scenarios (as seen in Fig. 7):

- Self-marking is fastest on 1–2 marks but slowest on 4-mark questions. This could also arise in case teachers hand-picked the simplest questions to mark themselves.
- Staged explanations on 3–4 mark questions keep the time moderate and below self-marking; it is a better default for higher-mark questions.
- Summaries are quickest on 1–2 marks (especially when accepting), but become costly on 3–4 marks - particularly when teachers open AI and then overturn (+15–20s vs Staged).

6.5.4 NASA-TLX (raw workload). Both UIs produced low to moderate workload in the NASA TLX composite with no meaningful difference (Staged **3.32** vs Summarized **3.27**, paired Δ=**+0.05**, 95% CI [0.85, 0.95], n.s.). Teachers reported **slightly more mental demand** with Staged (5.86) than Summarized (4.86) due to a larger amount of content reviewed with Staged. The **inverted Performance (i.e., lower perceived success) was higher for Summarised (3.29) vs Staged (1.93).** Question-wise survey analysis can be found in Appendix B.4.2.

6.5.5 Post-task Self-reported trust. Teachers **preferred Summarized explanations slightly higher** on *attitudinal* trust and understandability, even though behaviorally they **delegated more with Staged**. This attitude–behavior gap is commonly studied [72, 74]. Exact questions can be found in Appendix B.4.1.

6.6 Qualitative Findings

Two paper authors conducted a thematic analysis of post-task interview transcripts from all sessions and resolved disagreements through discussion. The quotes below are lightly paraphrased for clarity. We use themes to explain common behaviors and derive design implications.

(1) **Perceived marking effort dictates delegation**:
  (a) Verbatim: "I read the question and the student's answer to see if it's quick to grade (final answer is clearly right/wrong) or if it needs effort (final answer is wrong and the working is not exactly like the solution). Then I look at the confidence and decide."
    (i) High confidence + perceived low-effort: "I just accept the AI; it's faster."
    (ii) Low confidence + perceived low-effort: "I'll mark it myself, it's faster than reading an explanation."
    (iii) High/low confidence + perceived high-effort "It's basically a robot marking based on the given solution, so I don't expect it to correctly mark answers that are different from the solution, I'd rather mark this case myself."
  (b) Implication: This aligns with progressive disclosure and appropriate reliance: teachers scale review depth with perceived risk [5, 104].

(2) **Micro-steps serve accountability:**
  (a) Verbatim: "If a parent appeals, I can point exactly where the mark came from when AI showed that it followed our rubric. Otherwise, I will need to mark again, then what's the point?"
  (b) Implication: Staged, rubric-aligned steps function as an audit trail that can help to justify decisions later, and are not just useful while marking.

(3) **Summarised vs Staged explanations depend on question familiarity**:
   (a) Verbatim: "On legacy questions, a summary is enough because I am familiar with all sorts of student answers on those questions by now. But on new question types, I want to see the steps."
   (b) Implication: Staged explanations were very useful in understanding how to grade new questions, and possibly training new teachers for marking tasks.
(4) **Why teachers overdelegate with Staged and overcorrect Summarised**:
   (a) Verbatim (Staged): "AI is using the correct rubric and guideline here, so it seems like its mark should be correct."
   (b) Verbatim (Summarised): "It's not clear why it awarded these marks, so I need to check the (student's) answer steps."
   (c) Implication: This signals an over-reliance on staged explanations and calls for taking appropriate measures, e.g., regular training/testing sessions that require teachers to assess a mix of correctly and incorrectly AI-marked cases.
(5) **Confidence is a cue, but fragile**:
   (a) Verbatim (Low confidence): "Low confidence is the AI's red flag, so I will mark myself because the AI made a mistake."
   (b) Verbatim (High confidence): "If your AI isn't 100% accurate, how is its confidence accurate?"
   (c) Implications: Confidence modulates reliance, but teachers want evidence of near-perfect confidence calibration; as miscalibration can erode trust more than silence [104].
(6) **Marking guidelines are a helpful anchor**:
   (a) Verbatim: "I misremembered the unit rule because the primary math guidelines have changed recently, and this guideline helped me." "It's useful when it shows what value of π to use for this question, otherwise I would have to look for it in the marking guidelines document."
   (b) Implication: Extracted guideline snippets help align to local policies (units, rounding, method marks) and reduce policy-driven disagreements, and are helpful for all questions.

7 Discussion

Our work demonstrates that a multi-agent LLM framework grounded in teacher cognition can exceed human expert performance in grading complex, multi-step problems. We now discuss the theoretical and practical implications of this finding, focusing on how such systems can reshape collaborative grading workflows, enhance educational fairness, and the critical considerations for their responsible deployment.

7.1 Design Implications for Real-world Deployment

The following properties of TC-MAG collectively facilitate real-world deployment. These findings correlate well with the teachers' expectations and, therefore, allow a smooth ride to adopt the framework for a real-world setting.

(1) **Confidence-adaptive progressive disclosure**
   (a) **High confidence** default to Staged micro-steps to preserve correct marks and reduce unnecessary reversals (harmful flips lower with Staged vs Summarised: 0.14 vs 0.24), matching teachers' accountability need ("I need to defend the mark").
   (b) **Low confidence** use Summarized explanations to surface risk quickly (beneficial flips **66.7%**)(Low confidence → "I'd re-mark it myself") and progressive disclosure (with one-click expansion to Staged) to let them view extracted marking guidelines if needed ('marking guideline is helpful').
(2) **Lead with the decisive criterion, grounded in evidence**
   (a) Start explanations with rubric criteria-embedded evidence that moved the grade (e.g., 'Step 1: unit missing: -1') to cut triage times, then offer collapsible detail for in-depth evaluation [49, 51].
(3) **Flip-aware review and learning loops**
   (a) Because **LR+ of a flip is high**, capture a brief reason when teachers overturn AI. This is a rich signal for prompt refinement [5].
(4) **Guardrails against harmful flips at High confidence** When overturning a High-confidence mark, consider subtle 'are you sure?' nudges to reduce over-correction without blocking teacher autonomy [5, 10].

7.2 Superiority of TC-MAG over Human Baselines and its Implications in Human-AI Collaboration

7.2.1 Reasoning for Superior Performance. Our error analysis of TC-MAG and behavioural insights from **RQ2** suggest that TC-MAG outperforms experienced teachers on complex questions by systematically mitigating the cognitive biases inherent in human evaluation. When faced with multi-step, partial-credit problems that overload working memory, humans may resort to heuristics and are susceptible to biases like the Halo Effect or Confirmation Bias [43]. While our cognitive task analysis described an ideal grading process, teachers in practice may inadvertently deviate by over-relying on memory, misapplying guidelines, or overlooking details in lengthy answers. TC-MAG framework operationalizes an ideal grading process by deconstructing it into discrete, auditable micro-steps: guideline extraction, rubric generation, criterion-level marking by independent agents with justifications, and discrepancy arbitration. Via modularization of each step (executed by a separate LLM agent), TC-MAG avoids shortcuts and ensures every component of a student's answer is evaluated against explicit criteria.

7.2.2 Implications for Human-AI Collaborative Grading Workflows. Our results position TC-MAG as a **grading collaborator** for teachers for automating the repetitive aspects of grading. Teachers can delegate high-confidence, simple-answer questions to the AI and focus their expert attention on low-confidence cases or questions with higher mark values. This emergent behavior was observed in our **RQ2** study, where teachers under time pressure chose to offload grading for simple answers where the AI's confidence was high.

7.3 Broader Impact on Educational Practice

7.3.1 Enhancing Grading Consistency and Fairness. A key implication of TC-MAG is its potential to improve **grading consistency and fairness** at scale. By anchoring every marking decision to formal guidelines and rubric criteria, TC-MAG forces a collaborating teacher to re-trace the ideal grading logic, and therefore reduces the idiosyncratic grading practices that can vary between teachers. As participants in our RQ2 study noted, the explicit guideline snippets served as valuable reminders of official policy. This standardization is crucial for high-stakes contexts like national or district-wide assessments, contributing to greater equity by ensuring students are evaluated against the same standard, regardless of the assigned grader.

7.3.2 Supporting Student Learning and Agency. In our current design, students do not directly interact with TC-MAG, and teachers mediate results and any feedback. Ethically, it is vital to disclose the AI's role in grading. Providing a clear explanation (even if simplified) for why marks were deducted could help students accept and learn from their grades. We recommend that any classroom deployment include a channel for students to challenge grades, in accordance with the principles of contestability and explanation in algorithmic systems [1].

7.3.3 Managing Stakeholder Expectations and Algorithmic Aversion. Deploying a system like TC-MAG, even with its superior accuracy (than our human teacher participants), must confront a deep-seated human bias: **algorithmic aversion** [21]. An AI's mistake is often perceived as a systemic failure, whereas a similar error from a human teacher is a forgivable lapse. Bridging this perceptual gap requires communication strategies that frame the AI not as an infallible judge, but as a powerful, yet, probabilistic partner in the grading process. The conversation with parents, students, and teachers should center on the new possibilities AI grading unlocks: the chance for more frequent assessments, rapid feedback that accelerates learning, and more time for teachers to devote to personalized instruction. By empowering all stakeholders to contest the AI's decisions against solution keys and TC-MAG's explanations, we transform them from passive recipients into active participants in the system's ongoing refinement, building trust through transparency [79].

7.4 Towards Responsible and Equitable Deployment

7.4.1 The Imperative of Human-Led Audits for Fairness. TC-MAG will faithfully reflect any biases embedded in its source materials. Outdated or skewed prompts, rubrics, or guidelines will lead to systematically flawed grading. However, TC-MAG 's modularity facilitates audits because each mark is tied to a visible criterion, making it easier to spot and rectify systemic errors, such as unjustly penalizing an unconventional but correct problem-solving method. For responsible deployment, we recommend a 'human-in-the-loop' governance model where educators periodically review and refine both the source guidelines and the AI-generated rubrics - a critical step, given that incorrect rubrics were the primary source of TC-MAG's errors (37%) in our RQ1 analysis.

7.4.2 Mitigating Automation Bias and De-skilling. As AI graders approach and exceed human accuracy, the risk of **automation bias** and teacher de-skilling becomes salient [85]. Our RQ2 findings hinted at this, with some teachers occasionally over-delegating to the AI when presented with staged explanations. To counteract this, we advocate for interfaces and training protocols that promote active human engagement. For instance, institutions could implement calibration exercises where teachers must identify and correct AI grading errors, analogous to pilots training to handle autopilot failures. Such practices ensure that the teacher remains a vigilant and skilled supervisor in the collaborative loop.

7.4.3 Addressing Practical Barriers: Cost, Connectivity, and Compute. Our multi-agent approach, especially with powerful foundation models, is computationally expensive and requires reliable computer hardware and internet, posing significant barriers for **low-resource contexts**. We have provided token counts, API call costs per question (for both real-time processing and batch processing), and processing times per question for all the LLM baselines used in our RQ1 study.

Our results indicate that multiple models (GPT-4o and GPT-o3) benefited from the TC-MAG framework (yielding improvements over vanilla and CoT prompting), suggesting a potential to improve AI grading accuracy using TC-MAG framework even while using cheaper or offline models. However, we recommend using inferior inference models with great caution, as the accuracy drop can prohibit delegation of marking to AI (as seen in our RQ1 study 5.5, the macro κ drop from GPT-o3 to GPT-4o on our dataset is from 0.936 to 0.811). Institutions deploying human-AI collaborative grading should ensure teachers' capacity to address students' grade review requests, especially in low-resource settings [32].

7.4.4 Design choices to enable wider school-led deployment. We focused our baselines on closed-weight LLM variants (no fine-tuning) so that schools that may lack labeled data, governance capacity, or compute/engineering resources needed to fine-tune safely and maintainably, still be able to adopt AI-based grading [20, 36, 86]. However, fine-tuning can improve instruction-following and task adherence [71, 94]. Some single-agent errors, as identified in our motivational study, stem from non-adherence to prompts, and fine-tuning could reduce such "prompt drifts". If successful, it would provide a lower-latency alternative to a multi-agent LLM framework [41, 42]. However, fine-tuning is not guaranteed to improve performance, can introduce regressions or forgetting, and requires a robust, large-scale dataset to ensure generalization [46, 48].

7.4.5 Generalizability beyond Mathematics and Singapore Curriculum. Our findings generalize most directly to assessments with: (i) explicit marking guidelines, (ii) decomposable scoring criteria, and (iii) verifiable correct answers. Extension to holistic assessments (e.g., creative writing, argumentation essays) requires additional validation. When extending our TC-MAG framework beyond mathematics or beyond the Singapore curriculum, we expect accuracy-enhancing and confidence-gating design patterns to remain invariant (multiple passes of marker agents, arbitration, confidence estimator). However, the design patterns based on micro-decisions uncovered in the CTA Section 3.1 (marking guideline extraction,

rubric generator, marking based on criteria) may vary across curricula and subjects. Unlike the U.S. Common Core, which emphasizes multiple solution strategies equally, Singapore's curriculum privileges the Model Method for specific problem types [44, 45, 60, 61]. Our rubric generator is prompted to adhere to this preference and may require adaptation for curricula that emphasize solution diversity. In domains where rubrics cannot fully specify content quality (e.g., interpretive essays), TC-MAG would shift from step-wise solution decomposition to criterion-level evidence, exemplar-anchored moderation, and stronger uncertainty signaling.

7.4.6 Broader Justice and Unintended Negative Consequences. Prior work on automated scoring shows that high inter-rater reliability can coexist with systematic score differences across demographic groups [77, 101]. These disparities can arise when prompts, rubrics, and reference solutions operationalize local norms about what counts as "good reasoning," and those abstractions are treated as universal, thereby disadvantaging legitimate alternative solution paths [80]. In response, deployments should (i) report disaggregated performance across relevant learner groups and contexts, (ii) document evaluation conditions, and (iii) preserve contestability through clear appeal and override procedures aligned with established testing standards and education governance guidance [4]. A further unintended consequence is pedagogical: if AI systems assess most student responses, teachers may lose opportunities to notice misconceptions and personalize instruction.

7.5 Limitations and Future Work

We note the limitations of the proposed framework and the possible future works. First, the framework is evaluated on Singapore's primary mathematics curriculum, which we aim to generalize in terms of subjects, educational systems in our future work. Second, we carried out analysis involving experienced teachers. However, the findings may differ for novices or teachers with varying technological fluency. Third, our analysis is based on typed responses, whereas classrooms often rely on handwritten work, diagrams, and equations. A full *capture–convert–grade* pipeline can make the framework more generalized. Finally, error may stem from imperfect AI-generated rubrics and limited context in marking guidelines that we aim to address in our future work. Finally, our evaluation uses commercial OpenAI models (GPT-4o and GPT-o3) [65, 68], and model behavior can change due to provider updates, even for the same model name, complicating exact long-term reproducibility [12, 13, 70]. We partially mitigate this by reporting inference settings and validating that the TC-MAG framework improves performance across two distinct models.

8 Conclusion

We introduce TC-MAG, a multi-agent LLM framework that utilizes expert teacher cognition to improve accuracy and foster more effective teacher-AI collaboration in AI grading. The agents embedded in the framework perform distinct tasks (e.g., guideline extraction, rubric creation, or arbitration) to have a deployment ready trustworthy autograder. Our work provides two key contributions: First, we demonstrated that our teacher-cognition approach achieves greater reliability on complex, partial-credit questions (quadratic-weighted kappa (κ)=0.936), significantly outperforming both human markers ($\Delta\kappa$ = +0.046) and strong LLM baselines. Second, we established that UI affordances are critical for earning teacher trust and fostering effective collaboration. Specifically, with staged step-by-step explanations, teachers maintained higher precision in accepting AI grades (LR+ 11.5 vs 4.60 with summarised explanations). By functioning as a grading collaborator, TC-MAG can improve grading consistency and fairness at scale, freeing educators to focus on more nuanced pedagogical tasks. Future work should adapt this framework to other subjects, accommodate handwritten student answers, and conduct longitudinal studies to assess long-term impact on teacher workload and student feedback cycles.

9 Ethics and Responsibility Statement

The research protocol, including the use of student data, received written approval from the partner edtech organization's equivalent institutional review board. TC-MAG's full prompts and datasets remain proprietary to our partner organization. All student response data were deidentified before analysis (i.e., no names, school affiliations, or other Personally Identifiable Information). Written consent was obtained from all research participants, who were briefed on the purpose, data usage, their right to withdraw and anonymity safeguards of the study, were salaried employees of the partner organization, and participated as part of their regular duties without additional compensation, and were assured their responses would not affect their employment and were encouraged to provide critical feedback. We protect data privacy by using OpenAI's GPT-4o and GPT-o3 through the API in data opt-out settings (no student data was retained or used for model training, according to OpenAI's enterprise privacy policy). Two co-authors are employees of the partner edtech organization that provided the dataset and may have a commercial interest in automated grading. A third co-author is an independent academic collaborator who contributed only to research design, methodological oversight, and analysis of de-identified data. Several design choices were implemented to mitigate potential conflicts, including the use of pre-existing marking protocols with blind adjudication by a senior teacher and pre-planned analyses aligned with stated research questions and hypotheses. Additionally, results for all tested conditions are reported, and the partner organization had no veto power over publication decisions.

10 GenAI Usage Disclosure

We used OpenAI's GPT-5 [67] to refine portions of the text (the abstract and related works) and Anthropic's Claude Sonnet 4 [7] to generate Python code for the TC-MAG agent workflows and baseline models. We also used ChatGPT Codex [66] to generate code for the prototype used in our RQ2 user study. We used OpenAI's GPT-5 for the initial draft of the design elements of Fig. 1.

Acknowledgments

This research was funded and supported by Geniebook Pte. Ltd., Singapore. We are grateful to the co-founders, Neo Zhizhong (CEO) and Alicia Cheong (COO), for their vision in advancing AI-assisted education and for providing the resources, infrastructure, and access to proprietary datasets that made this work possible. We are grateful to the 14 primary school mathematics teachers who participated

in our user study, and to the teachers who created and verified the gold-standard labels used in our evaluation. Surjya Ghosh would also like to acknowledge the unrestricted research gift received from Google.

References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3173574.3174156

[2] Rudaiba Adnin, Atharva Pandkar, Bingsheng Yao, Dakuo Wang, and Maitraye Das. 2025. Examining Student and Teacher Perspectives on Undisclosed Use of Generative AI in Academic Work. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1071, 17 pages. doi:10.1145/3706598.3713393

[3] Anthony K. Akobeng. 2007. Understanding diagnostic tests 3: Likelihood ratios, pre- and post-test probabilities and their use in clinical practice. *Acta Paediatrica* 96, 4 (2007), 487–491. doi:10.1111/j.1651-2227.2006.00179.x

[4] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf Open-access PDF.

[5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300233

[6] Maryam Amirizaniani, Jihan Yao, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tanya Roosta, and Chirag Shah. 2024. LLMAuditor: A Framework for Auditing Large Language Models Using Human-in-the-Loop. *arXiv preprint arXiv:2402.09346* (2024).

[7] Anthropic. 2025. System Card: Claude Opus 4 & Claude Sonnet 4. System card. https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf Accessed 2025-10-10.

[8] Muniza Askari. 2025. Reliable but supervised: evaluating a generative AI-rubric model for consistent and fair assessment in postgraduate education. *Assessment & Evaluation in Higher Education* (2025), 1–18.

[9] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. doi:10.1145/3411764.3445717

[10] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287

[11] Imran Chamieh, Torsten Zesch, and Klaus Giebermann. 2024. LLMs in Short Answer Scoring: Limitations and Promise of Zero-Shot and Few-Shot Approaches. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 309–315. https://aclanthology.org/2024.bea-1.25/

[12] Lingjiao Chen, Tracy Cai, Matei Zaharia, and James Zou. 2021. Did the Model Change? Efficiently Assessing Machine Learning API Shifts. *arXiv preprint arXiv:2107.14203* (2021). doi:10.48550/arXiv.2107.14203

[13] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009* (2023). doi:10.48550/arXiv.2307.09009

[14] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *Transactions on Machine Learning Research* (2023).

[15] Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-yi Lee. 2024. Large Language Model as an Assignment Evaluator: Insights, Feedback, and Challenges in a 1000+ Student Course. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 2489–2513. doi:10.18653/v1/2024.emnlp-main.146

[16] Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21 (01 2020). doi:10.1186/s12864-019-6413-7

[17] Yucheng Chu, Hang Li, Kaiqi Yang, Harry Shomer, Yasemin Copur-Gencturk, Leonora Kaldaras, Kevin Haudek, Joseph Krajcik, Namsoo Shin, Hui Liu, and Jiliang Tang. 2025. A LLM-Powered Automatic Grading Framework with Human-Level Guidelines Optimization. In *Proceedings of the 18th International Conference on Educational Data Mining*, Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (Eds.). International Educational Data Mining Society, 31–41. doi:10.5281/zenodo.15870201

[18] Yucheng Chu, Hang Li, Kaiqi Yang, Harry Shomer, Hui Liu, Yasemin Copur-Gencturk, and Jiliang Tang. 2024. A LLM-powered automatic grading framework with human-level guidelines optimization. *arXiv preprint arXiv:2410.02165* (2024).

[19] Jonathan J. Deeks and Douglas G. Altman. 2004. Diagnostic tests 4: Likelihood ratios. *BMJ* 329, 7458 (2004), 168–169. doi:10.1136/bmj.329.7458.168

[20] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html

[21] Berkeley Dietvorst, Joseph Simmons, and Cade Massey. 2014. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General* 144 (11 2014), 114–126. doi:10.1037/xge0000033

[22] Hyo Jin Do, Michelle Brachman, Casey Dugan, Qian Pan, Priyanshu Rai, James M. Johnson, and Roshni Thawani. 2024. Evaluating What Others Say: The Effect of Accuracy Assessment in Shaping Mental Models of AI Systems. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 373 (Nov. 2024), 26 pages. doi:10.1145/3686912

[23] Afrizal Doewes, Nughthoh Kurdhi, and Akrati Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *16th International Conference on Educational Data Mining, EDM 2023*. International Educational Data Mining Society (IEDMS), 103–113.

[24] Afrizal Doewes, Nughthoh Arfawi Kurdhi, and Akrati Saxena. 2023. Evaluating Quadratic Weighted Kappa as the Standard Performance Metric for Automated Essay Scoring. In *Proceedings of the 16th International Conference on Educational Data Mining*, Mingyu Feng, Tanja KÃ¤ser, and Partha Talukdar (Eds.). International Educational Data Mining Society, Bengaluru, India, 103–113. doi:10.5281/zenodo.8115784

[25] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 263–274. doi:10.1145/3301275.3302316

[26] Rafael Ferreira Mello, Cleon Pereira Junior, Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Newarney Costa, Geber Ramalho, and Dragan Gasevic. 2025. Automatic Short Answer Grading in the LLM Era: Does GPT-4 with Prompt Engineering beat Traditional Models?. In *Proceedings of the 15th international learning analytics and knowledge conference*. 93–103.

[27] Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 420, 10 pages.

[28] Skyler Grandel, Douglas C Schmidt, and Kevin Leach. 2024. Applying large language models to enhance the assessment of parallel functional programming assignments. In *Proceedings of the 1st International Workshop on Large Language Models for Code*. 102–110.

[29] Christian Grévisse. 2024. LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education* 24, 1 (2024), 1060.

[30] Rebecca Grier. 2015. How high is high? A metanalysis of NASA TLX global workload scores. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 59. doi:10.1177/1541931215591373

[31] Sandra Hart. 2006. Nasa-task load index (Nasa-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. doi:10.1177/154193120605000909

[32] Wayne Holmes, Fengchun Miao, et al. 2023. *Guidance for generative AI in education and research*. Unesco Publishing.

[33] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1904.09751

[34] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=VtmBAGCN7o

[35] Silas Hsu, Tiffany Wenting Li, Zhilin Zhang, Max Fowler, Craig Zilles, and Karrie Karahalios. 2021. Attitudes Surrounding an Imperfect AI Autograder. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 681, 15 pages. doi:10.1145/3411764.3445424

[36] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=nZeVKeeFYf9

[37] Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. 2024. Uncertainty in Language Models: Assessment through Rank-Calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 284–312. doi:10.18653/v1/2024.emnlp-main.18

[38] Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024. Calibrating Long-form Generations From Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 13441–13460. doi:10.18653/v1/2024.findings-emnlp.785

[39] Hugging Face. 2025. Transformers: Generation strategies. https://huggingface.co/docs/transformers/en/generation_strategies. Accessed 9 Oct 2025.

[40] Lan Jiang and Nigel Bosch. 2024. Short answer scoring with GPT-4. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 438–442. doi:10.1145/3657604.3664685

[41] Lucas Joos, Daniel A. Keim, and Maximilian T. Fischer. 2025. Cutting Through the Clutter: The Potential of LLMs for Efficient Filtration in Systematic Literature Reviews. In *EuroVis Workshop on Visual Analytics (EuroVA)*, Hans-Jörg Schulz and Anna Villanova (Eds.). The Eurographics Association. doi:10.2312/eurova.20251105

[42] Lucas Joos, Daniel A. Keim, and Maximilian T. Fischer. 2025. Leveraging LLMs for Semi-Automatic Corpus Filtration in Systematic Literature Reviews. doi:10.48550/arXiv.2510.11409 arXiv:2510.11409 [cs.LG] arXiv preprint.

[43] D. Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux. https://books.google.com.sg/books?id=SHvzzuCnuv8C

[44] Berinderjeet Kaur. 2004. Teaching of Mathematics in Singapore Schools. In *Proceedings of ICME-10*. https://home.sandiego.edu/~pmyers/singapore/ICME-10-RG-paper-patsy.pdf

[45] Berinderjeet Kaur. 2014. Mathematics Education in Singapore—An Insider's Perspective. *IndoMS-JME* 5, 1 (2014), 1–16. https://files.eric.ed.gov/fulltext/EJ1079596.pdf

[46] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. 5637–5664. https://proceedings.mlr.press/v139/koh21a.html

[47] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. 2023. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Comput. Surv.* 55, 13s, Article 283 (July 2023), 39 pages. doi:10.1145/3582272

[48] Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2024. Understanding Catastrophic Forgetting in Language Models via Implicit Inference. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=VrHiF2hsrm

[49] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 126–137. doi:10.1145/2678025.2701399

[50] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/2207676.2207678

[51] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models. In *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC*. doi:10.1109/VLHCC.2013.6645235

[52] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. doi:10.1145/3287560.3287590

[53] J Richard Landis, Tonya S King, Jai W Choi, Vernon M Chinchilli, and Gary G Koch. 2011. Measures of agreement and concordance with clinical research applications. *Statistics in Biopharmaceutical Research* 3, 2 (2011), 185–209.

[54] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Scale Language Model Society. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://proceedings.neurips.cc/paper_files/paper/2023/file/a3621ee907def47c1b952ade25c67698-Paper-Conference.pdf

[55] Hang Li, Yucheng Chu, Kaiqi Yang, Yasemin Copur-Gencturk, and Jiliang Tang. 2025. LLM-based Automated Grading with Human-in-the-Loop. *arXiv preprint arXiv:2504.05239* (2025).

[56] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172 [cs.CL] https://arxiv.org/abs/2307.03172

[57] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173. doi:10.1162/tacl_a_00638

[58] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 384, 23 pages. doi:10.1145/3491102.3501825

[59] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 46534–46594. https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf

[60] Ministry of Education, Singapore. 2021. Mathematics Syllabus: Primary 1–6. https://www.moe.gov.sg/-/media/files/primary/mathematics_syllabus_primary_1_to_6.pdf Official syllabus; framework centers on mathematical problem solving.

[61] Ministry of Education, Singapore. 2025. Primary school subjects and syllabuses. https://www.moe.gov.sg/primary/curriculum/syllabus Last updated 14 Oct 2025.

[62] Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (Eds.). Association for Computational Linguistics, Portland, Oregon, USA, 752–762. https://aclanthology.org/P11-1076/

[63] Jakob Nielsen and Thomas K. Landauer. 1993. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 206–213. doi:10.1145/169059.169166

[64] OpenAI. 2023. How to make your completions outputs consistent with the new seed parameter. https://cookbook.openai.com/examples/reproducible_outputs_with_the_seed_parameter. Accessed: 2026-01-06.

[65] OpenAI. 2024. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/ Accessed: 2025-10-09.

[66] OpenAI. 2025. Addendum to GPT-5 System Card: GPT-5-Codex. OpenAI system card addendum. https://cdn.openai.com/pdf/97cc5669-7a25-4e63-b15f-5fd5bdc4d149/gpt-5-codex-system-card.pdf Accessed 2025-10-10.

[67] OpenAI. 2025. GPT-5 System Card. OpenAI system card. https://cdn.openai.com/gpt-5-system-card.pdf Accessed 2025-10-10.

[68] OpenAI. 2025. OpenAI o3 and o4-mini System Card. https://openai.com/index/o3-o4-mini-system-card/ Accessed: 2025-10-09.

[69] OpenAI. 2026. Advanced usage — Reproducible outputs. https://platform.openai.com/docs/guides/advanced-usage. Accessed: 2026-01-06.

[70] OpenAI. 2026. Deprecations. OpenAI API documentation. https://platform.openai.com/docs/deprecations Accessed: 2026-01-06.

[71] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[72] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 35 (March 2022), 33 pages. doi:10.1145/3495013

[73] Aditya Pathak, Rachit Gandhi, Vaibhav Uttam, Arnav Ramamoorthy, Pratyush Ghosh, Aaryan Raj Jindal, Shreyash Verma, Aditya Mittal, Aashna Ased, Chirag Khatri, et al. 2025. 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*. 181–195.

[74] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. doi:10.1145/3411764.3445315

[75] David M. W. Powers. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv:2010.16061 [cs.LG] https://arxiv.org/abs/2010.16061

[76] Vatsal Raina, Adian Liusie, and Mark Gales. 2023. Assessing Distractors in Multiple-Choice Tests. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, Daniel Deutsch, Rotem Dror, Steffen Eger, Yang Gao, Christoph Leiter, Juri Opitz, and Andreas Rücklé (Eds.). Association for Computational Linguistics, Bali, Indonesia, 12–22. doi:10.18653/v1/2023.eval4nlp-1.2

[77] Chaitanya Ramineni and David M. Williamson. 2018. Understanding Mean Score Differences Between the e-rater® Automated Scoring Engine and Humans for Demographically Based Groups in the GRE® General Test. *ETS Research Report Series* 2018, 1 (2018), 1–31. doi:10.1002/ets2.12192

[78] Philipp Reinhard, Mahei Manhai Li, Matteo Fina, and Jan Marco Leimeister. 2025. Fact or Fiction? Exploring Explanations to Identify Factual Confabulations in RAG-Based LLM Systems. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 274, 13 pages. doi:10.1145/3706599.3720249

[79] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. doi:10.1145/2939672.2939778

[80] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. doi:10.1145/3287560.3287598

[81] Chamuditha Senanayake and Dinesh Asanka. 2024. Rubric based automated short answer scoring using large language models (LLMs). In *2024 international research conference on smart computing and systems engineering (SCSE)*, Vol. 7. IEEE, 1–6.

[82] Burr Settles and Brendan Meeder. 2016. A Trainable Spaced Repetition Model for Language Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Katrin Erk and Noah A. Smith (Eds.). Association for Computational Linguistics, Berlin, Germany, 1848–1858. doi:10.18653/v1/P16-1174

[83] Lei Sha and Thomas Lukasiewicz. 2024. Text Attribute Control via Closed-Loop Disentanglement. *Transactions of the Association for Computational Linguistics* 12 (2024), 190–209. doi:10.1162/tacl_a_00640

[84] Yang Shi, Tian Gao, Xiaohan Jiao, and Nan Cao. 2023. Understanding Design Collaboration Between Designers and Artificial Intelligence: A Systematic Literature Review. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 368 (Oct. 2023), 35 pages. doi:10.1145/3610217

[85] Linda Skitka, Kathleen Mosier, and MARK BURDICK. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51 (11 1999), 991–1006. doi:10.1006/ijhc.1999.0252

[86] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3645–3650. doi:10.18653/v1/P19-1355

[87] Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. 2019. Improving short answer grading using transformer-based pre-training. In *International Conference on Artificial Intelligence in Education*. Springer, 469–481.

[88] Texas Education Agency. 2024. Texas STAAR RLA Spring 2024 Administration: Automated Scoring Methods and Results. https://tea.texas.gov/student-assessment/reports-and-studies/2024-staar-hybrid-scoring-study.pdf

[89] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. arXiv:2305.04388 [cs.CL] https://arxiv.org/abs/2305.04388

[90] Rama Adithya Varanasi, Nicola Dell, and Aditya Vashistha. 2024. Saharaline: A Collective Social Support Intervention for Teachers in Low-Income Indian Schools. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.

[91] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171 [cs.CL] https://arxiv.org/abs/2203.11171

[92] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=1PL1NIMMrw

[93] Yun Wang, Zhaojun Ding, Xuansheng Wu, Siyue Sun, Ninghao Liu, and Xiaoming Zhai. 2025. AutoSCORE: Enhancing Automated Scoring with Multi-Agent Large Language Models via Structured Component Recognition. *arXiv preprint arXiv:2509.21910* (2025).

[94] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=gEZrGCozdqR

[95] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=_VjQlMeSB_J

[96] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *CoRR* abs/2201.11903 (2022). arXiv:2201.11903 https://arxiv.org/abs/2201.11903

[97] David M Williamson, Xiaoming Xi, and F Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice* 31, 1 (2012), 2–13.

[98] C. Willmott and K Matsuura. 2005. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Climate Research* 30 (12 2005), 79. doi:10.3354/cr030079

[99] Scott Wood, Erin Yao, Lisa Haisfield, and Susan Lottridge. 2021. Establishing Standards of Best Practice in Automated Scoring. ACT Research. Technical Brief. *ACT, Inc.* (2021).

[100] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv preprint arXiv:2308.08155* (2023). doi:10.48550/arXiv.2308.08155

[101] Kaixun Yang, Mladen Raković, Yuyang Li, Quanlong Guan, Dragan Gašević, and Guanliang Chen. 2024. Unveiling the Tapestry of Automated Essay Scoring: A Comprehensive Investigation of Accuracy, Fairness, and Generalizability. doi:10.48550/arXiv.2401.05655 arXiv:2401.05655 [cs.CL]

[102] Kexin Bella Yang, Vanessa Echeverria, Zijing Lu, Hongyu Mao, Kenneth Holstein, Nikol Rummel, and Vincent Aleven. 2023. Pair-Up: Prototyping Human-AI Coorchestration of Dynamic Transitions between Individual and Collaborative Learning in the Classroom. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 453, 17 pages. doi:10.1145/3544548.3581398

[103] Calvin Yeung, Jeff Yu, King Chau Cheung, Tat Wing Wong, Chun Man Chan, Kin Chi Wong, and Keisuke Fujii. 2025. A Zero-Shot LLM Framework for Automatic Assignment Grading in Higher Education. *arXiv preprint arXiv:2501.14305* (2025).

[104] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300509

[105] Weizhe Yuan, Pengfei Liu, and Matthias Gallé. 2024. LLMCrit: Teaching Large Language Models to Use Criteria. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7929–7960. doi:10.18653/v1/2024.findings-acl.472

[106] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. doi:10.1145/3351095.3372852

[107] Chenyan Zhao, Mariana Silva, and Seth Poulsen. 2025. Language models are few-shot graders. In *International Conference on Artificial Intelligence in Education*. Springer, 3–16.

[108] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL] https://arxiv.org/abs/2306.05685

[109] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang,

Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2306.05685

[110] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. arXiv:2205.10625 [cs.AI] https://arxiv.org/abs/2205.10625

A RQ1

A.1 Dataset

Field	Type	Remarks
Question (text + image)	multimodal	avg. 39 words, avg. 0.2 figures
Model solution	text	avg. 26 words
Student answer	text	varies with mark (see Table 2)
Maximum mark	int	{1, 2, 3, 4}
Gold score	int	0–Maximum mark
Grade level	ordinal	Primary 1 – Primary 6

Table 9: Schema of each datum used in evaluation.

Questions and model solutions All questions follow the syllabus and question format of the Singapore national curriculum (prescribed by the Ministry of Education, Singapore). However, these questions were not created or approved by the Ministry of Education.

Student answers All student answers were entered as text (no images) by students studying at their respective levels through the online interface of our partner edtech organization.

Marking Scheme used in gold labels and human marks Gold labels and human marks were produced by human teachers, who followed a customized marking scheme used by our partner edtech organization. The marking scheme largely follows Singapore's Primary School Leaving Exam (PSLE) rubric, with one key modification in awarding full marks for multi-mark questions. In our partner's marking scheme, a student receives full points for a correct final answer, regardless of whether they showed intermediate work, whereas the standard PSLE marking scheme awards full marks only if the final answer is correct and proper working steps are shown. Marching scheme used,

For 1-mark questions

- 1 mark is given if the final answer is fully correct (including correct units and answer format) and 0 if not.

For 2–4 mark questions

- Full marks are given if the final numerical answer is correct (even without intermediate steps) under our scheme.
- If the final answer is correct but minor formatting elements (units, etc.) are missing, we deduct 1 mark.
- If the final answer is incorrect, we award partial credit: starting from 0, the student earns 1 mark for each correct intermediate step result they provided. The PSLE marking scheme would similarly give partial credit for each correct step if the final answer is wrong.

We adopted this customized marking scheme in our TC-MAG model due to the nature of the marked dataset available in the partner organization. Still, due to the modular nature of our TC-MAG architecture, we can change the desired marking scheme to fit the official PSLE marking rubric, or other global curricula. The Singapore Ministry of Education specifies the national curriculum syllabus, placing mathematical problem-solving and reasoning processes at the core of instruction alongside content mastery [44, 45, 60, 61]. This context naturally favors stepwise, partial-credit grading schemes, which the TC-MAG framework incorporates through marking guidelines and agent prompts.

A.2 Evaluation metrics

A.2.1 Formulas used. Confusion matrices: Human and AI accuracy The normalized confusion matrices shown above are derived as follows. For each class i (true mark) and class j (predicted mark) we compute

Notation. Let y_i be the gold (true) mark for item i, $\hat{y}_i^{(m)}$ the mark given by the model, and $\hat{y}_i^{(h)}$ the mark given by the human marker. Each dataset is partitioned into strata corresponding to the maximum possible marks $K \in \{1, 2, 3, 4\}$.

$$\begin{aligned} n &= \text{number of items per stratum}, \\ \mathbf{y} &= (y_1, y_2, \ldots, y_n), \\ \hat{\mathbf{y}}^{(m)} &= (\hat{y}_1^{(m)}, \hat{y}_2^{(m)}, \ldots, \hat{y}_n^{(m)}), \\ \hat{\mathbf{y}}^{(h)} &= (\hat{y}_1^{(h)}, \hat{y}_2^{(h)}, \ldots, \hat{y}_n^{(h)}). \end{aligned}$$

Evaluation Metrics.

Exact Accuracy.

$$\text{ExactAcc} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[\hat{y}_i = y_i].$$

Within-1 Accuracy.

$$\text{Within-1Acc} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[|\hat{y}_i - y_i| \le 1].$$

Mean Absolute Error (MAE).

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|.$$

Quadratic Weighted Kappa (QWK).

$$\text{QWK} = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}, \quad w_{ij} = \frac{(i-j)^2}{(K-1)^2}.$$

Here O_{ij} is the observed rating frequency and E_{ij} the expected frequency under independence.

Matthews Correlation Coefficient (MCC).

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.$$

Cohen's κ.

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad p_o = \sum_i P(y_i = \hat{y}_i), \quad p_e = \sum_k P(y_i = k) P(\hat{y}_i = k).$$

Mark stratum	Number of unique student responses	Mean words in question (SD)	Mean words in model answer (SD)	Mean words in students' answers (SD)	Range of marks awarded to students (gold standard)
1-mark	500	27.06 (± 12.15)	4.00 (± 1.80)	1.54 (± 0.73)	Binary (0, 1)
2-mark	500	36.90 (± 16.88)	19.42 (± 7.33)	3.19 (± 1.60)	Ordinal (0, 1, 2)
3-mark	500	43.73 (± 13.63)	35.55 (± 6.10)	13.57 (± 5.64)	Ordinal (0, 1, 2, 3)
4-mark	500	47.00 (± 15.18)	47.27 (± 7.40)	24.21 (± 12.53)	Ordinal (0, 1, 2, 3, 4)

Table 10: Dataset statistics by mark stratum.

F1 Score.

$$F1 = \frac{2TP}{2TP + FP + FN}.$$

Bootstrap Confidence Intervals and Significance Testing. We used a stratified paired bootstrap with $B = 200$ resamples per stratum. For each bootstrap replicate b, items were sampled with replacement within strata and all metrics recomputed.

$$\hat{\theta}_b^{(m)}, \hat{\theta}_b^{(h)} \text{ for } b = 1, \ldots, B$$

The 95% confidence interval for metric θ is taken as:

$$[\text{quantile}_{2.5\%}, \text{quantile}_{97.5\%}] \text{ of } \{\hat{\theta}_b\}.$$

A one-sided paired bootstrap test compared TC-MAG vs. human baselines:

$$H_0 : \theta_h \geq \theta_m, \quad H_A : \theta_m > \theta_h.$$

The p-value is the fraction of bootstrap replicates where $\hat{\theta}_b^{(m)} \leq \hat{\theta}_b^{(h)}$. Holm–Bonferroni correction was applied across baselines.

A.3 Findings: LLM Baseline accuracies of 2-4 mark questions

A.3.1 2-mark. Please refer to Table 11.

Baseline	N	QWK	Exact-acc	Within-1	MAE
GPT-4o simple	500	0.839	0.708	0.960	0.322
GPT-4o CoT	500	0.853	0.723	0.958	0.329
GPT-4o TC-MAG	500	0.814	0.770	0.992	0.238
GPT-o3 simple	500	0.932	0.844	0.998	0.174
GPT-o3 CoT	500	0.935	0.847	0.994	0.174
GPT-o3 TC-MAG (our model)	500	0.945	0.889	0.998	0.113

Table 11: 2-mark (ordinal) results.

A.3.2 3-mark. Please refer to Table 12.

Baseline	N	QWK	Exact-acc	Within-1	MAE
GPT-4o simple	500	0.737	0.614	0.928	0.462
GPT-4o CoT	500	0.768	0.632	0.918	0.464
GPT-4o TC-MAG	500	0.777	0.684	0.930	0.392
GPT-o3 simple	500	0.881	0.782	0.974	0.246
GPT-o3 CoT	500	0.902	0.820	0.982	0.202
GPT-o3 TC-MAG (our model)	500	0.933	0.864	0.990	0.148

Table 12: 3-mark (ordinal) results.

Baseline	N	QWK	Exact-acc	Within-1	MAE
GPT-4o simple	500	0.744	0.562	0.884	0.588
GPT-4o CoT	500	0.780	0.518	0.870	0.652
GPT-4o TC-MAG	500	0.780	0.600	0.910	0.524
GPT-o3 simple	500	0.919	0.730	0.986	0.286
GPT-o3 CoT	500	0.914	0.740	0.970	0.292
GPT-o3 TC-MAG (our model)	500	0.955	0.840	0.992	0.168

Table 13: 4-mark (ordinal) results.

A.3.3 4-mark. Please refer to Table 13.

A.4 Representative Questions

This section contains 12 representative sample questions (3 per mark stratum: 1-, 2-, 3-, and 4-mark questions), along with their model solutions. These examples span different primary grade levels and question types to illustrate the diversity of our evaluation corpus.

A.4.1 4-mark questions.

(1) **[4 marks] Level:** Primary 6 **Topic:** Gap and Difference Strategies
Question. Mrs Tay baked some muffins and packed them into boxes containing 6 muffins each. The next day, she baked 96 more muffins and she re-packed all her muffins into boxes containing 10 muffins each and realised that she needed 4 more boxes. How many muffins did she bake altogether?
Solution.

$$\text{Muffins in the 4 more boxes} = 4 \times 10 = 40$$
$$\text{Difference per box} = 10 - 6 = 4$$
$$\text{Muffins added to original boxes} = 96 - 40 = 56$$
$$\text{Number of original boxes} = 56 \div 4 = 14$$
$$\text{Number of boxes the next day} = 14 + 4 = 18$$
$$\text{Total number of muffins} = 18 \times 10 = 180$$

Answer: 180 muffins

(2) **[4 marks] Level:** Primary 6 **Topic:** Model drawing strategies
Question. Jolene and Michelle had some money. Jolene had \$485 more than Michelle. After Jolene spent $\frac{1}{5}$ of her money on a watch and Michelle spent $\frac{1}{3}$ of her money on a mobile phone, Jolene had \$584 more than Michelle. How much money did they have altogether at first? Refer to Figure 8.
Solution. Use a common unit: $\frac{1}{5} = \frac{3}{15}$ and $\frac{1}{3} = \frac{5}{15}$.

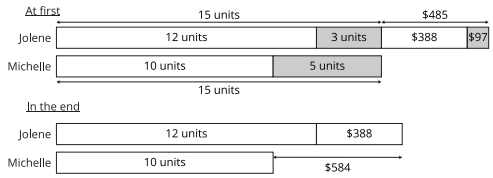


Figure 8: 4 mark - question 2

	Spent	Left
Jolene	3 units	12 units
Michelle	5 units	10 units

Jolene's difference over Michelle decreases because Jolene has fewer units left than before.

$$\frac{1}{5} \times \$485 = \$97$$

$$\begin{aligned}
\text{New difference due to Jolene spending} &= \$485 - \$97 = \$388 \\
2 \text{ units} &= \$584 - \$388 = \$196 \\
1 \text{ unit} &= \$196 \div 2 = \$98 \\
\text{Total at first} &= 30 \text{ units} + \$485 \\
&= 30 \times \$98 + \$485 \\
&= \$2940 + \$485 \\
&= \$3425
\end{aligned}$$

Answer: \$3425

(3) **[4 marks] Level:** Primary 3 **Topic:** Money
Question. A party shop is selling balloons at 5 for \$2. Mrs Tay wants to buy 100 balloons for a party. If she only has ten-dollar notes in her wallet, how many ten-dollar notes does she need to give the cashier? She needs to give the cashier _____ ten-dollar notes.
Solution.

$$\begin{aligned}
\text{Number of groups of 5 balloons} &= 100 \div 5 = 20 \\
\text{Cost of 20 groups} &= 20 \times \$2 = \$40 \\
\text{Number of ten-dollar notes} &= \$40 \div \$10 = 4
\end{aligned}$$

Answer: 4

A.4.2 3-mark questions.

(1) **[3 marks] Level:** Primary 5 **Topic:** Four Operations of Whole Numbers
Question. Kyla bought 5 boxes of oranges from her supplier for her fruit stall. Each box contained 56 oranges. She then repacked them into packets of 4 and sold them at \$2.50 per packet. She sold all the packets during the weekends. How much money did she collect from her sales?
Solution.

$$\begin{aligned}
\text{Number of oranges} &= 5 \times 56 = 280 \\
\text{Number of packets} &= 280 \div 4 = 70 \\
\text{Money collected} &= 70 \times \$2.50 = \$175
\end{aligned}$$

Answer: \$175

(2) **[3 marks] Level:** Primary 5 **Topic:** Whole Number Strategies

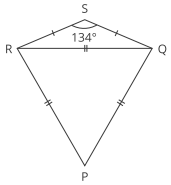


Figure 9: 3 mark - question 3

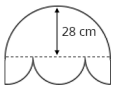


Figure 10: 2 mark - question 1

Question. Teacher Ryan arranged all his students in his class to line up to go for Chinese New Year dancing event. Every 4 girls was standing in between every 2 boys. There were 28 girls participating. How many students did he have in his class?
Solution. Between every 2 boys there is one "spacing", and each spacing contains 4 girls.

$$\begin{aligned}
\text{Number of spacings} &= 28 \div 4 = 7 \\
\text{Number of boys} &= 7 + 1 = 8 \\
\text{Total students} &= 28 + 8 = 36
\end{aligned}$$

Answer: 36 students

(3) **[3 marks] Level:** Primary 5 **Topic:** Angles
Question. The figure below, not drawn to scale, is made up of an isosceles triangle SQR and an equilateral triangle PQR. $\angle RSQ = 134^\circ$. Find $\angle SQP$. Refer to figure 9. **Solution.**

$$\angle SQR = \frac{180^\circ - 134^\circ}{2} = 23^\circ \quad \text{(isosceles triangle)}$$

$$\angle PQR = \angle PRQ = \angle RPQ = 60^\circ \quad \text{(equilateral triangle)}$$

$$\angle SQP = 23^\circ + 60^\circ = 83^\circ$$

Answer: 83°

A.4.3 2-mark questions.

(1) **[2 marks] Level:** Primary 6 **Topic:** Circles and Composite Figures
Question. The figure below is made up of a big semicircle and 4 small quarter circles. (Take $\pi = \frac{22}{7}$.) Find the area of the figure 10. **Solution.**

$$\text{Radius of small quarter circle} = 28 \div 2 = 14 \text{ cm}$$

$$\text{Area of big semicircle} = \frac{1}{2} \times \frac{22}{7} \times 28 \times 28 = 1232 \text{ cm}^2$$

$$\text{Area of 4 quarter circles} = \frac{22}{7} \times 14 \times 14 = 616 \text{ cm}^2$$

$$\text{Area of figure} = 1232 + 616 = 1848 \text{ cm}^2$$

Answer: 1848 cm^2

(2) **[2 marks] Level:** Primary 1 **Topic:** Numbers to 10

Figure 11: 2 marks - question 2

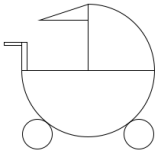


Figure 12: 1 mark - question 1

Question. What are the missing numbers? Refer to figure 11.
Solution. Since the numbers are increasing, count on starting from 2:

$$2,\ 3,\ 4,\ 5,\ 6$$

Answer: 4, 5

(3) **[2 marks] Level:** Primary 2 **Topic:** Fractions
Question. Emily had $4\frac{1}{3}$ m worth of ribbon. Every present that she wrapped used up $\frac{1}{5}$ m of ribbon. Find the maximum number of presents that she could wrap and the fraction of ribbon left (in metres).
Solution.

$$4\frac{1}{3} \div \frac{1}{5} = \frac{13}{3} \times \frac{5}{1} = \frac{65}{3} = 21\frac{2}{3}$$

So she can wrap 21 presents.

$$\text{Ribbon used} = 21 \times \frac{1}{5} = 4\frac{1}{5}\ \text{m}$$

$$\text{Ribbon left} = 4\frac{1}{3} - 4\frac{1}{5}$$

$$= 4\frac{5}{15} - 4\frac{3}{15} = \frac{2}{15}\ \text{m}$$

Answer: 21 presents and $\frac{2}{15}$ m

A.4.4 1-mark questions.

(1) **[1 mark] Level:** Primary 2 **Topic:** Shapes and Patterns
Question. Look at figure 12. There are _____ rectangles and triangles altogether in the figure. **Solution.**

$$\text{Number of rectangles} = 2$$

$$\text{Number of triangles} = 1$$

$$\text{Total} = 2 + 1 = 3$$

Answer: 3

(2) **[1 mark] Level:** Primary 3 **Topic:** Time
Question. Perform the following subtraction of time:

5 h 5 min − 1 h 25 min = _____ h _____ min

Solution.

$$5\text{ h }5\text{ min} - 1\text{ h }25\text{ min} = 4\text{ h }65\text{ min} - 1\text{ h }25\text{ min}$$

$$= 3\text{ h }40\text{ min}$$

Answer: 3 h 40 min

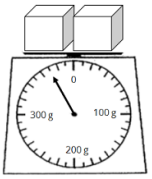


Figure 13: 1 mark - question 3

(3) **[1 mark] Level:** Primary 3 **Topic:** Mass
Question. Two identical cubes are placed on the weighing scale. What is the mass of each cube? Refer to figure 13.
Solution.

$$\text{Mass of 2 identical cubes} = 370\text{ g}$$

$$\text{Mass of each cube} = 370 \div 2 = 185\text{ g}$$

Answer: 185 g

A.5 Prompt Constructs

Agent 1: Guideline Extractor

- Purpose: Extract only the marking-guideline rules needed to grade one question, preserving rule IDs for later citation.
- Inputs: {question}, {solution}, {question_topic}, {full_marking_guideline_document}
- Output: {guideline_snippet}
- Hard constraints
  - Copy only relevant rules.
  - Do not paraphrase rule IDs.
  - Anticipate likely variations in student answers and provide topic-specific guidelines to evaluate them, including the rules needed to decide between variants.
  - Output in ≤250 words

Agent 2: Rubric Creator

- Purpose: Convert the official solution into an analytic step-wise rubric with step codes S1…Sk and Final answer, plus unit requirements.
- Inputs: {question}, {solution}, {question_topic}, {max_marks}, {guideline_snippet}
- Output: {rubric_json}
  - List of rubric steps, each with step_code, step_summary, evidence_required, units_required, and mark_value.
- Hard constraints
  - Decompose the solution into intermediate steps; assign 1 mark per intermediate step; the final answer has {max_marks} marks.
  - Enforce step-count constraints so students cannot receive full marks solely via intermediate steps.
  - Two worked examples of different scenarios of step decomposition
  - Evidence_required must be a value/conclusion (no calculations); step evidence must be mutually exclusive.
  - Intermediate step units_required default to "none"; final step units_required determined by explicit rules (e.g., fill-in-blank unit already given in question vs not).
  - The output must match the provided JSON schema.

Agents 3 and 4: Marker Agents (Marker A and Marker B)

- Purpose: Score the student's answer against the rubric using criteria-level evaluation.
- Inputs: {question}, {solution}, {max_marks}, {student_answer}, {rubric_json}, {guideline_snippet}
- Output: {markerA_output}, and {markerB_output} from respective agents
  - Step_marks: step_code, criteria scores, marks contribution, short justification
  - Final_marks: marks
- Hard constraints
  - Authority constraints ("don't deviate from the provided solution")
  - Criteria definitions (equivalence/unit/format).
  - Final answer is evaluated first under three criteria: 1) Answer equivalence; 2) Unit accuracy; 3) Format/clarity (using guideline snippet).
  - If final answer meets all criteria → {max_marks}; if only equivalence passes but unit/format fail → {max_marks-1}; if equivalence fails → 0 from final answer.
  - If final answer yields 0, sum 1-mark contributions for intermediate steps whose evidence_required can be unambiguously matched in the student work.
  - Scoring algorithm (final answer precedence, partial-credit logic)
  - The output must match the provided JSON schema.

Agent 5: Arbitrator Agent

- Purpose: Resolve disagreements between Marker A/B at the criterion level, selecting the better evaluation and citing the exact rubric step and/or guideline rule IDs that justify the decision.
- Inputs: All agent 3/4 inputs + {markerA_output}, {markerB_output}
- Output: {arbitrator_output}
  - final_marks, preferred_marker, lists of rubrics_missed and rule_ids_missed, and brief justification.
- Hard constraints
  - Must choose one marker output and justify briefly.
  - Must cite the missed criterion by step_code or guideline rule_ids referred to for marker choice.
  - The output must match the provided JSON schema.

Agent 6: Confidence Analyzer

- Purpose: Provide a conservative High/Low confidence tag for the final mark, used for human routing.
- Inputs:. All agent 5 inputs + {arbitrator_output} (if any).
- Output: {confidence_output}
- Confidence bucket ∈ {High, Low}, and brief justification.
- Hard constraints
  - Conservative rule: if any credible uncertainty remains, output "Low confidence."
  - Must explicitly comment on each rubric/rule grounding quality.
  - ≤100-word justification.
  - The output must match the provided JSON schema.

B RQ2

B.1 Pre-task survey

B.1.1 List of questions and results . 7-point Likert; 1=Strongly disagree, 7=Strongly agree. Please refer to Table 14.

B.2 User study Part 1

B.2.1 Distribution of questions used in Part 1 . Please refer to Table 15.

B.3 User study Part 2

B.3.1 Distribution of questions used in Part 2. Please refer to Table 16.

Concurrent think-aloud Before beginning Part 2, participants were given with prompt, "Please verbalize what you are noticing about the rubric, explanation steps, and confidence." We captured *verbatim quotes* and timestamp alignments to telemetry.

B.4 User study Part 3

B.4.1 Survey questions. Survey questions for each architecture (**Staged** and **Summarized**): **Trust & willingness to delegate:**

- "I **trust** this system to assign **criterion-aligned** marks most of the time." (7-point)
- "I would **delegate** marking of similar questions to this system." (7-point + binary "Would you delegate to this system for your next unit?")
- "I felt I could **understand** why this system assigned each mark." (7-point; perceived transparency)
- "I could **predict** when the system might be wrong." (7-point; predictability)
- Open-ended: "Briefly, **why** would/wouldn't you delegate to this system?"

Confidence cue perception:

- "The **confidence indicator** helped me decide whether to accept or re-mark." (7-point)
- "When the system showed **Low** confidence, I was **less likely** to accept its mark." (7-point)

Architecture preference:

- "If you had to pick one for your class next term, which would you choose, and why?"

B.4.2 NASA-TLX survey questions and results. Please refer to Table 17 for the survey questions used and Table 18 for the results.

B.5 Performance metrics

Formulas used.

Delegation rate.

$$\text{Delegation} = \frac{\text{AI marked-not changed (delegated)}}{\text{Total instances}}$$

Selective Delegation rate. Confusion matrix (flip test vs AI correctness; for only when teacher viewed AI marks) Let:

- **TP** = AI wrong **and** flipped
- **FN** = AI wrong **and** kept
- **TN** = AI right **and** kept
- **FP** = AI right **and** flipped

Survey question	Mean ± SD	Top–box (6–7) %
Human graders are generally more reliable than AI graders on my subject	5.43 ± 1.09	42.9
I would accept an AI-generated mark if it follows the correct marking rubric.	6.36 ± 0.93	85.7
If I disagree with an AI's mark, I expect that I am more likely to be correct.	5.64 ± 1.40	64.28
I am comfortable delegating routine marking to an AI if I can audit exceptions.	6.00 ± 1.04	64.3
An AI grader can consistently apply marking rubric-level rules.	5.64 ± 0.84	57.1
I can predict how the AI will assign marks across similar questions.	5.29 ± 0.73	42.9
I trust the intentions of developers who build grading AI for educational institutes.	6.21 ± 0.80	78.6
In general, I trust automated systems when they are tested before releasing.	6.07 ± 0.48	92.9
I tend to try AI systems myself before deciding to trust them.	5.50 ± 2.03	71.4

Table 14: Pre-task survey items - 7-pt Likert, $N = 14$

Mark stratum	Correct-High	Correct-Low	Incorrect-High	Incorrect-Low	Total	Achieved accuracy	Realised accuracy in RQ1
1-mark	2	0	0	0	2	6/6 = **100%**	**98.4%**
2-mark	1	1	0	0	2	5/6 = **83%**	**93.0%**
3-mark	1	0	1	0	2	9/10 = **90%**	**86.4%**
4-mark	1	0	0	1	2	8/10 = **80%**	**84.0%**

Table 15: Split of questions used in Part 1 study to ensure a mix of incorrect/correct TC-MAG marking instances and high/low confidence labels.

Mark stratum	Correct-High	Correct-Low	Incorrect-High	Incorrect-Low	Achieved accuracy	Realised accuracy in RQ1
1-mark	3	1	0	0	4/4 = **100%**	**98.4%**
2-mark	3	0	1	0	3/4 = **75%**	**93.0%**
3-mark	3	0	0	1	3/4 = **75%**	**86.4%**
4-mark	2	0	1	1	2/4 = **50%**	**84.0%**

Table 16: Split of questions used in Part 2 study to ensure a mix of incorrect/correct TC-MAG marking instances and high/low confidence labels.

Category	Adapted question (used)	Original question
Mental Demand	While marking student answers with the AI, how mentally demanding was the task?	How mentally demanding was the task?
Physical Demand	While marking student answers with the AI, how physically demanding was the task?	How physically demanding was the task?
Temporal Demand	While marking student answers with the AI, how hurried or rushed did you feel?	How hurried or rushed was the pace of the task?
Performance	While marking student answers with the AI, how successful were you in accomplishing what you were asked to do?	How successful were you in accomplishing what you were asked to do?
Effort	While marking student answers with the AI, how hard did you have to work to accomplish your level of performance?	How hard did you have to work to accomplish your level of performance?
Frustration	While marking student answers with the AI, how insecure, discouraged, irritated, stressed, and annoyed did you feel?	How insecure, discouraged, irritated, stressed, and annoyed were you?

Table 17: Adapted NASA-TLX workload questions, which participants rated on a scale of 0–10.

$$\textbf{Sensitivity (TPR)} = \frac{TP}{TP + FN}$$

$$\textbf{Specificity (TNR)} = \frac{TN}{TN + FP}$$

$$\textbf{LR}^{+} = \frac{\text{Sens}}{1 - \text{Spec}} = \frac{\text{Sens}}{\text{FPR}}$$

$$\textbf{LR}^{-} = \frac{1 - \text{Sens}}{\text{Spec}} = \frac{\text{FNR}}{\text{Spec}}$$

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Dimension	Staged: mean (SD)	Summarised: mean (SD)
Mental Demand	5.86 (2.85)	4.86 (2.18)
Physical Demand	3.57 (2.17)	3.71 (2.81)
Temporal Demand (hurried/rushed)	3.07 (1.86)	2.86 (2.03)
Effort (work required)	3.50 (2.62)	3.07 (2.02)
Frustration (affect)	2.00 (2.42)	1.86 (1.61)
Performance (inverted)	**1.93 (2.02)**	**3.29 (3.27)**
TLX_official_mean (0–10)	**3.32 (1.55)**	**3.27 (1.59)**

Table 18: NASA-TLX workload (0–10) by explanation architecture; "Performance" is inverted.

C Discussion

C.1 Cost comparison of TC-MAG and Baselines

These costs are calculated based on the following OpenAI API prices available on their website as of October 4, 2025.

OpenAI API costs used for per question cost calculation. Please refer to Table 19.

Mean total cost for marking each question. Please refer to Table 20.

Mean total input+output tokens for marking each question. Please refer to Table 21.

Mean runtime of LLM API per question in seconds. Please refer to Table 22.

Scenario	Input $/1M	Output $/1M
GPT-4o Standard	2.5	10
GPT-4o Batch	1.25	5
o3 Standard	2	8
o3 Batch	1	4

Table 19: OpenAI API costs as of October 4, 2025.

Prompt strategy	Scenario	1-mark	2-mark	3-mark	4-mark	Mean (SD)	Mean (SD) for 10,000 questions
CoT	GPT-4o — Batch	0.0089 (0.0019)	0.0101 (0.0004)	0.0104 (0.0023)	0.0109 (0.0037)	0.0101 (0.0021)	100.63 (20.65)
	GPT-4o — Standard	0.0178 (0.0037)	0.0202 (0.0008)	0.0208 (0.0045)	0.0218 (0.0074)	0.0201 (0.0041)	201.26 (41.31)
	o3 — Batch	0.0071 (0.0015)	0.0086 (0.0005)	0.0090 (0.0019)	0.0091 (0.0030)	0.0085 (0.0017)	84.60 (17.09)
	o3 — Standard	0.0142 (0.0030)	0.0172 (0.0009)	0.0180 (0.0037)	0.0183 (0.0061)	0.0169 (0.0034)	169.20 (34.18)
TC-MAG	GPT-4o — Batch	0.0114 (0.0049)	0.0180 (0.0104)	0.0182 (0.0032)	0.0198 (0.0077)	0.0168 (0.0065)	168.49 (65.47)
	GPT-4o — Standard	0.0228 (0.0099)	0.0360 (0.0207)	0.0363 (0.0064)	0.0396 (0.0153)	0.0337 (0.0131)	336.97 (130.94)
	o3 — Batch	0.0114 (0.0050)	0.0187 (0.0085)	0.0190 (0.0036)	0.0205 (0.0071)	0.0174 (0.0061)	173.83 (60.63)
	o3 — Standard	0.0227 (0.0100)	0.0374 (0.0170)	0.0379 (0.0073)	0.0410 (0.0142)	0.0348 (0.0121)	347.66 (121.26)
Vanilla	GPT-4o — Batch	0.0079 (0.0019)	0.0079 (0.0004)	0.0081 (0.0023)	0.0087 (0.0037)	0.0081 (0.0021)	81.34 (20.60)
	GPT-4o — Standard	0.0158 (0.0037)	0.0158 (0.0008)	0.0161 (0.0045)	0.0174 (0.0074)	0.0163 (0.0041)	162.67 (41.20)
	o3 — Batch	0.0062 (0.0015)	0.0063 (0.0004)	0.0066 (0.0018)	0.0069 (0.0030)	0.0065 (0.0017)	65.08 (16.83)
	o3 — Standard	0.0124 (0.0030)	0.0127 (0.0008)	0.0132 (0.0037)	0.0139 (0.0060)	0.0130 (0.0034)	130.15 (33.66)

Table 20: Cost per question (and per 10,000) for TC-MAG and baselines under standard vs. batch settings, by mark stratum and averaged.

Prompt strategy	Scenario	1-mark	2-mark	3-mark	4-mark
CoT	GPT-4o	4442.206 (201.2445)	5053.254 (936.1298)	5189.982 (1132.2522)	5440.55 (1860.9829)
	o3	4449.486 (282.4776)	5382.606 (933.7186)	5612.002 (1165.6006)	5705.57 (1891.1585)
TC-MAG	GPT-4o	7608.0763 (1608.3931)	12013.8788 (2474.2308)	12110.7938 (3835.5834)	13196.5938 (5175.5666)
	o3	9468.1188 (2275.8994)	15583.4475 (3125.3922)	15795.7675 (4449.5073)	17096.7525 (5307.2182)
Vanilla	GPT-4o	3938.816 (190.2211)	3943.976 (936.4563)	4030.376 (1133.4951)	4354.194 (1859.6119)
	o3	3864.67 (242.2269)	3964.992 (930.5844)	4109.93 (1147.0398)	4329.758 (1887.6148)

Table 21: Total tokens per question for TC-MAG and baselines (by model and mark stratum).

Prompt strategy	Model	1-mark	2-mark	3-mark	4-mark
CoT	GPT-4o	10.9761 (8.8865)	13.5221 (16.5404)	16.8521 (11.9888)	17.4214 (13.1744)
	o3	18.2643 (8.706)	28.0869 (9.2168)	30.363 (16.1338)	35.071 (19.9536)
Simple	GPT-4o	8.6512 (3.4773)	9.2087 (3.7395)	10.077 (4.1627)	11.8114 (5.1043)
	o3	21.0464 (10.1919)	24.8229 (24.4807)	25.6753 (18.2845)	29.346 (14.7187)
TC-MAG	GPT-4o	16.1122 (26.8534)	21.6127 (6.2684)	23.6487 (6.023)	25.3682 (5.8533)
	o3	26.9498 (7.456)	37.6004 (10.745)	41.2038 (10.1782)	42.0307 (11.8474)

Table 22: LLM API runtime per question for TC-MAG and baselines under real-time vs. batch processing (by model and mark stratum).