MASGrader: A Multi-Agent Framework for Automated Subjective Question Grading

Anonymous ACL submission

Abstract

Automated grading of subjective questions remains a significant challenge in educational assessment. Traditional manual grading is inefficient and inconsistent, while existing AI-based methods lack flexibility and robustness in handling diverse answers. This paper introduces MASGrader, an innovative multi-agent framework for grading subjective answers, consisting of four agents: the Overview Agent, which performs macroscopic evaluation; the Detail Review Agent, which conducts microscopic reviews; the Logical Validation Agent, which checks semantic and logical consistency; and the Supervisory Agent, which coordinates debates and makes final decisions. MASGrader enhances grading accuracy, stability, and transparency by simulating human-like debate and reflection mechanisms. Experiments on a dataset of 500 subjective answers demonstrate that MASGrader improves weighted Kappa scores and accuracy by 5-10% compared to a single-agent baseline while generating detailed scoring rationales that increase interpretability. By introducing dynamic collaboration, logical validation, and iterative self-improvement, this multi-agent framework provides a reliable solution for high-stakes educational assessment.

1 Introduction

The emergence of automated grading for subjective questions aims to solve the inefficiency and fairness issues of manual grading (Bennett, 2011).Early systems like Oxford-UCLES (Sukkarieh et al., 2004) relied on keyword matching but struggled with semantic variations (Leacock and Chodorow, 2003). Later methods, such as the Latent Semantic Analysis (LSA) of AutoTutor (Wiemer-Hastings et al., 1999), improved coherence detection but had difficulty with causal relationships. While deep learning models like RNNs, CNNs (Surya et al., 2019), and LSTMs (Graves, 2013) captured sequential patterns in short answers, they struggled with logical consistency in longer responses.

Modern pre-trained models (e.g., BERT, GPT) (Devlin et al., 2019; Radford, 2018) achieve near-human-level accuracy through task-specific fine-tuning (Condor et al., 2021), but they still face issues such as dependence on large datasets and opaque scoring mechanisms (Ribeiro et al., 2016), including poor portability and lack of interpretability, which undermine exam fairness.

To address the limitations of existing methods, this paper introduces an automated grading approach based on a multi-agent (Guo et al., 2024) framework. The framework consists of four agents: the Overview Agent, which grades from a macroscopic perspective; the Detail Review Agent, which ensures scoring accuracy by examining specific details; the Logical Validation Agent, which checks semantic consistency and logical coherence; and the Supervisory Agent, which coordinates debate and decision-making among the agents. The agents collaborate and debate to form the final grading decision, enhancing the accuracy and consistency of the scores (Stone and Veloso, 2000). This framework simulates expert-level debates and reflective processes, improving system stability, flexibility, and interpretability while providing grading rationales, ensuring transparency and fairness.

To validate the effectiveness of our method, we constructed a free-text subjective question-answer dataset consisting of 500 responses per question. Experimental results show the multi-agent framework significantly outperforms traditional large language model-based grading methods, with performance improvements of 5-10%. The contributions of this paper are as follows:

- We propose a novel multi-agent framework for automated grading, addressing the limitations of existing methods in flexibility, portability, and interpretability.
- We validate this framework on our custom-built dataset, with 500 answers per question,

comparing it with traditional large language model methods. The results demonstrate significant improvements in grading accuracy, stability, and interpretability;

- We demonstrate the effective application of multi-agent collaboration in automated grading, expanding the scope of automated grading systems and offering significant potential for educational applications.

2 Related Work

Rule-Based Automated Grading Research on automated grading has evolved from rule-driven to data-driven approaches. Early methods, like the Oxford-UCLES system by Sukkarieh et al. (Sukkarieh et al., 2004), used pattern matching with defined keywords and synonyms. Similarly, Mitchell et al. (Mitchell et al., 2002) scored answers based on keyword presence. However, these methods struggled with semantic diversity and linguistic structures, such as synonyms and word order variations (Leacock and Chodorow, 2003).

Progression from Statistical to Semantic Analysis Advances in NLP led to statistical methods based on semantic similarity. AutoTutor (Wiemer-Hastings et al., 1999) used Latent Semantic Analysis (LSA) and the Bag-of-Words (BOW) to assess answer quality but ignored word order and causal relationships (Landauer and Dumais, 1997). Later improvements, such as Explicit Semantic Analysis (ESA) (Gabrilovich et al., 2007) and SELSA (Kanejiya et al., 2003), enhanced semantic representation but lacked fine-grained feedback, limiting interpretability (Mohler and Mihalcea, 2009).

Challenges with Deep Learning and Model Interpretability Deep learning has greatly improved grading accuracy. Models based on RNNs and CNNs enhance accuracy by capturing text sequences (Surya et al., 2019). Pre-trained models (e.g., BERT, GPT) perform better with fine-tuning but require large annotated datasets and have limited transferability (Condor et al., 2021; Zhang et al., 2022). Existing methods lack interpretability and fail to justify scores (Nielsen et al., 2009).

3 Method

This section introduces the multi-agent grading framework, comprising the Overview Agent, Detail Reviewer Agent, Logic Validator Agent, and Supervisor Agent, each with distinct roles and tasks.

3.1 Overview and Detail Reviewer Agents

The *Overview Agent* assesses the overall structure and logical coherence of answers, ensuring key points are addressed and argumentation is clear. It adheres to grading guidelines, awarding points only when requirements are fully met. Conversely, the *Detail Reviewer Agent* evaluates answers based on accuracy, word usage, and clarity, ensuring rigor through detailed grading. During the debate phase, it challenges the scores of the Overview Agent and provides detailed justifications, enhancing grading rationality and system transparency.

The evaluation results of both agents are based on the semantic understanding of the answer A by the large language model (LLM). The grading results are represented as a list of tuples, each containing a score and its corresponding rationale, mapped to each scoring point $s_i \in S$ in the grading criteria S. The grading result is defined as: $\text{Result} = [(\text{Score}_i, \text{R}_i) \mid s_i \in \text{S}]$ where Score_i represents the score for s_i, and R_i is the rationale associated with s_i.

This result is generated through a multi-stage process involving semantic matching and score generation. First, semantic matching establishes a mapping between the answer A and the scoring criteria s_i through deep semantic understanding by the LLM, where the function $\text{Match}(A, s_i)$ indicates that A meets the requirements of s_i. Based on this match, quantitative scoring is performed. If $\text{Match}(A, s_i)$ is true, the score for s_i, denoted as $Score_i$, is equal to the full score f_i; otherwise, it is zero, as shown in Equation 1 and 2.

After each grading process, the agent engages in self-reflection, using the overview agent as an example, comparing the initial grading result $Result^0_{over}$ with the final grading result $Result_{final}$. The reflection outcome F_{over} is then added to the experience repository E to optimize future grading, as shown in Equation 3.

$$\text{Match} = \text{LLM}(A, s_i) \tag{1}$$

$$Score_i = \begin{cases} f_i & \text{if Match}(A, s_i) = True, \\ 0 & otherwise. \end{cases} \tag{2}$$

$$F_{over} = \text{LLM}(Result^0_{over}, Result_{final}, A, S) \tag{3}$$

3.2 Logic Validator Agent

The *Logic Validator Agent* ensures logical and semantic consistency by checking the internal consistency of R and the consistency between R and the

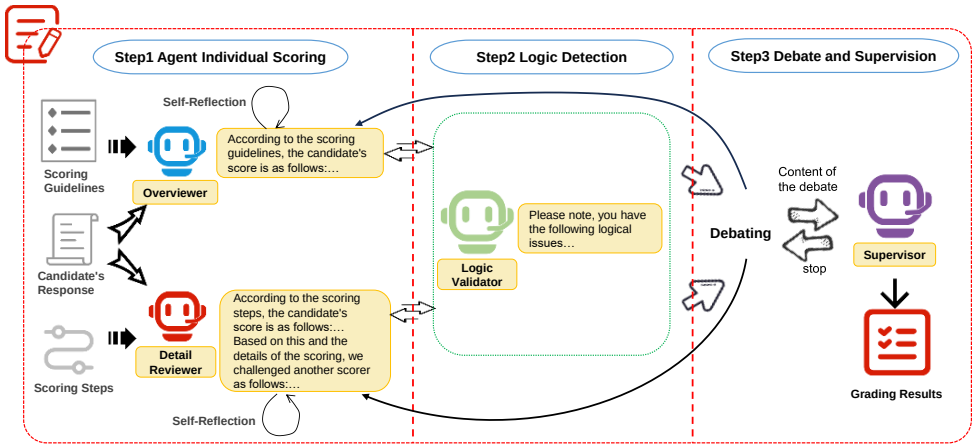


Figure 1: Overall architecture of the multi-agent grading framework.

answer A. This is formalized as:

$$\text{Valid}(R) = \text{LLM}(R, A, \\ task = \text{"Logic detection"}) \quad (4)$$

where $\text{Valid}(R, A) = \text{True}$ allows progression to the debate phase.

3.3 Supervisor Agent

The *Supervisor Agent* oversees the debate between the Overview and Detail Reviewer Agents, generating the final grading result. It monitors the debate until consensus is reached:

$$Consensus = \text{LLM}(Result^t_{over}, Result^t_{Detail}, \\ task = \text{"check consensus"}) \quad (5)$$

If $\text{Consensus} = \text{True}$, the final result is produced:

$$Result_{Final} = \text{LLM}(Result^t_{over}, Result^t_{Detail}, \\ task = \text{"generate final score"}), \quad (6)$$

3.4 Multi-Agent Grading Framework

The framework improves grading accuracy, stability, and interpretability through role allocation and collaboration, as shown in Figure 1. The Overview Agent performs a macro-level evaluation of the student's answer, generating an initial score and rationale, mimicking the human grader's quick assessment. In contrast, the Detail Reviewer Agent focuses on the answer's specifics, identifying issues or highlights, and challenging the Overview Agent's score. This role setup simulates the complementary process in human grading, where the Overview Agent ensures the overall direction, while the Detail Reviewer guarantees depth through meticulous scrutiny.

Subsequently, the Logic Validator Agent ensures logical consistency, improving reliability. After validation, the two agents engage in a debate, where the Supervisor Agent monitors the entire debate process and notifies the agents to stop debating once a consensus is reached. The final grading result is then determined and formatted for output. After the result is generated, the two grading agents reflect on it, and the reflection mechanism simulates human learning to continuously improve the system's accuracy and efficiency. Further details regarding the prompt settings and other configurations can be found in Figure 2 in the appendix.

4 Experiments

4.1 Experimental Setups

Benchmark The dataset consists of real exam questions and 500 student answers from a provincial institution in China, including a subjective question with two sub-questions (10 and 12 points) and manually annotated scores. Due to the lack of publicly available datasets with similar questions, standard answers, and scoring guidelines, this non-public dataset was chosen (confidentiality restrictions apply). Evaluation metrics include the Kappa coefficient, consistency rate, threshold agreement rate, and stability analysis, used to compare the multi-agent framework with the baseline model.

Experimental Configuration All agents are implemented with the *Qwen2.5-32B-Instruct-GPTQ-Int4* model, running on four NVIDIA L20 GPUs. The generation parameters are temperature 0.9, top-p 0.9, and repetition penalty 1.05.

Baseline model The baseline model is a single *Qwen/Qwen2.5-32B-Instruct-GPTQ-Int4* model,

	Question A					Question B				
Model	WK	Acc	Th 1	Th 2	Th 3	WK	Acc	Th 1	Th 2	Th 3
Single Agent	0.59	0.58	0.60	0.62	0.93	0.56	0.40	0.57	0.67	0.89
MASGrader	0.62	0.63	0.67	0.69	0.94	0.62	0.46	0.62	0.67	0.90

Table 1: Performance comparison of different models on two questions includes metrics such as Weighted Kappa (WK), Accuracy (Acc), and Threshold Agreement (Th).

	Question A					Question B				
Model	WK	Acc	Th 1	Th 2	Th 3	WK	Acc	Th 1	Th 2	Th 3
MASGrader	0.62	0.63	0.67	0.69	0.94	0.62	0.46	0.62	0.67	0.90
MASGrader - LV	0.57	0.61	0.65	0.68	0.93	0.59	0.43	0.62	0.66	0.87
MASGrader - DR	0.61	0.44	0.63	0.68	0.88	0.50	0.36	0.48	0.57	0.87
MASGrader - OA	0.57	0.59	0.65	0.67	0.93	0.54	0.38	0.55	0.65	0.90

Table 2: The performance comparison of the ablation study includes metrics such as Weighted Kappa (WK), Accuracy (Acc), and Threshold Agreement (Th).MASGrader: Full model; MASGrader - LV: Without Logic Validator; MASGrader - DR: Without Detail Reviewer; MASGrader - OA: Without Overview Agent.

which independently scores the answers using a prompt-based method.

4.2 Experimental Results

4.2.1 Quantitative Analysis

Table 1 shows the performance comparison of the two models on the two questions. MASGrader achieves a WK of 0.62 on both Question A and B, higher than the Single Agent's 0.59 and 0.56, indicating better agreement with human scores. Its accuracy on the two questions is 0.63 and 0.46, outperforming the Single Agent's 0.58 and 0.40. MASGrader also surpasses the Single Agent at all threshold agreement levels (Th 1, Th 2, Th 3).

4.2.2 Visualization Analysis

Figure 3 (appendix) shows the better accuracy and stability of MASGrader versus Single Agent. The scatter plot shows MASGrader scores cluster closer to human scores within a 2-point threshold, while the line chart demonstrates its scoring stability.

4.2.3 Summary of Results

The experimental results show MASGrader outperforms the Single Agent model in consistency, accuracy, and stability, especially at lower thresholds. Its higher weighted kappa, accuracy, and threshold agreement rates demonstrate better alignment with human grading, underscoring the effectiveness of the multi-agent collaborative mechanism. Visualization analysis further confirms improved scoring stability, highlighting the advantage of using multiple agents for consistent results across diverse conditions. This indicates the role-based structure and collaboration of MASGrader significantly enhance grading reliability and reduce volatility. The framework not only improves accuracy but also ensures a more reliable, transparent grading process, essential for subjective tasks like exam scoring.

4.3 Ablation Study

The ablation study reveals a performance drop when key components (Logic Validator, Detail Reviewer, Overview Agent) are removed. Specifically, removing the Logic Validator results in a 5% decrease in WK for Question A and a 3% decrease for Question B, highlighting its critical role in ensuring logical consistency. These results highlight the complementary functions of each agent, improving grading accuracy and reliability. A detailed analysis is provided in the Appendix.

Conclusion

This paper presents MASGrader, a multi-agent framework that enhances automated subjective grading through agent collaboration, logical validation, and reflection. Its innovative architecture mirrors human grading dynamics, balancing macro and micro perspectives. MASGrader outperforms single-agent models in scoring effectiveness, interpretability, and robustness, offering transparent rationales and reducing biases.

Future work will aim to extend MASGrader into multilingual contexts and integrate domain-specific knowledge graphs to improve semantic understanding, opening new avenues for more nuanced and culturally diverse educational assessment tools.

Limitations

While MASGrader demonstrates promising results, several limitations warrant consideration. First, the evaluation of the framework relies on a non-public dataset from a specific domain (Chinese civil service exams), which may limit generalizability to other educational contexts or linguistic/cultural settings. Second, the dependence on a particular large language model (Qwen/Qwen2.5-32B) and substantial computational resources (4×NVIDIA L20 GPUs) could hinder scalability and accessibility for resource-constrained institutions. Finally, the self-reflection mechanism's long-term impact—such as potential bias accumulation from iterative updates or overfitting to specific grading patterns—requires further investigation. These limitations highlight areas for future research to enhance adaptability and robustness.

Acknowledgements

References

Randy Elliot Bennett. 2011. Formative assessment: A critical review. *Assessment in education: principles, policy & practice*, 18(1):5–25.

Aubrey Condor, Max Litster, and Zachary Pardos. 2021. Automatic short answer grading with sbert on out-of-sample questions. *International Educational Data Mining Society*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Evgeniy Gabrilovich, Shaul Markovitch, et al. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Dharmendra Kanejiya, Arun Kumar, and Surendra Prasad. 2003. Automatic evaluation of students' answers using syntactically enhanced lsa. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing*, pages 53–60.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37:389–405.

Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge. 2002. Towards robust computerised marking of free-text responses.

Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 567–575.

Rodney D Nielsen, Wayne Ward, and James H Martin. 2009. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering*, 15(4):479–501.

Alec Radford. 2018. Improving language understanding by generative pre-training.

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*, pages 1135–1144.

Peter Stone and Manuela Veloso. 2000. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 8:345–383.

Jana Z Sukkarieh, Stephen G Pulman, and Nicholas Raikes. 2004. Auto-marking 2: An update on the ucles-oxford university research into using computational linguistics to score short, free text responses. *International Association of Educational Assessment, Philadephia*.

K Surya, Ekansh Gayakwad, and MJIJRTE Nallakaruppan. 2019. Deep learning for short answer scoring. *Int. J. Recent. Technol. Eng.(IJRTE)*, 7(6).

Peter Wiemer-Hastings, Katja Wiemer-Hastings, and Arthur Graesser. 1999. Improving an intelligent tutor's comprehension of students with latent semantic analysis. In *Artificial intelligence in education*, volume 99.

Lishan Zhang, Yuwei Huang, Xi Yang, Shengquan Yu, and Fuzhen Zhuang. 2022. An automatic short-answer grading model for semi-open-ended questions. *Interactive learning environments*, 30(1):177–190.

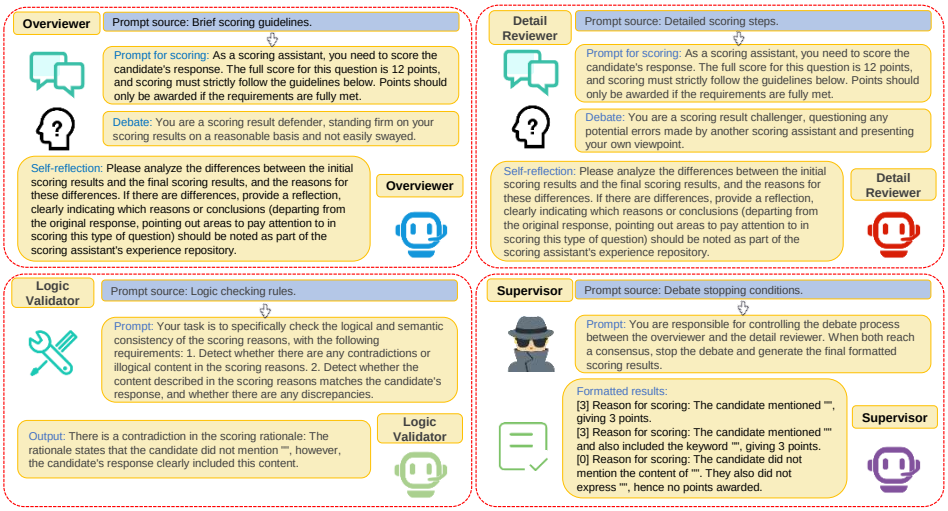


Figure 2: Roles and responsibilities of each agent in the grading process.

A Agent Roles and Grading Algorithm

This appendix provides a detailed breakdown of the roles and responsibilities of each agent in the grading process, as illustrated in Figure 2. The figure highlights the prompt sources for each agent, including the Logic Validator, Detail Reviewer, and Overview Agent, showcasing their specific contributions to the grading workflow.

Additionally, the appendix presents the algorithmic workflow of the Multi-Agent Grading Framework, which outlines the step-by-step process of how agents collaborate to evaluate subjective answers, ensuring logical consistency, accuracy, and transparency.

Algorithm 1 Multi-Agent Grading Framework

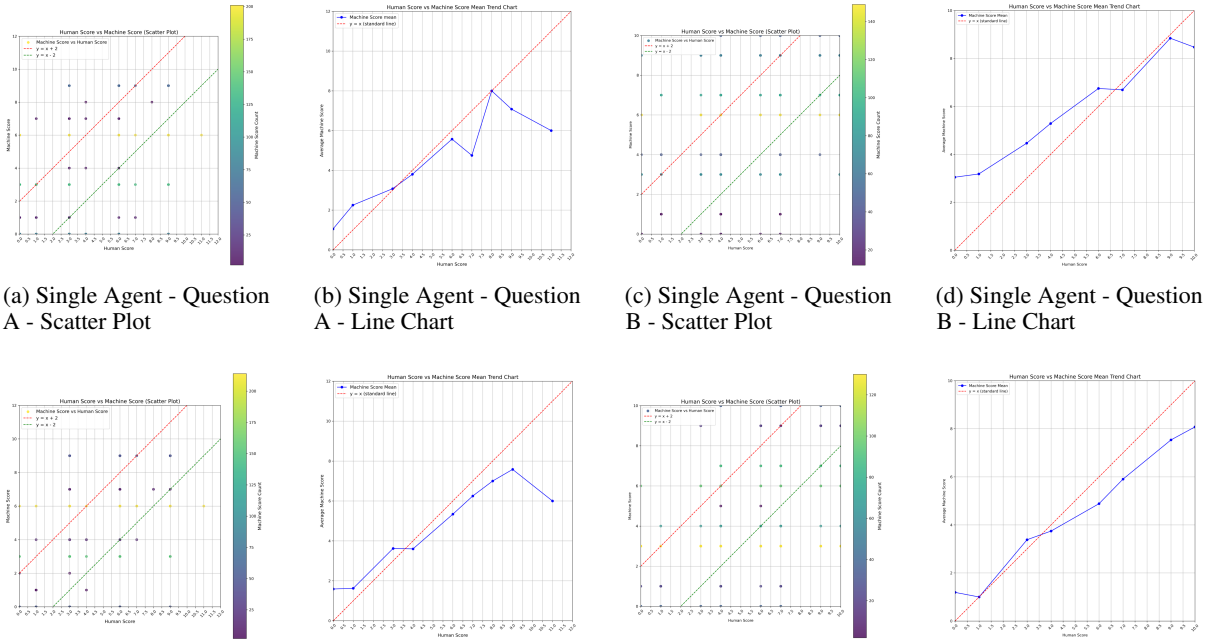
Initialization:
Initialize Agents: O (Overview Agent), DR (Detail Reviewer Agent), LV (Logic Validator Agent), S (Supervisor Agent)
Initialize Experience Repositories: $E_{Overview}$, E_{Detail}
for each Answer A in All_Answers **do**
  Grading Phase:
  $Result_O^0 \leftarrow [(Score_i, Rationale_i) | \forall s_i \in S : Match(A, s_i) \rightarrow Score_i = f_i \text{ if True else } 0]$
  $Result_{DR}^0 \leftarrow [(Score_i, Rationale_i) | \forall s_i \in S : Match(A, s_i) \rightarrow Score_i = f_i \text{ if True else } 0]$
  $Result_O, Result_{DR} \leftarrow Result_O^0, Result_{DR}^0$
  Logic Validation Phase:
  $Valid_O \leftarrow LV.Check(Result_O, A)$
  $Valid_{DR} \leftarrow LV.Check(Result_{DR}, A)$
  while not ($Valid_O$ and $Valid_{DR}$) **do**
    if not $Valid_O$ **then**
      $Result_O \leftarrow O.Revise(Result_O, A)$
    end if
    if not $Valid_{DR}$ **then**
      $Result_{DR} \leftarrow DR.Revise(Result_{DR}, A)$
    end if
    $Valid_O \leftarrow LV.Check(Result_O.Rationale_i, A)$
    $Valid_{DR} \leftarrow LV.Check(Result_{DR}.Rationale_i, A)$
  end while
  Debate and Consensus Phase:
  $Consensus \leftarrow$ False
  while not $Consensus$ **do**
    $Consensus \leftarrow S.Check(Result_O, Result_{DR})$
    if not $Consensus$ **then**
      $Result_O \leftarrow O.Reflect(Result_O, Result_{DR}, E_{Overview})$
      $Result_{DR} \leftarrow DR.Reflect(Result_{DR}, Result_O, E_{Detail})$
    end if
  end while
  Final Scoring Generation:
  $Result_{Final} \leftarrow S.GenerateFinalResult(Result_O, Result_{DR})$
  Self-Reflection Mechanism:
  $F_{Overview} \leftarrow LLM(Result_O^0, Result_{Final}, A, S)$
  $F_{Detail} \leftarrow LLM(Result_{DR}^0, Result_{Final}, A, S)$
  Update Experience Repositories:
  $E_{Overview} \leftarrow E_{Overview} \cup \{F_{Overview}\}$
  $E_{Detail} \leftarrow E_{Detail} \cup \{F_{Detail}\}$
end for
Return the final scoring results for all answers
return $Result_{Final}$ for all Answers

B Experimental Setups and Results

B.1 Dataset and Configuration

Dataset. The dataset used in this experiment comes from real exam essay questions and candidates' answers from a provincial institution in China. The dataset includes one essay question (comprising two sub-questions, which can be treated as two independent questions) and 500 corresponding candidate answers along with their scores. The full scores for the two questions are 10 and 12 points, respectively. The scores are based on detailed grading guidelines and were manually annotated. The characteristics of the dataset are as follows:

- **Subjectivity**: The questions require candidates to read and comprehend provided materials before answering, resulting in highly open-ended and subjective responses.



(a) Single Agent - Question A - Scatter Plot

(b) Single Agent - Question A - Line Chart

(c) Single Agent - Question B - Scatter Plot

(d) Single Agent - Question B - Line Chart

(e) MASGrader - Question A - Scatter Plot

(f) MASGrader - Question A - Line Chart

(g) MASGrader - Question B - Scatter Plot

(h) MASGrader - Question B - Line Chart

Figure 3: Experimental results visualization. Scatter plots show the relationship between human scores and machine scores, with the x-axis representing human scores and the y-axis representing machine scores. Line charts show the trend of mean human scores versus mean machine scores.

- **Standard Answers and Grading Guidelines**: Each question is accompanied by official, formal standard answers and detailed grading guidelines, ensuring the authority and practicality of the scoring process.

- **Expert Guidance**: The scoring process was guided by grading experts, further ensuring the accuracy and reliability of the scores.

Due to the lack of publicly available datasets with similarly open-ended questions, standard answers, and detailed grading guidelines, this non-public dataset was chosen for the experiment. For confidentiality reasons, the dataset is not publicly available.

Experimental Configuration. All agents in the experiment are implemented based on the **Qwen/Qwen2.5-32B-Instruct-GPTQ-Int4** model. The hardware configuration includes 4 NVIDIA L20 GPUs.

Baseline Model. The baseline model for this experiment is a single-agent grading model, implemented using Qwen/Qwen2.5-32B-Instruct-GPTQ-Int4. This model directly grades candidate answers without involving multi-agent collaboration or debate mechanisms. By comparing with the baseline model, we can evaluate the improvements in grading consistency, stability, and accuracy brought by the multi-agent grading framework.

Evaluation Metrics. To comprehensively evaluate the grading performance of the models, the following evaluation metrics are used:

1. **Kappa Coefficient**: Measures the agreement between model scores and human scores. The formula is:

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the observed agreement rate (the proportion of agreement between model scores and human scores), and P_e is the expected agreement rate (the proportion of agreement expected by chance). The Kappa coefficient ranges from [-1, 1], where 1 indicates perfect agreement, 0 indicates random agreement, and negative values indicate disagreement.

2. **Agreement Rate**: The proportion of cases where the model scores exactly match the human scores.

3. **Threshold Agreement Rate**: The proportion of cases where the difference between model scores and human scores falls within a certain range. This metric reflects how close the model scores are to human scores.

4. **Grading Stability Analysis**: Used to determine whether the model adheres to a unified standard during grading. This includes:

- **Mean Trend Chart**: A line chart showing the trend of mean human scores versus mean machine scores, with the x-axis representing human score ranges and the y-axis representing mean machine scores.
- **Scatter Plot**: A scatter plot showing the relationship between human scores and machine scores, with the x-axis representing human scores and the y-axis representing machine scores. The color of the points indicates the frequency of scores.

Using these evaluation metrics, we can comprehensively assess the performance of the multi-agent grading framework in subjective grading tasks and compare it with the baseline model.

B.2 Experimental Results and Visualizations

Figure 3 shows the visualization of grading results for both models on the two questions, including scatter plots and line charts. From Figure 3, we observe:

- **Scatter Plots**: The scatter plots show the relationship between human scores and machine scores. MASGrader's scatter plots are more concentrated, indicating better agreement with human scores.
- **Line Charts**: The line charts show the trend of mean human scores versus mean machine scores. MASGrader's lines are smoother, indicating better grading stability compared to Single Agent.

B.3 Ablation Study Analysis: Impact of Key Components on Scoring Performance

Necessity of the Logic Validator Agent After removing the logic validation module, the model's WK value decreased by 5% on Question A and 3% on Question B, with slight declines in threshold agreement metrics. This indicates that the logic validator effectively identifies causal reasoning errors in answers, reducing logical misjudgments in grading and contributing to system performance.

Key Role of the Detail Reviewer Agent When the Detail Reviewer Agent was removed, the model's accuracy dropped significantly, with a 19% decrease on Question A and a 10% decrease on Question B. The agreement rates at Th 1 and Th 2 also dropped significantly on Question B, proving that this module captures key details such as keywords and data references through fine-grained semantic analysis, playing a decisive role in grading accuracy.

Global Regulation by the Overview Agent After removing the Overview Agent, the model's grading consistency declined across the board, indicating that this module constructs a macro-level semantic framework for answers, coordinating the local judgments of other agents, and is crucial for maintaining unified grading standards.

The ablation study reveals that the Logic Validator, Detail Reviewer, and Overview Agents form a complementary triangular structure—the Logic Validator ensures reasoning rationality, the Detail Reviewer ensures fine-grained feature capture, and the Overview Agent maintains macro-level standard unity. Through the dynamic coordination of the Supervisor Agent, the framework ultimately achieves grading performance that surpasses single-agent models.