Fusing Graph Logic and LLMs: A Novel Framework for STEM Diagram Evaluation

1, *R.KEERTISH KUMAR, 2,*NANDITHA.N, 3.M.ESHA REDDY
4,*MOHAMMED FAIZAL
Department of Computer Science and Technology, Dayananda Sagar University, Bangalore, India

1,*Corresponding Author

5,*Dr. M. SHAHINA PARVEEN

Department of Computer Science and Technology Dayananda Sagar University, Bangalore, India

ABSTRACT

Evaluating diagrams like flowcharts in STEM education needs careful methods to ensure both structural correctness and logical clarity. Current automated grading systems often focus on text accuracy but miss important logical connections. This research presents a new hybrid framework that merges graph-based logic analysis with large language model (LLM) support for contextual evaluation. By using graph traversal techniques, the system checks the logical flow and structure of diagrams, while LLMs interpret the meaning of the text. This combined method fills existing gaps, improving the accuracy and thoroughness of the grading process. The proposed system shows great promise for enhancing feedback quality, leading to better learning results in STEM fields.

1 INTRODUCTION

The growing use of technology in STEM education has changed how teaching and assessment are done. Automated grading tools, especially those that assess diagrams like flowcharts and schematics, have become essential for reducing manual work and maintaining consistency. However, these tools often struggle to evaluate important aspects of diagrams, such as logical flow and problem-solving methods, which are key to STEM learning.

While improvements in artificial intelligence, especially Large Language Models (LLMs), have allowed for better understanding of diagram parts, they still have limitations in assessing logical connections. On the other hand, graph-based methods are good at analyzing structure and validating logical flow but do not effectively interpret the meaning and context of diagram elements.

To overcome these challenges, this research suggests a combined approach that merges graph-based logic analysis with LLM-supported contextual evaluation. By integrating these two methods, the system aims to create a more thorough evaluation framework that captures both the structural and semantic aspects of STEM diagrams. This development could improve the accuracy of automated grading systems, enhance student understanding, and lead to better teaching methods in STEM education.

2 LITERATURE SURVEY

Paper Name with Year	Authors	Methodology	Drawbacks
Automated Assessment of Multimodal Answer Sheets in the STEM domain (2024)	Rajlaxmi Patil, et al.	Utilizes LLMs to evaluate diagrams and answer sheets in the STEM domain.	Limited to using LLMs for evaluating diagrams, missing logical relationships and flow consistency.
Recognition of Handwritten Flowcharts using Convolutional Neural Networks (2022)	C. David Betancourt Montellano, et al.	Employs CNNs for recognizing handwritten flowcharts.	Focuses on flowchart recognition but lacks advanced logical evaluation of flow consistency.
Online recognition of sketched arrow-connected diagrams (2016)	Martin Bresler, Daniel Prusa, et al.	Recognizes sketched arrow-connected diagrams using diagram recognition methods.	Does not incorporate logical evaluation of the process or decision flows.
[GraphEval: A Knowledge-Graph Based LLM Hallucination Evaluation Framework (2024)]	Hannah Sansford, Nicholas Richardson, Hermina Petric Maretic, Juba Nait Saada	Leverages Knowledge Graphs (KGs) to detect and evaluate hallucinations in LLM outputs, improving explainability and accuracy.	Relies on the quality of KG construction methods, which may limit its effectiveness if the methods are not robust.
[The State of the Art in Empirical User Evaluation of Graph Visualizations (2024)]	Michael Burch, Weidong Huang, Mathew Wakefield, Helen C. Purchase, Daniel Weiskopf, Jie Hua	Classifies literature on graph visualization into graph interpretation, memorability, and creation tasks.	Limited work on memorization and creation aspects; challenges with spatio-temporal evaluation techniques.

3 PROBLEM STATEMENT

Traditional grading methods for diagrams like flowcharts demand extensive teacher involvement to assess both structural correctness and logical integrity. Current automated systems largely focus on textual evaluation, often overlooking logical relationships and flow. This shortfall highlights the need for a more robust hybrid approach that integrates graph-based analysis with LLM-based contextual evaluation, offering a comprehensive and efficient solution to these limitations.

4 PROPOSED METHODOLOGY

To significantly enhance the evaluation of diagrams in education, we propose a hybrid evaluation system that integrates graph-based analysis with LLM-based contextual evaluation. This solution will focus on improving logical flow assessment, error detection, and overall grading accuracy.

5 ARCHITECTURE DIAGRAM

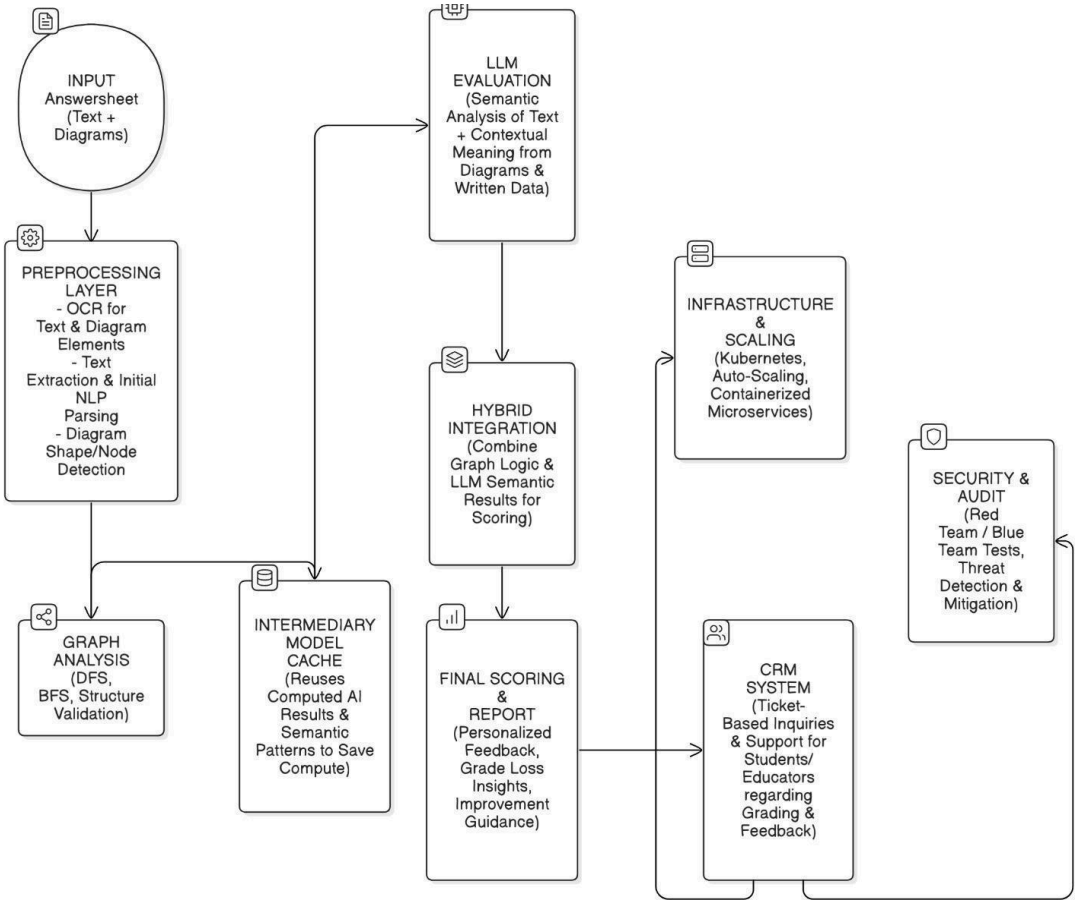

Fig 1. Architecture diagram of proposed system

The proposed system architecture for evaluating answer sheets integrating both text and diagrams leverages a sophisticated multi-layered approach to ensure accuracy and efficiency. The architecture begins with an input layer that collects the answer sheets, followed by a preprocessing layer utilizing Optical Character Recognition (OCR) to extract text and diagrams, alongside initial Natural Language Processing (NLP) for basic parsing. Advanced graph analysis techniques, including depth-first and breadth-first searches, are applied to validate the structure of diagrams. A Large Language Model (LLM) is employed for in-depth semantic evaluation of the textual and graphical elements. To enhance efficiency, an intermediary model cache stores computed AI results, optimizing resource utilization. The hybrid integration layer combines graph logic and LLM results to produce comprehensive scoring, while the final scoring and reporting layer delivers personalized feedback and improvement guidance. The infrastructure is designed to scale with Kubernetes, incorporating auto-scaling and containerized microservices for robust performance. Additionally, the system emphasizes security, implementing thorough testing and threat detection measures, and integrates a Customer Relationship Management (CRM) system for effective handling of inquiries and support. This architecture represents a cutting-edge approach to automating the complex task of evaluating answer sheets, combining multiple advanced technologies to achieve reliable and detailed assessments.

6 IMPLEMENTATION

In the implementation of the proposed system, we first utilize OCR (Optical Character Recognition) and computer vision techniques to convert handwritten diagrams into a structured digital format (JSON). The output JSON contains raw, unorganized data representing the diagram's components. This data is then parsed into nodes and edges, where each node corresponds to an element of the diagram (such as a decision or process), and each edge represents the relationship or flow between these elements.

To improve the organization and structure of the data, we apply a Breadth-First Search (BFS) algorithm. BFS is used to traverse and establish a hierarchy among the nodes and edges, ensuring the logical flow of the diagram is accurately represented. This step is crucial for creating a clear representation of the diagram's structure before further analysis.

Once the diagram's structure is organized, we evaluate it in two stages. Initially, we use only a Large Language Model (LLM) to assess the semantic correctness of the text contained in the diagram, such as labels or descriptions. The LLM helps verify that the diagram's textual content aligns with the intended logical or problem-solving context.

In the second stage, we integrate graph-based analysis into the evaluation process. The graph analysis refines the LLM's output by reweighting the neural network's predictions based on the diagram's logical structure and flow, which was assessed via the graph traversal. This hybrid approach, combining both semantic and structural evaluations, significantly improves the accuracy of automated grading, ensuring that both the logical consistency and contextual meaning of the diagram are considered, leading to better assessment and enhanced learning outcomes.

1.1 Handwritten to JSON Data Conversion

Using Optical Character Recognition (OCR) and computer vision, handwritten diagram data is transformed into a JSON format. This process facilitates easier handling and digital analysis of the content.

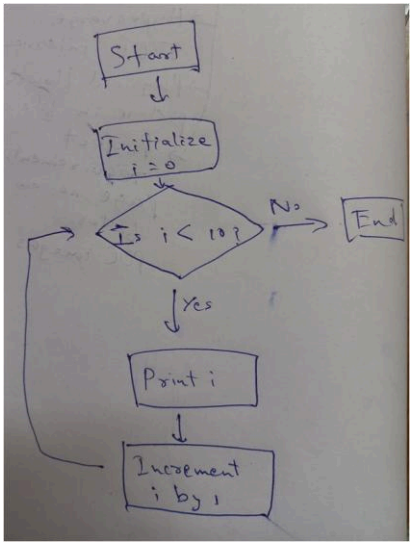

Fig 2. Sample input data which contents a simple loop to display the first 10 digits which is handwritten.

```
{
    "nodes": [
      {"id": "start", "type": "start", "text": "Begin algorithm"},
      {"id": "n1", "type": "process", "text": "Initialize i to 0"},
      {"id": "n2", "type": "decision", "text": "Is i < 10?"},
      {"id": "n3", "type": "process", "text": "Print i"},
      {"id": "n4", "type": "process", "text": "Increment i by 1"},
      {"id": "end", "type": "end", "text": "End algorithm"}
    ],
    "edges": [
      {"from": "start", "to": "n1"},
      {"from": "n1", "to": "n2"},
      {"from": "n2", "to": "n3", "condition": "yes"},
      {"from": "n3", "to": "n4"},
      {"from": "n4", "to": "n2"},
      {"from": "n2", "to": "end", "condition": "no"}
    ]
}
```


Fig 3. Handwritten data to json data format using OCR (Optical Character Recognition) and Computer vision

1.2 Graph Structure Creation

The converted JSON data is parsed to extract nodes and edges, constructing a basic graph structure. Initially, this structure is disorganized, requiring further processing.

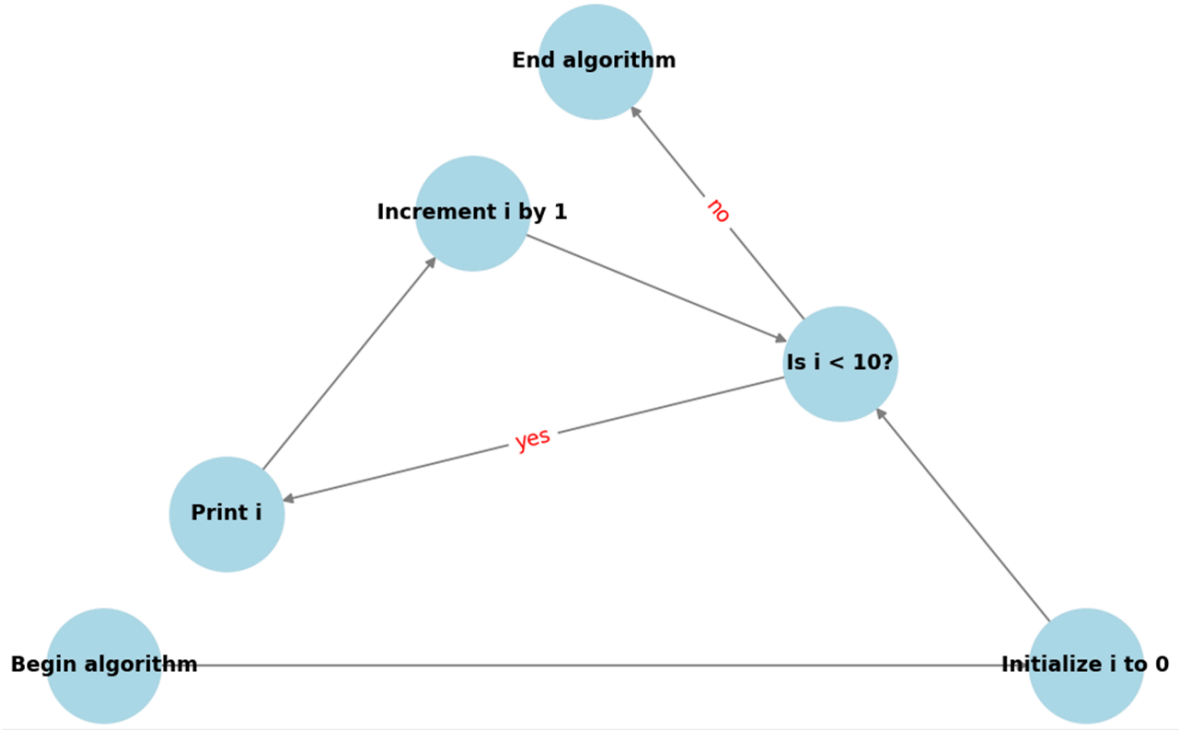

Fig 4. The json data file is converted into nodes and edges. It is currently unorganized.

1.3 Hierarchy Formation

Breadth-First Search (BFS) is employed to organize the graph into a hierarchical structure. This step ensures logical sequencing and improves the clarity of the diagram representation.

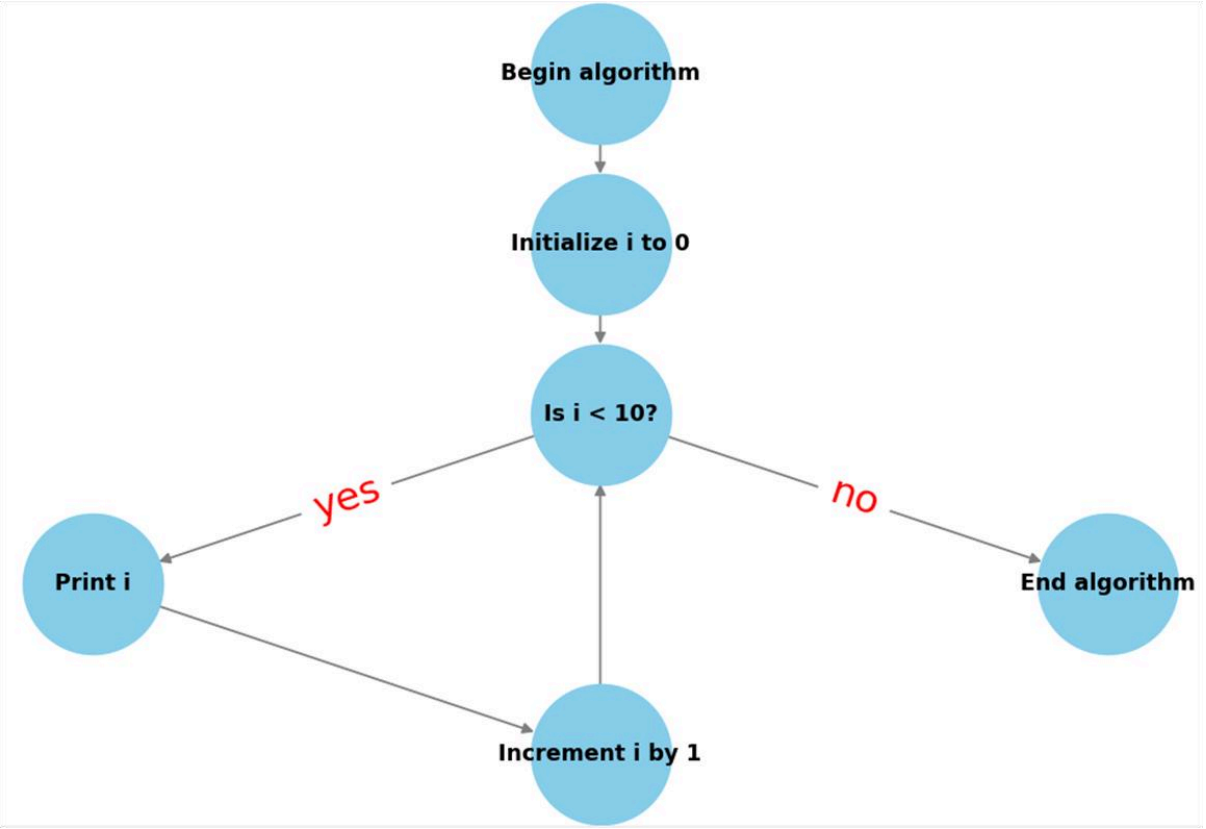

Fig 5. BFS is used to create a hierarchy of the nodes and edges.

1.4 Scoring System

Integrating graph-based logic analysis with LLM-assisted evaluation significantly improves grading outcomes. The system demonstrates a score of 40/100 when relying solely on LLM evaluation, whereas combining it with graph analysis elevates the score to 72/100, showcasing enhanced logical and structural assessments.

7 RESULT

The results of the analysis highlight the effectiveness of combining large language models (LLMs) with graph analysis for evaluating flowcharts. When using only LLMs (`analyze_diagram.py`), the student's flowchart received a semantic score of 40/100, indicating incomplete logic representation. However, the hybrid approach (`analyze_diagram_with_graph_llm.py`) achieved a higher final combined score of 72/100, with structural correctness verified by graph analysis. This demonstrates the importance of verifying both the structure and semantics of flowcharts to achieve accurate assessments. The structural score of 100 further underscores the significance of detailed graph analysis in achieving comprehensive evaluations.

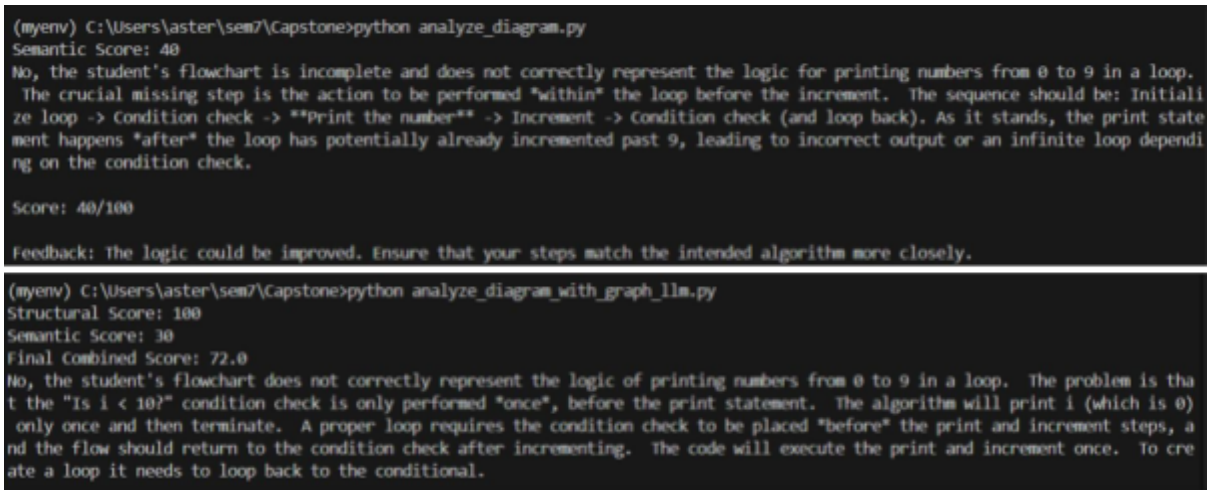

Fig 6. Diagram score is only 40/100 when using only LLM while diagram score is 72/100 when graph analysis verifies the structure

8 CONCLUSION

The proposed hybrid evaluation system effectively bridges the gap between structural and contextual analysis in automated diagram grading. By integrating graph-based logic analysis with LLM-assisted contextual understanding, this approach addresses key limitations of existing methods. The results demonstrate significant improvements in grading accuracy and logical flow assessment, paving the way for enhanced educational outcomes. This system not only supports educators by reducing manual effort but also benefits students through more accurate feedback on their diagrammatic problem-solving skills. Future research could explore real-time implementation and scalability across diverse educational domains.

9 FUTURE ENHANCEMENTS

Future enhancements to this system could focus on further improving accuracy and scalability while seamlessly integrating advanced caching and security measures. Leveraging distributed computing and parallel processing techniques can ensure efficient handling of large volumes of answer sheets, thereby enhancing scalability. Incorporating sophisticated caching mechanisms, such as multi-layer caching and intelligent invalidation strategies, will optimize resource utilization and reduce latency in processing. Furthermore, implementing robust security frameworks, including advanced encryption, access controls, and real-time threat detection, will safeguard sensitive data and maintain the integrity of the evaluation process. By integrating these concepts from the architecture diagram, the system can achieve a higher level of precision and reliability, ensuring comprehensive and secure assessments at scale.

10 REFERENCES

1. Rajlaxmi Patil, Aditya Ashutosh Kulkarni, Ruturaj Ghatage, Sharvi Endait, Dr. Geetanjali Kale, Raviraj Joshi. "Automated Assessment of Multimodal Answer Sheets in STEM." arXiv preprint arXiv:2409.15749, 2024. Available: https://arxiv.org/abs/2409.15749
2. C. David Betancourt Montellano, C. Onder Francisco Campos Garcia, Roberto Oswaldo Cruz Leija. "Recognition of Handwritten Flowcharts using Convolutional Neural Networks." International Journal of Computer Applications, Vol. 184, Issue 1, 2022. Available: https://www.ijcaonline.org/archives/volume184/number1/2921969-2022921
3. Martin Bresler, Daniel Pruša, Vaclav Hlavac. "Online Recognition of Sketched Arrow-Connected Diagrams." International Journal on Document Analysis and Recognition (IJDAR), Vol. 19, 2016. Available: https://link.springer.com/article/10.1007/s10032-016-0269-z
4. Sansford, Hannah, et al. "Grapheval: A knowledge-graph based llm hallucination evaluation framework." arXiv preprint arXiv:2407.10793 (2024).
5. Burch M, Huang W, Wakefield M, Purchase HC, Weiskopf D, Hua J. The state of the art in empirical user evaluation of graph visualizations. IEEE Access. 2020 Dec 28;9:4173-98.
6. Liu EZ, Yuan D, Ahmed A, Cornwall E, Woodrow J, Burns K, Nie A, Brunskill E, Piech C, Finn C. A Fast and Accurate Machine Learning Autograder for the Breakout Assignment. InProceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 2024 Mar 7 (pp. 736-742).
7. Kumar A, Pandey A, Gadia R, Mishra M. Building knowledge graph using pre-trained language model for learning entity-aware relationships. In2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON) 2020 Oct 2 (pp. 310-315). IEEE.
8. Min S, Krishna K, Lyu X, Lewis M, Yih WT, Koh PW, Iyyer M, Zettlemoyer L, Hajishirzi H. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. arXiv preprint arXiv:2305.14251. 2023 May 23.
9. Mündler N, He J, Jenko S, Vechev M. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. arXiv preprint arXiv:2305.15852. 2023 May 25.
10. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. InProceedings of the 40th annual meeting of the Association for Computational Linguistics 2002 Jul (pp. 311-318).