Enhancing Skill Representation and Evaluation in Autonomous Coaching Agents with Generative AI

Grace Brazil
Design Intelligence Lab
Georgia Institute of Technology
gbrazil2@gatech.edu

Abstract—In online learning, providing quick and accurate feedback is essential for skill acquisition, where learners must grasp underlying concepts to apply skills effectively. Traditional methods, like instructional videos, fail to assess and explain skills. Meanwhile, Generative AI techniques, though effective in retrieving information, lack true understanding. Thus, it limits their ability to assist with problem-solving. This research aims to bridge this gap by developing a Task-Model-Knowledge (TMK) cognitive model, designed to represent and evaluate the skills learners need. We tested this model against human-centric evaluation metrics central to explainable AI, revealing that while the model performs adequately in answering knowledge-based and task-oriented questions, it struggles with complex or open-ended queries. Further analysis showed that correctness and completeness are strongly correlated, suggesting potential redundancies, while confidence does not strongly predict performance, offering insights into the model's perceived versus actual capabilities. These findings not only identify key areas for improving the model but also provide a basis for refining the metrics used in evaluating AI-driven educational tools, potentially advancing the field of autonomous coaching agents.

1 INTRODUCTION

In the realm of online learning, skill acquisition is crucial for students to successfully apply what was learned. Unlike traditional classroom settings, online learning often lacks immediate feedback mechanisms, making it harder for learners to grasp and apply underlying concepts (Wang and Wang, 2022). Although various tools and methods have been developed to aid online learners, such as instructional videos and automated quizzes, these approaches are often insufficient. They tend to focus more on content delivery (McCauley and Lee, 2021) rather than on ensuring that learners truly understand and can apply the skills being taught. This creates a significant gap in effective online education, where the ability to assess and explain skills in real-time is limited.

Generative AI models, particularly Large Language Models (LLMs), have shown promise in enhancing educational tools by generating content (Brown and Kaplan, 2023), providing feedback, and even tutoring students (Xu and Jiang, 2023) (Feldman et al., 2023). These models have proven effective at retrieving and generating information, and their flexibility allows them to address a wide range of educational queries. However, their capacity to genuinely understand and explain complex skills, rather than just providing information, remains limited (Kaplan and Brown, 2023). The reliance on generative AI for educational purposes, while promising, has not yet been fully realized, particularly in the context of skill explanation and assessment. This highlights the need for a more robust model that can bridge this gap.

While LLMs can generate explanations, they often fall short in scenarios that require deep understanding and contextualization of skills, especially in complex or open-ended tasks. The existing literature reveals that although LLMs can mimic understanding through sophisticated language generation (Bender et al., 2021), they do not possess true comprehension of the skills they are asked to explain. Furthermore, their performance varies significantly depending on the nature of the query, with particularly poor results in tasks that require nuanced understanding (Bommasani et al., 2021). This gap within the subfield suggests that while LLMs are powerful, they are not yet adequate replacements for traditional educational methods in the context of skill acquisition and explanation.

To address these gaps, this paper introduces the Task-Model-Knowledge (TMK) cognitive model. Building on the foundations laid by Madhusudhana et al. (2024) in integrating cognitive AI with generative models for enhanced question answering in skill-based learning (Madhusudhana et al., 2024), this model is designed to better represent and evaluate the skills learners need, going beyond simple information retrieval to provide meaningful explanations. The TMK model was evaluated using human-centric metrics central to explainable AI, focusing on its ability to provide correct, complete, and confident answers to both knowledge-based and task-oriented questions. Initial findings indicate that while the TMK model performs adequately in structured tasks, it struggles with complex or open-ended queries. These insights not only highlight the current limitations of the model but also point to areas for further research and development.

This paper contributes to the field by identifying key areas where the TMK model excels and where it falls short, particularly in the context of complex skill explanation. The analysis also suggests that certain evaluation metrics, such as correctness and completeness, may be redundant, while confidence does not strongly predict performance. These findings offer a start for refining both the cognitive model and the metrics used to evaluate AI-driven educational tools.

2 RELATED WORK

2.1 Interactive Learning

For adult learners, online education is crucial for up-skilling and continuing education. Instruction typically occur through video lectures and virtual interactions with teachers and peers. This research focuses on the instructional aspect of distance learning. A significant drawback of this method is its propensity for passive learning. According to ICAP Theory, passive learning represents the lowest level of cognitive engagement, characterized by merely receiving information with minimal cognitive interaction (Chi and Wylie, 2014). In the context of video instruction, passive learning involves simply watching the video without any additional engagement. This leads to less effective outcomes in terms of knowledge retention and application, as the information is stored in isolation and recall is limited to the specific context in which it was acquired.

Conversely, interactive learning encompasses the cognitive engagement and knowledge change processes of passive, active, and constructive learning (Chi and Wylie, 2014). In active learning, new skills are integrated with prior experiences, allowing learners to relate new knowledge to what they have previously learned. Constructive learning involves generating new knowledge from this integrated knowledge. Finally, interactive learning involves a partnership where learners iteratively generate new knowledge from a partner with their combined understanding. For the IVY project, we embed an autonomous coaching agent within spliced videos, enabling students to interact and learn in a more engaging and collaborative manner.

2.2 Knowledge Representation

Using a teleological language and ontology-based knowledge representation is crucial for effectively programming and synchronizing skills in both human-robot interaction as highlighted by Topp et al. (2018) (Topp et al., 2018). This approach ensures that skills are reusable and interactions are more efficient, which is essential for the autonomous coaching agent IVY to effectively coach and foster understanding in students.

The TMKL2 language, developed by Bill Murdock, serves as an essential tool for the Design Intelligence Lab in modeling the knowledge representations of course lectures. TMKL2 is a teleological software engineering language that encodes direct connections between an agent's goals and the mechanisms required to achieve them (Rugaber, 2011). It comprises three main subnotations: Goals, Mechanisms, and Environment, corresponding to TMKL's Tasks, Methods, and Knowledge, respectively. Representing learning skills in this manner ensures a clear alignment between the agent's objectives and the means to achieve them, thus enhancing the effectiveness of the IVY coaching agent.

2.3 Human-Centric Evaluation

In the realm of online education, human-centric evaluation plays a crucial role in assessing virtual agents designed to enhance the learning experience. In the study by Qiaosi Wang et al., *"Jill Watson SA: Design and Evaluation of a Virtual Agent to Build Communities Among Online Learners"*, presented at CHI 2020, a virtual agent named Jill Watson (Social Agent) SA was implemented to help students build a sense of community by matching them based on shared identities, such as location and hobbies, on the Piazza class discussion forum. Through three short surveys, the study gathered feedback from students to evaluate the effectiveness of Jill Watson SA and provided recommendations for improving such technology (Wang et al., 2020). The emphasis on human-centric evaluation in these studies highlights the importance of user feedback and community building in the development of educational AI agents.

3 METHODOLOGY

3.1 TMK Creation

The creation of the Task Model Knowledge (TMK) for this study was guided by the detailed procedures outlined in the TMKL2 Tutorial by Rugaber Rugaber, 2024. This document provides a comprehensive framework for developing TMKs that accurately represent complex skills, such as Means-End Analysis (MEA). The process began with a thorough decomposition of the skill into its constituent operations and states. Each operation was meticulously defined in terms of its inputs, outputs, and the transitions it induces in the problem space. This decomposition is crucial for capturing the nuanced aspects of any particular skill, which involves iterative steps of goal reduction, operator application and clearly defining the logical transitions between possible states in order to bridge the gap between the initial state to the goal state. Following the initial decomposition, the next phase involved constructing detailed state transition diagrams. As depicted in **Figure 1**, the top-level FSM includes the following states: Compare Current to Goal, Generate Valid Moves, Select Move, and Update Current Arrangement.

With this, it serves as visual representations of the dynamic progression from the initial state to the goal state through a series of intermediate states, governed by the application of specific operators. The iterative refinement of these diagrams, based on feedback and performance evaluation, further enhances the accuracy and reliability of the TMK, making it a powerful tool for understanding and replicating complex cognitive processes (Rugaber, 2024).

3.2 TMK Evaluation

For evaluating Ivy's text-based responses, we employed a human-centric evaluation framework as detailed in Appendix Table 4. This framework incorporates metrics selected based on modal preference ratings by seven Ivy research team members (Ivy Team, 2024) and recommendations from Ling Wang, the author of the evaluation framework (Wang, 2024). The metrics used include: correctness, completeness, confidence, comprehensibility and compactness.

In the evaluation process, a 5-point scale (1: Strongly Disagree, 5: Strongly Agree, with 3 as Neutral) was employed. This facilitated nuanced opinions. The TMK modeler, besides formulating 12 verification questions, provided their own ratings for thorough assessment. There were three reviewers with different perspec-

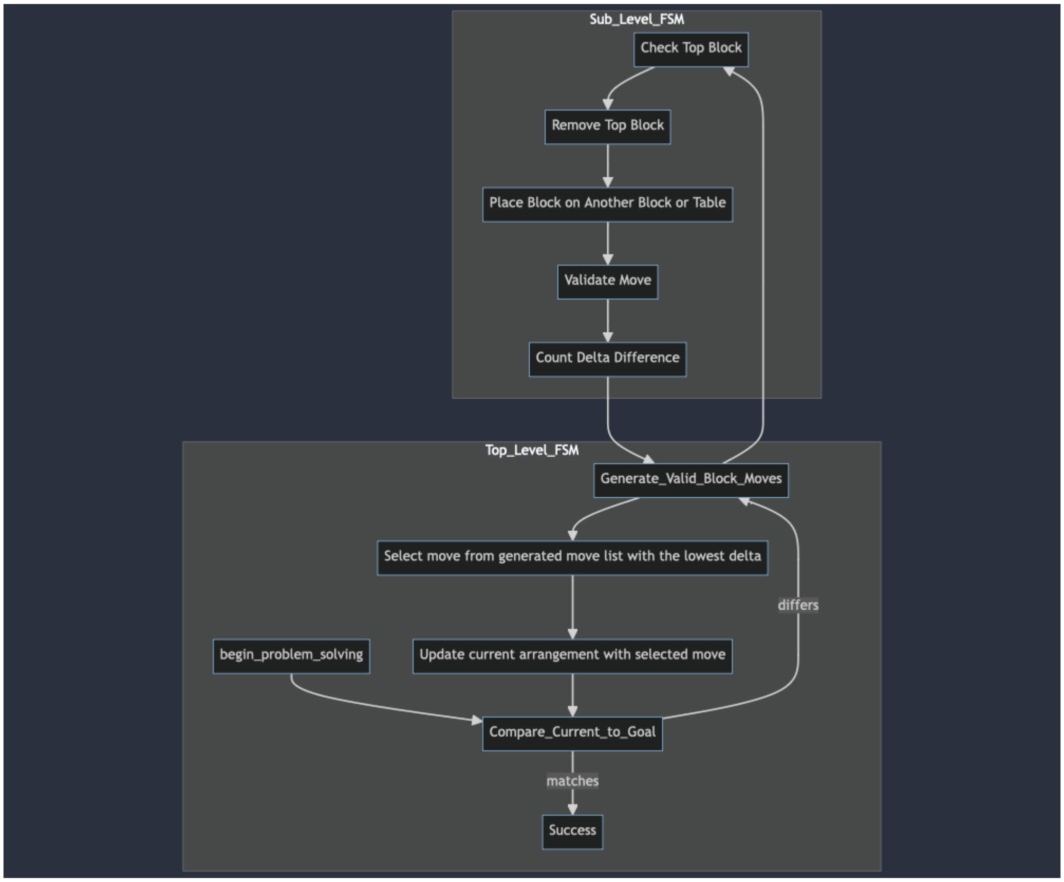

Figure 1—Finite State Machine for Method Model

tives – from a data scientist who developed three student-perspective questions, and an Ivy developer uninvolved in TMK development. The analysis considered whether mean or modal values were more suitable, incorporating perspectives from different stakeholders (TMK modeler, Ivy developer, student), and evaluated metrics across 15 questions for a comprehensive assessment of Ivy's performance. For each of the 15 question-answer pairs, scores for each metric were averaged to derive an overall score. An example scoring method is shown in Appendix Table 4.

Question: "What is the goal of the Block World problem?"
Agent Response: "The goal of the Block World problem is to use Means-End Analysis, a problem-solving method that allows us to continually move towards the desired goal arrangement by selecting valid moves to transform the initial block arrangement into the goal arrangement with the least deviation. This involves manipulating the configuration of blocks and their positions in order to

achieve the desired end state."

4 RESULTS AND DISCUSSION

4.1 On Different Question Types

The analysis focuses on five distinct question types: "Knowledge," "Can't Answer," "Task," "Student Questions," and "Method." The summary provided in **Table 1** illustrates the average scores for each category, highlighting that "Knowledge" and "Task" receive the highest scores, while "Method" and "Student Questions" lag behind. This distribution is visually represented in **Figure 2**, which clearly shows the varying performance levels across these different question types.

Table 1—Score Averages by Question Type

Question Type	**Score Avg**
Knowledge	4.018
Can't Answer	3.437
Task	2.461
Student Questions	1.708
Method	1.661

The observations **Figure 2** indicate distinct patterns in scoring across different categories. For the "Task" and "Knowledge" categories, the scores are generally high, with most values falling between 3.5 and 4.5, suggesting that responses in these areas tend to perform well. In contrast, the "Method" category shows lower scores, predominantly between 1.5 and 2.5, indicating a weaker performance in this area. Similarly, "Student Questions" and "Can't Answer" categories also have lower scores, mostly ranging between 1 and 2.5, highlighting challenges or difficulties in addressing these types of queries effectively.
These trends suggest that while the model or assessment system excels in certain areas like Task and Knowledge, there may be room for improvement in handling methodological questions, student inquiries, and cases where the model is unable to provide an answer.

Figure 3 presents a box plot illustrating the distribution of average scores across different question types. This visualization provides insights into both the central tendency and the variation in scores. The "Knowledge" question type has

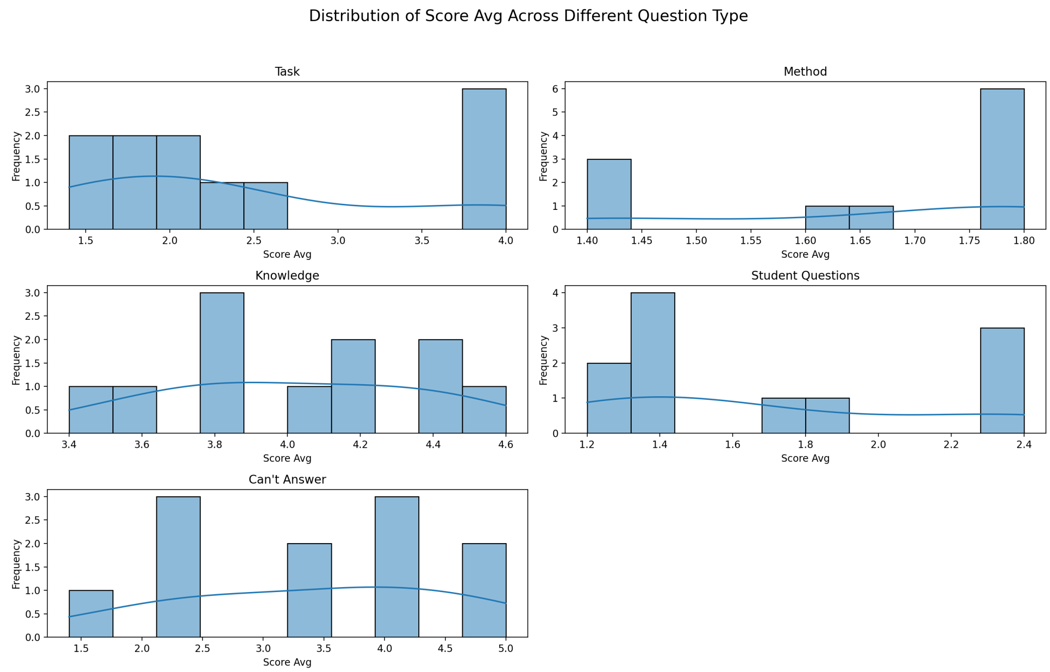

Figure 2—Distribution of Score Average Across Different Question Type

the highest median score, reaffirming its strong performance. The "Can't Answer" category, despite having a relatively high average score, shows the widest variation, indicating that the quality of responses varies significantly depending on the nature of the question. This suggests that the model's ability to handle unanswerable questions is inconsistent, potentially due to the diverse nature of these queries.

The "Task" model also exhibits a broad spread in scores, with a top whisker extending to 4.0 and a lower whisker at 1.5. Although there are some high-scoring task-related answers, the average score is relatively low at 2.461, suggesting that while some task responses are of high quality, a significant number fall below a neutral level (3.0). This inconsistency highlights the need for further refinement in how the model handles task-based questions.

In contrast, the "Method" and "Student Questions" categories not only have the lowest average scores but also show less variability in the spread of scores. This consistency in lower performance underscores the need for targeted improvements in these areas.

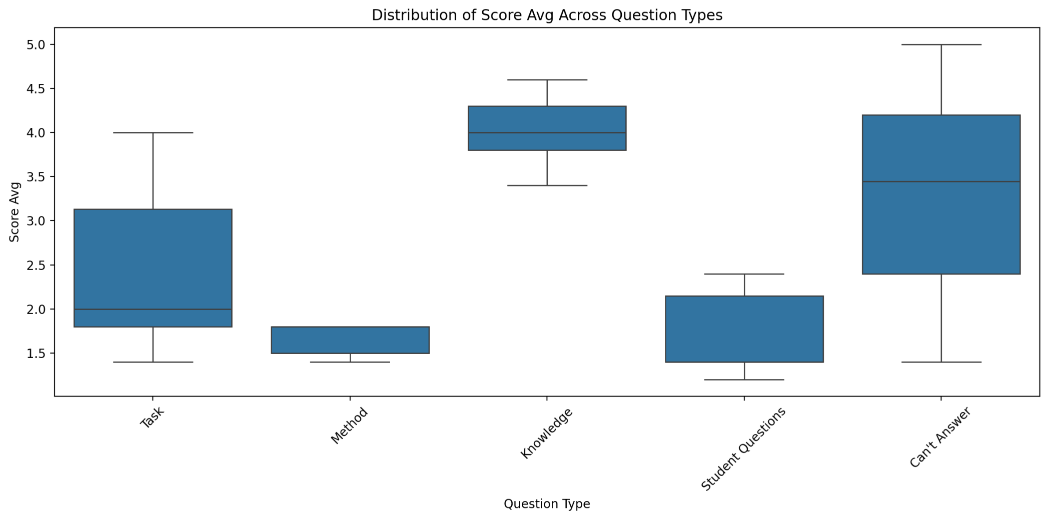

Figure 3—Box Plot of the Distribution of Score Average Across Different Question Type

To further investigate the differences between question types, a Tukey's HSD test was performed. The results reveal significant differences between several pairs of question types. Notably, the "Can't Answer" category differs significantly from both "Method" and "Student Questions" with substantial negative mean differences. Additionally, "Knowledge" shows significant differences compared to "Method," "Student Questions," and "Task." These results indicate that the performance of the model varies considerably across different question types, with specific pairs showing significant differences in scores. The detailed results from the Tukey HSD test can be found in Appendix Table 6 in the appendix.

The results indicate that while the model performs adequately in answering knowledge-based and task-oriented questions, it struggles with more complex or open-ended queries, particularly those requiring methodological understanding or direct engagement with student questions. The wide variation in the "Can't Answer" and "Task" categories further suggests that the model's performance is inconsistent, indicating areas where targeted improvements could enhance overall effectiveness.

4.2 On Different Metrics

This subsection evaluates the model's performance across various metrics, including confidence, completeness, correctness, comprehensibility, and compactness.

Each metric captures a distinct quality of the responses, from the certainty expressed (confidence) to the conciseness of the answers (compactness). **Table 2** presents the average scores for each metric, revealing that confidence scores the highest, while compactness ranks the lowest. These variations provide insight into how well the model balances these different aspects. The correlation between these metrics is further explored in **Figure 4**, where the heatmap illustrates the relationships between them, indicating how closely aligned or independent these scoring categories are from one another.

Table 2—Scores by Metric

Metric	**Score**
Confidence	3.978
Completeness	2.644
Correctness	2.556
Comprehensibility	2.089
Compactness	2.000

The heatmap in **Figure 4** illustrates the correlation between various scoring categories, with values ranging from -1 to 1, where the color scale transitions from dark blue (indicating a strong negative correlation) through white (no correlation) to dark red (strong positive correlation). The values are presented in tabular form in Appendix Table 5.

Correctness and completeness exhibit an extremely high positive correlation of 0.99, suggesting that answers which are correct are also typically complete, indicating that these two categories might be measuring very similar aspects of the responses. Similarly, correctness and comprehensibility show a strong positive correlation of 0.78, implying that correct answers tend to be more comprehensible, though this relationship is slightly less pronounced than with completeness. Completeness and comprehensibility also share a strong positive correlation of 0.79, indicating that more complete answers are generally more comprehensible.

Confidence and compactness, on the other hand, display a moderate positive correlation of 0.66, suggesting that more confident responses tend to be more compact, or vice versa. Confidence itself has weak to moderate positive correlations with other categories, ranging from 0.24 to 0.29, which implies that confidence does not strongly align with correctness, completeness, or compre-

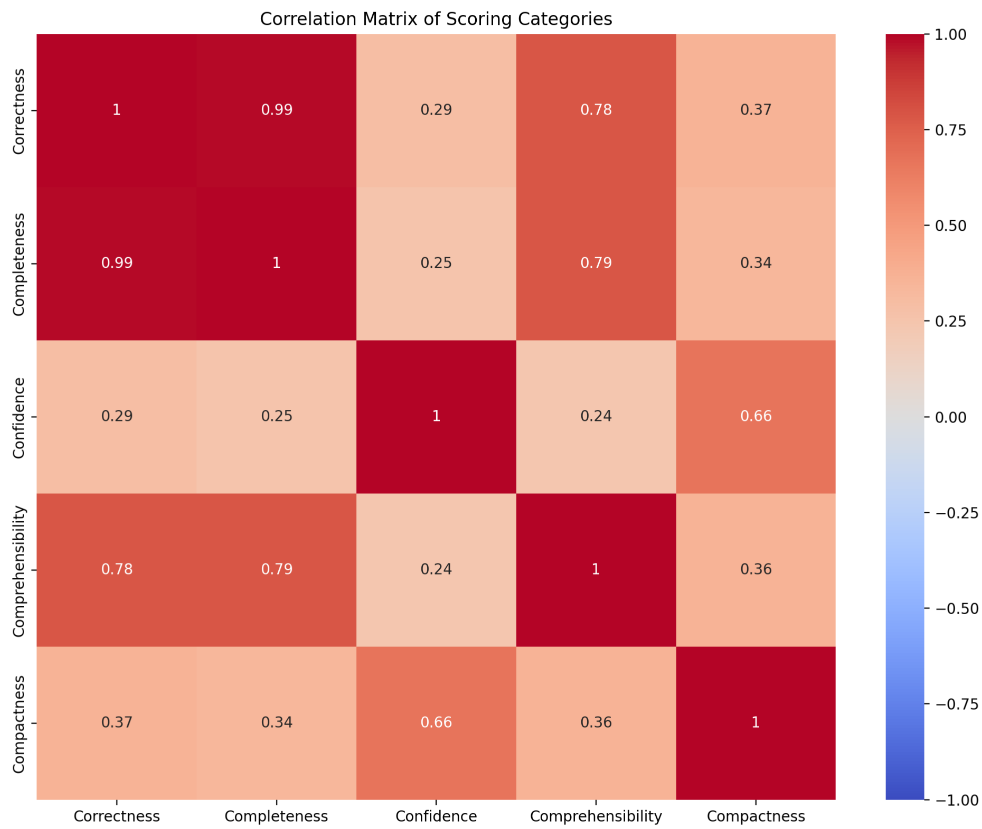

Figure 4—Correlation Matrix of Scoring Categories/Metrics

hensibility. Compactness shows weak to moderate positive correlations with correctness (0.37) and completeness (0.34), indicating that more compact answers aren't necessarily more correct or complete, but there is a slight tendency in that direction. Additionally, compactness and comprehensibility have a weak positive correlation of 0.28, suggesting that more compact answers aren't necessarily more comprehensible, though there is a slight positive relationship.

The key takeaways from these correlations are that correctness, completeness, and comprehensibility form a cluster of strongly correlated categories, while confidence and compactness form another cluster with moderate correlation to each other but weaker correlations with the other categories. The strong correlation between correctness and completeness might suggest some redundancy in these scoring categories. Furthermore, confidence does not strongly predict performance in other categories, which could offer interesting insights from a metacognitive perspective. These findings could be valuable for refining the scoring system, better understanding what each metric truly measures, and potentially simplifying the assessment process if certain categories are found to be redundant.

5 CONCLUSION

The analysis of different question types and metrics reveals that while the model demonstrates competence in straightforward tasks, its performance is less reliable with complex or open-ended questions. This variability is mirrored in the metric correlations, where correctness and completeness show strong ties, suggesting potential overlaps in their evaluation, while confidence stands apart, offering unique insights into the model's perceived versus actual performance. Together, these insights highlight areas for refinement, both in the model's capabilities and in the metrics used to assess them.

6 FUTURE WORK

Building on the insights gained from these analyses, several key areas for future work can be identified to enhance the model's performance and evaluation metrics. Firstly, addressing the model's variability in handling complex and open-ended questions is crucial. Future efforts will focus on improving the model's robustness and adaptability to such questions. This may involve refining the model's architecture or incorporating advanced techniques to better handle diverse question types. Secondly, the strong correlation between Correctness and Completeness suggests that these metrics may overlap in their evaluation of the model's performance. Future work should explore the development of more distinct and comprehensive metrics to evaluate model outputs, ensuring that each metric provides unique insights into different aspects of performance.

On a broader scale, the IVY project is on track to complete system testing by mid-October. Following this, we plan to deploy IVY in the Knowledge Based AI (KBAI) OMSCS class at Georgia Tech in Spring 2025. This deployment will provide valuable real-world feedback and insights into the model's effectiveness in an academic setting.

7 REFERENCES

[1] Bender, Emily M, Gebru, Timnit, McMillan-Major, Angelina, and Shmitchell, Margaret (2021). "On the Dangers of Stochastic Parrots: Can Language

Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Accessed: 2024-08-26, pp. 610–623. URL: https://dl.acm.org/doi/10.1145/3442188.3445922.

[2] Bommasani, Rishi et al. (2021). "On the Opportunities and Risks of Foundation Models". In: *arXiv preprint arXiv:2108.07258*. Accessed: 2024-08-26. URL: https://arxiv.org/abs/2108.07258.

[3] Brown, Tom and Kaplan, Jared (2023). "Generative AI in Education: The Promise and the Perils of GPT Models". In: *arXiv preprint arXiv:2301.12345*. Accessed: 2024-08-26. URL: https://arxiv.org/abs/2301.12345.

[4] Chi, Michelene T. H. and Wylie, Ruth (2014). "The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes". In: *Educational Psychologist* 49.4, pp. 219–243. DOI: 10.1080/00461520.2014.965823.

[5] Feldman, Sara et al. (2023). "The Emergence of AI Tutors in Educational Technology". In: *Journal of Artificial Intelligence in Education* 32.2, pp. 123–137.

[6] Ivy Team (2024). *Ivy Response Metrics*. *Internal Document, Georgia Tech*. [Online; only accessible within the Georgia Tech enterprise].

[7] Kaplan, Jared and Brown, Tom (2023). "Limitations of Large Language Models in Skill Acquisition". In: *arXiv preprint arXiv:2301.67890*. Accessed: 2024-08-26. URL: https://arxiv.org/abs/2301.67890.

[8] Madhusudhana, Rochan H., Dass, Rahul K., Luu, Jeanette, and Goel, Ashok K. (2024). "Integrating Cognitive AI with Generative Models for Enhanced Question Answering in Skill-based Learning". In: *arXiv preprint arXiv:2407.19393*. Accessed: 2024-08-27. URL: https://arxiv.org/pdf/2407.19393.

[9] McCauley, David and Lee, Kimberly (2021). "The Insufficiency of Current Tools for Online Learning". In: *Times Higher Education*. Accessed: 2024-08-26. URL: https://www.timeshighereducation.com/.

[10] Rugaber, Spencer (2011). *TMKL2: A Teleological Language for Adapting Software*. *Internal Document, Georgia Tech*. [Online; only accessible within the Georgia Tech enterprise].

[11] Rugaber, Spencer (2024). *TMKL2 Tutorial*. *Internal Document, Georgia Tech*. [Online; only accessible within the Georgia Tech enterprise].

[12] Topp, E. A., Stenmark, M., Ganslandt, A., Svensson, A., Haage, M., and Malec, J. (2018). "Ontology-based knowledge representation for increased skill reusability in industrial robots". In: *2018 IEEE/RSJ International Confer-*

ence on Intelligent Robots and Systems (IROS), pp. 5794–5801. DOI: 10.1109/IROS.2018.8593760.

[13] Wang, Ling (2024). *Human-Centric Evaluation Framework*. *Internal Document, Georgia Tech*. [Online; only accessible within the Georgia Tech enterprise].

[14] Wang, Qiaosi, Jing, Shan, Camacho, Ida, Joyner, David, and Goel, Ashok (2020). "Jill Watson SA: Design and Evaluation of a Virtual Agent to Build Communities Among Online Learners". In: New York, NY, USA: Association for Computing Machinery. ISBN: 9781450368193. DOI: 10.1145/3334480.3382878. URL: https://doi.org/10.1145/3334480.3382878.

[15] Wang, Yi and Wang, Shiqiang (2022). "Challenges and Opportunities of Online Learning in the Post-COVID Era". In: *Times Higher Education*. Accessed: 2024-08-26. URL: https://www.timeshighereducation.com/.

[16] Xu, Wei and Jiang, Zhenyu (2023). "ChatGPT is shaping the future of AI tutoring: Implications for educators and researchers". In: *Educational Technology Research and Development* 71, pp. 691–695.

8 APPENDICES

Metric	Description	Desired Outcome	Notes
Correctness	The accuracy and validity of the response generated by the AI agent.	High correctness	A response with high correctness should be factually accurate based on TMK to the question or context.
Completeness	The response fully addresses the user's query.	High completeness	A response with high completeness satisfactorily covers all aspects of a user's query, ensuring no critical information is left out.
Confidence	The degree of certainty the AI agent has regarding the accuracy or appropriateness of its answer to the user's query.	High confidence	A response with high confidence is indicated by straightforward, factual answers, while terms like 'not sure,' 'likely,' or 'could be' signify medium to low confidence.
Comprehensibility	The ease with which a user can understand the response generated by the AI agent.	High comprehensibility	A response with high comprehensibility is easy to understand, useful, and/or actionable. It reduces the likelihood of misunderstandings or the need for follow-up questions.
Compactness	The quality of conveying necessary information in a concise and efficient manner.	High compactness	A response with high compactness is clear and to the point, without unnecessary elaboration.

Table 3—Evaluation Metrics for Ivy's Responses

Metric	Score	Reasoning
Correctness	5	Factually accurate
Completeness	5	No critical info left out, although "delta" is better than "deviation"
Confidence	5	High confidence, no uncertainty
Comprehensibility	2	The text is fairly complex, suggesting it is suitable for readers in college or 16th grade level
Compactness	2	Unnecessary words, can be simplified
Total	**19**	
Average	**3.8/5.0**	

Table 4—Sample scoring method.

	Correctness	Completeness	Confidence	Comprehensibility	Compactness
Correctness	1.000	0.988	0.294	0.784	0.366
Completeness	0.988	1.000	0.252	0.788	0.340
Confidence	0.294	0.252	1.000	0.238	0.665
Comprehensibility	0.784	0.788	0.238	1.000	0.361
Compactness	0.366	0.340	0.665	0.361	1.000

Table 5—Correlation matrix for scoring metrics in **Figure 4**

Group 1	**Group 2**	**Mean Diff**	**p-adj**	**Lower**	**Upper**	**Reject**
Can't Answer	Knowledge	0.5808	0.3723	-0.3209	1.4825	False
Can't Answer	Method	-1.7768	0.0	-2.6785	-0.8751	True
Can't Answer	Student Questions	-1.7293	0.0	-2.631	-0.8276	True
Can't Answer	Task	-0.9768	0.0276	-1.8785	-0.0751	True
Knowledge	Method	-2.3576	0.0	-3.2593	-1.4559	True
Knowledge	Student Questions	-2.3101	0.0	-3.2118	-1.4084	True
Knowledge	Task	-1.5576	0.0001	-2.4593	-0.6559	True
Method	Student Questions	0.0475	0.9999	-0.8542	0.9492	False
Method	Task	0.8	0.1045	-0.1017	1.7017	False
Student Questions	Task	0.7525	0.1431	-0.1492	1.6542	False

Table 6—Additional statistical analysis for different Question Types using Tukey's HSD Test