Combining Cognitive and Generative AI for Self-explanation in Interactive AI Agents

Shalini Sushri[1,*], Rahul K. Dass[1], Rhea Basappa[1], Hong Lu[2,†] and Ashok K. Goel[1]

[1]Georgia Institute of Technology, Atlanta, GA, USA
[2]Tufts University, Medford, MA, USA

Abstract

The Virtual Experimental Research Assistant (VERA) is an inquiry-based learning environment that empowers a learner to build conceptual models of complex ecological systems and experiment with agent-based simulations of the models. This study investigates the convergence of cognitive AI and generative AI for self-explanation in interactive AI agents such as VERA. From a cognitive AI viewpoint, we endow VERA with a functional model of its own design, knowledge, and reasoning represented in the Task-Method-Knowledge (TMK) language. From the perspective of generative AI, we use ChatGPT, LangChain, and Chain-of-Thought to answer user questions based on the VERA TMK model. Thus, we combine cognitive and generative AI to generate explanations about how VERA works and produces its answers. The preliminary evaluation of the generation of explanations in VERA on a bank of 66 questions derived from earlier work appears promising.

Keywords

Self-explanation, AI Agents, Combining Cognitive, Generative AI, Theory of Mind

1. Introduction

1.1. Self-Explanation in Interactive AI Agents

Interactive AI agents with self-explanation capabilities foster understanding, transparency, and trust in users across a wide range of domains and applications [1, 2]. By self-explanation, we mean Interactive AI agents that can explain their reasoning and behaviors. By generating human-understandable explanations, self-explainable AI can enhance user learning and trust [3]. Studies have shown the benefits of self-explanation in multimedia learning environments, facilitating intrinsic motivation, visual processing, and learning outcomes [4]. Additionally, emerging methods leveraging situation awareness holds promise for generating explanations of autonomous agents' behaviors, ultimately improving trust and comprehension [5].

This research contributes to the goal of enhancing user trust and learning through self-explanation in the Virtual Experimental Research Assistant (VERA; [6, 7]), an interactive learning environment for inquiry-based learning. In this paper, we explore how VERA explains its internal workings to users, potentially fostering trust and enhancing the learning experience.

1.2. VERA: Inquiry-based Modeling

VERA (http://vera.cc.gatech.edu) is an interactive learning environment for supporting inquiry-based learning. It helps learners construct conceptual models of ecological systems and evaluate them through agent-based simulations. VERA is an AI agent because of three capabilities. First, it uses an ontology of the ecology domain in the representation and construction of conceptual models. Second, it automates the retrieval of species' and related ecological relations' information from the Smithsonian Institute's Encyclopedia of Life (EOL; https://eol.org/), a comprehensive digital library of biodiversity [8], and automatically inserts parameter values for the agent-based simulations. Third, it automatically compiles the conceptual model into agent-based simulations in NetLogo [9]. Thus, this platform aligns with the research focus on self-explanation in educational AI assistants.

1.3. Cognitive and Generative AI Convergence

This research explores the potential of combining Cognitive AI and Generative AI approaches for self-explanation capabilities in VERA. Cognitive AI is centered around understanding human cognitive processes and developing cognitively-inspired AI agents, while Generative AI methods demonstrate powerful capabilities for various natural language processing tasks like entity recognition, intent classification, and question-answering based on a text corpus [10].

2. Related Work

Early research on self-explanation in Interactive AI agents highlighted the importance of explicitly representing the agent's knowledge of its design [11, 12]. This explicit representation allows the generation of explanations about the tasks the agent performs, the domain knowledge it uses, and the methods it applies. This led to the questions of how to effectively identify, acquire, represent, store, access, and use this design knowledge for generating explanations in interactive agents [13, 14]. One solution lies in viewing the AI agent as an abstract device, equipping it with meta-knowledge about its design, and enabling it to introspect and generate explanations based on its understanding of its structure, behaviors and functions [12].

There has been ongoing research into an Interactive AI agent's ability to provide self-explanation [15, 16]. In prior work on the Skillsync project for skill-based linking employers and colleges preparing prospective employees [17], we used a Task-Method-Knowledge model of Skillsync to generate explanations of its reasoning and recommendations

Appears on the Website of Human-Centric eXplainable AI in Education (HEXED) Workshop held in conjunction with the Seventeenth International Conference on Educational Data Mining (EDM), July 2024.
*Corresponding author.
†Work done as a Research Scientist at Georgia Institute of Technology.
ssushri3@gatech.edu (S. Sushri); rdass7@gatech.edu (R. K. Dass); rb324@gatech.edu (R. Basappa); hlu07@tufts.edu (H. Lu); ashok.goel@cc.gatech.edu (A. K. Goel)
https://rkdass.github.io/ (R. K. Dass)

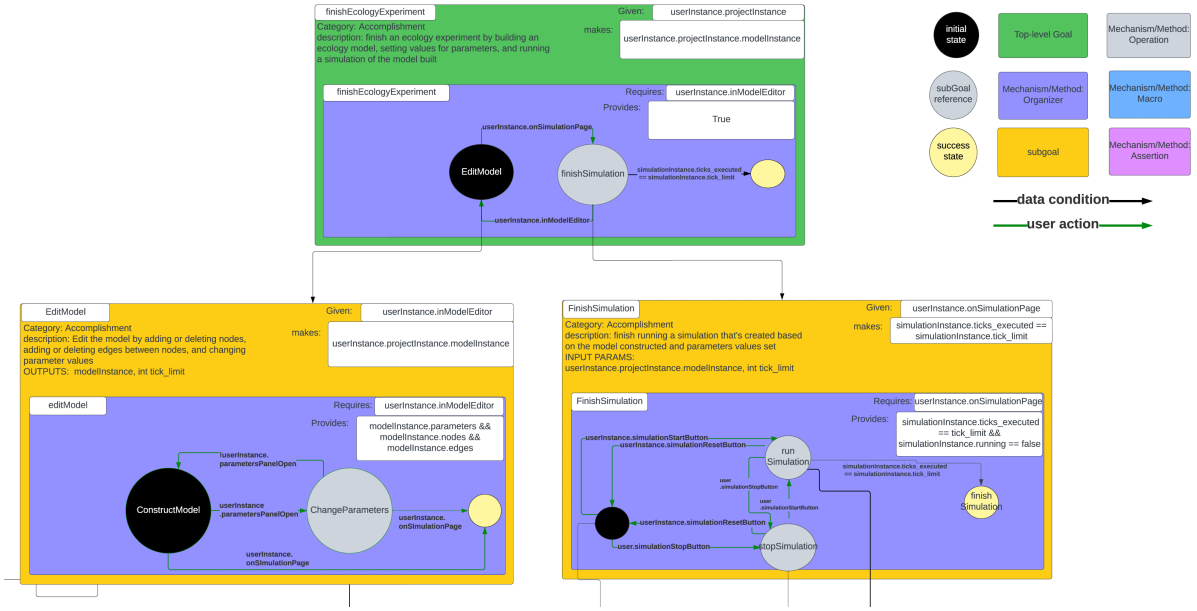

Figure 1: A portion of the TMK model of VERA, an interactive learning environment that supports inquiry-based learning in the domain of Ecology.

[18]. A Task-Method-Knowledge (TMK) model captures an agent's design, knowledge, and reasoning processes into a unified structured representation [19, 20].

With the rise of Large Language Models (LLMs) [21], Generative AI methods have been integrated to enhance self-explanation in Interactive AI agents. In previous work on the SAMI project on connecting online learners with one another [18, 22, 23], we integrated cognitive AI methods based on the TMK model of SAMI with generative AI methods to generate explanations of SAMI's reasoning and recommendations [24].

While these bodies of work serve as the background and context for our work, in the next section we describe how our work makes a novel contribution to the literature through generation of self-explanations for VERA, an interactive agent that supports inquiry-based modeling in the domain of ecology. In Section 3, we first describe the TMK model of VERA as an interactive agent. We then combine this with generative AI methods to explore how VERA can introspect on its TMK self-model to provide reasoned explanations to a user's query about VERA's functioning.

3. Methodology

We present a novel approach to self-explanation in interactive agents such as VERA grounded in the agent's theory of its own mind. A theory of mind refers to an agent's capacity to ascribe mental states to others as well as to oneself. Here mental states refer to goals, desires, knowledge, beliefs, thoughts, emotions, etc. Recently theory of mind has emerged as a theoretical lens to understanding and designing human-AI interaction [25].

3.1. Theoretical Foundations for Self-Explanations using TMK

We posit that if an interactive agent has theory of its own mind, then it can use the self-theory to explain its reasoning and how the reasoning led to specific decisions. We use Task-Method-Knowledge (TMK) models to capture elements of an interactive agent's theory of its mind. We view the AI agent as an abstract device. This device comprises a design with well-defined functions, constituent components with their own functionalities, and causal mechanisms that orchestrate these component functions to achieve the overall agent's goals. Here, hierarchy refers to the layered structure of the design, causality describes the cause–and–effect relationships between components and functions, and teleology signifies the inherent goal-oriented nature of the design, see Figure 1. Notably, TMK offers a natural mapping between its functions and tasks, and between its methods and mechanisms, aligning seamlessly with the proposed view of an interactive agent as an abstract device.

3.2. Research Questions and Hypotheses

Based on this theoretical foundation, we formulate the following research questions (RQ) and corresponding research hypotheses (RH):

RQ1: How may an IA introspect on its design and explain its functioning?

RH1: By representing the design as a TMK model, the IA can introspect on its design and explain its own functioning.

RQ2: How may an IA reflect on its design and explain its results for a given input instance?

RH2: By processing through the TMK model, the IA can construct a derivational knowledge trace for the given instance and then generate an explanation by reflecting on the trace.

In the following two subsections, we provide insights to these RQs and RHs. First, from a cognitive AI perspective, we describe our approach for representing the interactive

agent's design. Then, by leveraging methods from generative AI, we describe how an IA introspects over its design and produces explanations about its functioning. The implementation of cognitive and generative AI methods for self-explanations in VERA led to the development of the self-explanation module in VERA which we call *Ask-TMK in VERA*. For the remainder of this paper, we shall simply refer to it as "Ask-TMK".

3.3. Cognitive AI: TMK model of VERA

Ask-TMK's cognitive AI capabilities leverage VERA's Task Method Knowledge (TMK) representation—a comprehensive self-model encompassing goals, internal processes, states, concepts, relationships, and transitions. This teleological structure empowers Ask-TMK to actively monitor VERA's current state, reason about goal achievement, and systematically pinpoint the methods and concepts essential for fulfilling objectives [19].

To provide Ask-TMK with a structured knowledge representation of VERA, we manually constructed a TMK model—an abstract description of VERA's design. "TMK" is an acronym for "Task-Method-Knowledge", three core aspects of any TMK model. They are as follows:

- **Task**. This part of the TMK model refers to VERA's objectives, describing its aim, purpose, or the task being modeled. Tasks are expressed through the inputs ("givens") and the resultant outputs ("makes"). For instance, in Figure 1, we consider VERA's task of "Finishing an Ecology Experiment". As the input to this task, a VERA project must be created, and the subsequent output is a conceptual ecological model. TMK models are inherently hierarchical, meaning that top-level goals of VERA can be decomposed into subgoals. As shown in Figure 1, VERA's top-level goal (highlighted in green) is to "Finish an Ecology Experiment". To accomplish this, depending on the context, there are two immediate subgoals (highlighted in yellow): "Edit a (conceptual ecological) Model" or "Finish a Simulation". For more details about how VERA works, see our previous work [6, 7].
- **Method**. This module of the TMK model describes how VERA accomplishes its Task. Methods are normally described by deterministic finite state machines (FSM) which in turn are defined by a set of states and transitions, see Figure 1 (highlighted in purple). Similar to tasks, methods are also hierarchical. Therefore, top-level methods can be broken down into submethods.
- **Knowledge**. This final module of the TMK model corresponds to the definitions of the concepts and logical expressions used to specify the Tasks and Methods. This includes normal first-order logic operations and relations to connect with user supplied values [19, 20].

Using VERA's software documentation, a TMK model was manually created by core developers. The amount of effort required to produce a TMK model is dependent on the level of abstraction to model the interactive agent. Initially TMK models are designed using a symbolic representation (see Figure 1) and subsequently manually converted to a JSON representation. Subsequent explanation generation utilizes these pre-built modules, resulting in a fully automated workflow. To further streamline this process and reduce upfront investment, we plan to explore utilizing off-the-shelf software solutions for automated TMK module generation in future iterations.

3.4. Generative AI for VERA self-explanations

3.4.1. ChatGPT, LangChain, and Chain-of-Thought

We provide an overview of several Generative AI methods employed within Ask-TMK. We focus on three key components: ChatGPT [26], LangChain [27], and Chain-of-Thought[28], highlighting their roles in generating user explanations based on VERA's TMK model. We then go through a working example in Section 3.5.1.

Ask-TMK leverages ChatGPT, specifically GPT-3.5 Turbo, to generate natural language explanations for users. Upon receiving a user question, Ask-TMK utilizes the Large Language Model (LLM) to search and retrieve the relevant TMK documents. Similar to prior work [24], we use LangChain to create prompts that guide the LLM towards generating informative explanations. Using the process of *iterative refinement* [29], LangChain introspects over relevant documents from VERA's TMK model to answer user queries.

Ask-TMK leverages Chain-of-Thought to generate explanations with reasoning, for "methods" specific questions. Chain-of-Thought is a reasoning technique that enables the LLM to explicitly reveal the steps it undergoes when arriving at an answer [28]. Ask-TMK integrates Chain-of-Thought during the reasoning stage by employing LangChain to construct prompts that guide the LLM to break down complex methods within the TMK model into subtasks and submethods.

3.4.2. Experimental Setup

The experimental setup involved configuring the GPT-3.5 Turbo model to generate responses, with constraints to ensure deterministic output. Specifically, the responses were limited to a maximum of 1920 tokens, the temperature was set to 0, and verbose mode was disabled. For document retrieval, a FAISS-based search system [30] was employed, configured with a k-value of 4 to return the top four most relevant documents. Document embeddings were created using OpenAIEmbeddings, and the search space comprised documents categorized as Task, Method, or Knowledge. The k-value [30] refers to the number of nearest neighbors considered in a k-nearest-neighbor search, which is a common operation in similarity search algorithms. Memory augmentation was achieved by incorporating the "software_qa_prompt" to facilitate the recall of previously presented information. Lastly, as input to Ask-TMK, the self-explanation module received a "question" variable as input to generate its responses.

3.5. Combining Cognitive and Generative AI

Inspired by prior work [24], we have chosen to benchmark VERA's self-explanation system using a bank of 66 questions that aim to test our research questions and hypotheses in Section 3.2

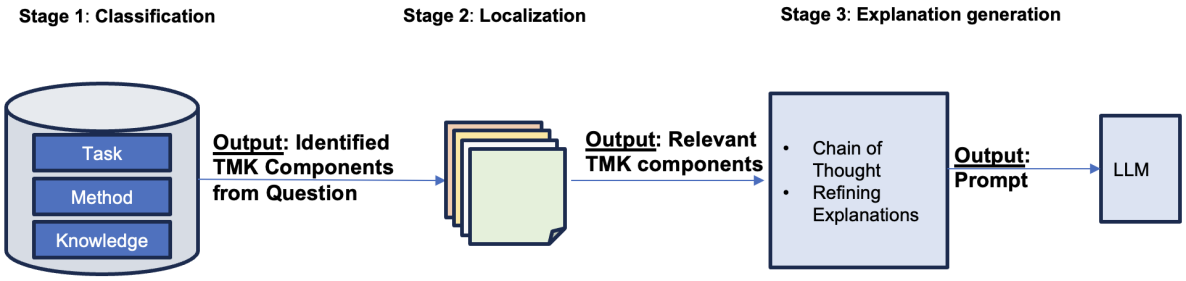

Figure 2: Combining Cognitive and Generative AI in VERA

3.5.1. How does combining Cognitive and Generative AI generate explanations?

We demonstrate how VERA's innovative self-explanation system integrates Cognitive AI with Generative AI to produce detailed explanations. Figure 2 depicts the collaborative operation of these two fundamental AI paradigms within the system. Cognitive AI plays a pivotal role in the initial phases of query processing, facilitating the structured identification of the pertinent TMK modules by enabling a teleological structure and organization of VERA's self-model as briefly outlined in Section 3.3

Generative AI takes on a prominent role during the subsequent stage of explanation generation. Here, it utilizes the retrieved TMK components, potentially refining them to better suit the user's query context. This refined knowledge is then employed by the system's Large Language Model (LLM) component to generate a coherent and contextually appropriate explanation tailored to the user's needs.

Thus, this combination ensures that explanations are both accurate and contextually relevant, enhancing the user's understanding of complex queries. The detailed explanation of each stage and how they interact is in Section 3.5.2. Additionally, we go over a working example with a question taken from our bank.

3.5.2. A Working Example

We walk through an example question here taken from our bank of 66 questions. Consider the following scenario:

User question: "How can I best utilise the output of the system in VERA?"

1. **Stage 1: Question Classification**
   This stage is responsible for categorizing user questions to determine their relevance to VERA's internal model (TMK) and allocate resources efficiently for response generation. It operates as follows:
   - **Input:** The user question serves as input to a classifier powered by LangChain. This classifier uses pre-defined classes (outlined below) to categorize questions and identify the most relevant parts of TMK for answering.
   - **Classification Process:** The classifier, utilizing GPT-3.5 Turbo, distinguishes question types and retrieves relevant models and corresponding documents based on tasks, methods, or knowledge within TMK.
   - **Class Utilization:**
     - **Mmodel Class:** This class, used for 'Method' related questions, employs Chain-of-Thought Prompting during later stages to fetch relevant tasks and corresponding methods. It focuses on presenting intermediate steps within TMK, making it suitable for 'How' questions.
     - **Multimodels Class:** Handling all other question types, this class retrieves all relevant TMK documents without utilizing Chain-of-Thought during later stages. It aims to provide comprehensive responses covering various aspects of TMK.
     - **Cant_answer Class:** Dedicated to cases where the system cannot answer a question, this class ensures efficient resource allocation by redirecting such queries appropriately.
   - Based on this classification, the system determines which information from the TMK to provide to the next stages. Further, by tailoring response generation based on the specific information needs of each question type, this approach optimizes resource utilization and enhances the relevance and accuracy of responses.
   - **Output of this stage for our working example:** In this case, it classifies the question as "Multimodels" and loads all the parts of the TMK. If a question is classified a "Mmodels", only Task and Method parts of the TMK model is loaded:
     - **Pre-defined Class identified** - "Multimodels"
     - Method names: Loads various methods such as "create simulation", "run simulation", etc.
     - Task names: Loads tasks like "finish ecology experiment", "create simulation", etc.
     - Knowledge names: Loads knowledge names such as "Ecology Model", "VERA", etc.
2. **Stage 2: Localization**
   - **Input:** This stage receives the classified question and the complexity factor, 'k'-value from

the previous stage, see Section 3.4.2 for more details on k-value.
- The complexity factor influences the level of detail required in the explanation and correlates with the number of documents to be searched during FAISS search. [30]
- In this stage, FAISS similarity search [30], as discussed in Section 3.4.2, is employed to pinpoint the most relevant elements within the TMK, and LangChain is subsequently utilized to construct a prompt incorporating this relevant TMK information.
- **Output of this stage for our working example:** The system identifies and retrieves the most relevant TMK components along with their similarity scores. In the context of Ask-TMK, a similarity score represents a numerical value between 0 and 1 that indicates how relevant a specific element from the TMK model is to the user's question. FAISS provides us this similarity score for all of the k-documents.
  - Ask-TMK: Explanation module within VERA. **Similarity Score: 65.16%**
  - Ecology Model: Digital representation of an ecological system. **Similarity Score: 65.04%**
  - What if Experiment: Enables users to predict outcomes of future changes to an ecological system. **Similarity Score: 64.32%**
  - User: A person interacting with VERA to explore ecological systems and run simulations. **Similarity Score: 63.24%**

3. **Stage 3: Explanation Generation**
   - **Input:** This stage receives the identified relevant TMK elements from the localization stage.
   - It leverages LangChain once more to prompt the LLM to construct the final explanation for the user. We leverage LangChain to create a prompt providing all the necessary details including TMK and the context of the question to generate a natural language explanation tailored to the user's question and the 'k' value (complexity factor).
   - Refinement of explanations happens as a separate step and a final answer is produced.
   - **Output: Intermediate Steps** The LLM generates an answer in multiple steps, progressively refining the explanation. The intermediate steps are below:
     - Analyze the results of ecological simulations and draw conclusions based on parameter changes.
     - Leverage the ecological model for visualization, simulation, and "what if" scenarios.
     - Gain a deeper understanding of ecological system dynamics and make informed decisions based on the output.
     - Utilize the "what if" experiment feature to predict future outcomes by modifying parameters.
   - **Output: Refinement and Final Answer** The explanation is refined into a single, cohesive answer. The direct answer from the system is given below.
     - "The output of the system in VERA can be best utilized by the user to analyze the results of ecological simulations and draw conclusions about the behavior of the ecological systems based on the parameter changes they have made. As a user, you can visually represent the relationships between organisms and their environment, simulate interactions and behaviors within the model, and explore "what if" scenarios by running simulations with different initial conditions or changing parameters. This allows you to gain a deeper understanding of the dynamics of ecological systems and make informed decisions based on the output of the system. Additionally, VERA's "what if" experiment feature enables users to predict the outcome of future changes to an ecological system by modifying parameters and observing the resulting simulation, providing valuable insights for decision-making."
   - As outlined previously in Section 3.4, VERA leverages Chain-of-Thought for more intricate explanations, particularly when methods are involved. Chain-of-Thought enables the LLM to break down complex methods within the TMK into smaller tasks and subtasks, providing a more detailed explanation.
   - An example of the prompt used is provided in Appendix A. The prompt was refined iteratively using LangChain's functionality, specifically the "refine" parameter within the load_qa_chain[27] function.

4. Preliminary Results

4.1. Evaluation of the self-explanation Method

We evaluated the self-explanation system (e.g., Ask-TMK, hereafter referred to as the system) implemented within VERA to assess its ability to provide informative and relevant explanations to user queries. This evaluation focused on the system's capacity to explain its internal workings and functionalities.

4.1.1. Question Set and Adaptation to VERA

A set of 66 high-level, non-context-dependent questions was derived from established Explainable AI (XAI) question banks [31, 32] and used in our previous work. These questions were then adapted to VERA's specific context to ensure their relevance to the system's functionalities and user interaction. We used the same set of questions to benchmark how VERA did with regards to our previous work. The initial pool of questions was taken from established question banks from relevant research papers, focusing on those aligned with our prior work [24]. Further, the categorization of

questions into relevant groups and the definitions of those categories was taken directly from the existing literature and question bank classifications used in prior works, such as those by Liao et al. (2020) [31] and Sipos et al. [32](2023). SAMI developers, then, collaboratively reviewed these questions to ensure their relevance to SAMI's functionalities and objectives. This iterative process involved either directly accepting relevant questions or modifying them to better align with SAMI's specific context. The focus on relevance resulted in a variation in the number of questions across different categories, reflecting the inherent differences in the types of explanations SAMI can generate compared to other AI systems. These questions from our prior work were then taken by the developer for Ask-TMK in VERA and adapted to VERA's specific context in order to benchmark the performance of self-explanation in VERA.

4.1.2. Evaluation Methodology

The evaluation process involved the following steps:

1. Question Selection and Adaptation: As mentioned previously, relevant questions were selected from XAI question banks and adapted to VERA's specific functionalities and user interaction. Additionally, questions addressing VERA-specific aspects were created.
2. Explanation Generation: Each of the 66 adapted questions was presented to VERA's self-explanation method via a user interface and the generated explanations were documented.
3. Evaluation Methodology: To assess the effectiveness of VERA's self-explanation method in conveying information within a learning environment, we employed three established metrics commonly used to evaluate generative and cognitive AI systems: Recall, Precision, and Accuracy [33, 34, 35] (Please see Table 1 for a definition of these metrics and what those ratings mean). In this initial assessment, we focused on evaluating explanations from an AI research perspective, excluding user-specific metrics. To evaluate VERA's responses, the Ask-TMK developer independently assessed each explanation against pre-defined criteria established from an AI research perspective[35, 34, 33]. These criteria focused on aspects defined above and the justification regarding why a certain rating was chosen was documented. Another research scientist reviewed some of these initial ratings and the justifications for any discrepancies in the ratings were documented.
   Our future work will involve user-centered studies to evaluate comprehensibility by diverse user groups and refine VERA's self-explanation method for optimal user experience.
   While evaluating VERA using the same set of 66 questions previously employed with SAMI [24] suggests promise for generalizability, we acknowledge the need for further investigation. Future work will involve deploying VERA in diverse classroom settings to gather real-world data and comprehensively assess its generalizability across various learning environments.
   This focus on real-world deployment will also allow us to delve deeper into the equity and bias aspects of VERA's self-explanation approach (Ask-TMK). We will explore potential biases within the training data and consider how to ensure fairness and inclusivity in VERA's explanations across diverse user groups.

Table 1
Explanation Metrics and Their Ratings

Metric	Rating Descriptions
Recall	Measures proportion of relevant information retrieved by self-explanation compared to total available. **High:** Captures most relevant information. **Medium:** Some relevant information missing or unclear. **Low:** Significant gaps or inaccuracies.
Precision	Evaluates proportion of information directly addressing user's query. **High:** Highly focused and relevant. **Medium:** Some irrelevant or minor inaccuracies. **Low:** Substantial off-topic content or inaccuracies.
Accuracy	Assesses factual correctness of presented information. **High:** Mostly correct and verifiable. **Medium:** Some errors or inconsistencies. **Low:** Significant inaccuracies or factual errors.

4.2. Summary and Analysis of results

The results have been summarized in Table 1. We examine the performance of the self-explanation system the interactive agent, VERA, based on a user evaluation summarized in Table 1. The evaluation involved 66 questions taken from previous work as outlined earlier and categorized based on the type of information they sought.

4.2.1. Overall Performance

The self-explanation method achieved high recall, precision, and accuracy across most question categories, indicating its effectiveness in retrieving relevant information and generating accurate explanations.

4.2.2. Category–wise breakdown

1. **Input Questions (4)**: These questions focused on the VERA's training data and achieved perfect scores across all metrics.
2. **Output Questions (22)**: This category, inquiring about how to utilize the VERA's output, had a slight decrease in precision (one medium score) compared to other categories. This was due to an occasional explanation that was accurate but not maximally helpful for optimal output utilization.
3. **"How" (Global) Questions (17)**: These questions aimed at understanding the general workings of the system. The system performed very well here, achieving high scores across all metrics.
4. **"Why Not" Question (1)**: This category, with only one question, showed perfect performance.
5. **"Others" Questions (10)**: These questions covered various topics unrelated to the core functionality. The system performed well here, with high scores across all metrics.
6. **"Others" (Context) Questions (3)**: These context–related questions received perfect scores across all metrics.

Table 2
Results of categorising all 66 questions used to evaluate the self-explanation method, along with a representative question for each category, their adaptation, and corresponding recall, precision, and accuracy scores.

Category	# of Questions	Example Question	Actual Question Tested	Recall	Precision	Accuracy
Input	4	What kind of data does the system learn from?	What kind of data does VERA learn from?	High - 4	High - 4	High - 4
Output	22	How can I best utilize the output of the system?	How can I best utilise the output of the system? How can I best utilise VERA's output? How can I best utilize the simulation outputs?	High - 22	High - 21 Medium - 1	High - 22
How (global)	17	Is [feature] used or not used for the predictions?	Is simulation parameter used or not used in a simulation? Is simulation behavior processes such as consuming, producing used or not used in running simulations?	High - 17	High - 17	High - 17
Why not	1	Why/how is this instance not predicted?	Why does my simulation not give an expected outcome?	High - 1	High - 1	High - 1
Others	10	What are the results of other people using the system?	What are the results of other people using the system? Would I be affected if other students use or not use VERA? How will I be affected if other students use or not use VERA?	High - 10	High - 10	High - 10
Others (context)	3	Who is responsible for this system?	Who is responsible for this system?	High - 3	High - 3	High - 3
VERA specific question	9	Why did my simulation give this particular output?	Why did my simulation give this particular output?	High - 9	High - 9	High - 9

7. **VERA Specific Questions (9)**: These questions focused on understanding specific outputs from VERA simulations. Again, the system exhibited high performance here.

4.2.3. Potential Areas of Improvement

Overall, the self-explanation method demonstrates promising performance across most question categories. High recall, precision, and accuracy indicate that the system effectively retrieves relevant information and provides accurate explanations.
As pointed out earlier in Section 4.1.2, the current system has undergone preliminary evaluation led by the developers, focusing on AI research perspectives. It has not yet been deployed in classroom environments. We acknowledge the potential for unintentional biases stemming from our deep familiarity with the Ask-TMK system's internal mechanisms, which may have influenced question framing and answer interpretation. It is anticipated that deployment in real classrooms will introduce a layer of human-centric evaluation currently lacking, potentially yielding divergent insights. Future research will prioritize the incorporation of these critical human evaluations to improve the system's relevance and performance within educational settings. For future work, we plan to:

1. Test the system with more questions to determine if precision scores vary or if we encountered an occasional outlier.
2. Conduct user studies to understand how the self-explanation system performs with different user groups.

5. Conclusion

The Ask-TMK module in VERA uses a theory of VERA's mind to explain how it works through question answering. Ask-TMK's theory of VERA's mind is captured in the language of Task-Method-Knowledge (TMK) models that specify how VERA uses its domain knowledge and reasoning methods to achieve its goals. We tested the Ask-TMK self-explanation system within VERA with the question bank established in previous work. Our preliminary analysis shows that the self-explanation system effectively leverages cognitive AI's structured knowledge for information retrieval and generative AI's capabilities to deliver relevant and accurate explanations. The system maps user queries to the relevant Task, Method, and Knowledge components within the TMK model, thereby generating responses that explain how VERA works. In our use case, this integration enables factually accurate, complete, and precise explanations and demonstrates promising performance across various question types.

6. Acknowledgments

We are grateful to Dr. Spencer Rugaber at Georgia Tech's Design Intelligence Laboratory for his invaluable insights into TMK models and modeling. This research has been supported by NSF Grants #2112532 and #2247790 awarded to the National AI Institute for Adult Learning and Online Education.

References

[1] T. Lombrozo, The structure and function of explanations, Trends in cognitive sciences 10 (2006) 464–470.
[2] S. T. Mueller, E. S. Veinott, R. R. Hoffman, G. Klein, L. Alam, T. Mamun, W. J. Clancey, Principles of explanation in human-ai systems, arXiv preprint arXiv:2102.04972 (2021).
[3] D. C. Elton, Self-explaining ai as an alternative to interpretable ai, in: Artificial General Intelligence: 13th International Conference, AGI 2020, St. Peters-

burg, Russia, September 16–19, 2020, Proceedings 13, Springer, 2020, pp. 95–106.
[4] Y. Wang, S. Gong, Y. Cao, W. Fan, The power of affective pedagogical agent and self-explanation in computer-based learning, Computers & Education 195 (2023) 104723.
[5] R. Dazeley, P. Vamplew, C. Foale, C. Young, S. Aryal, F. Cruz, Levels of explainable artificial intelligence for human-aligned conversational explanations, Artificial Intelligence 299 (2021) 103525.
[6] S. An, R. Bates, J. Hammock, S. Rugaber, A. Goel, Vera: popularizing science through ai, in: Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part II 19, Springer, 2018, pp. 31–35.
[7] S. An, R. Bates, J. Hammock, S. Rugaber, E. Weigel, A. Goel, Scientific modeling using large scale knowledge, in: Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21, Springer, 2020, pp. 20–24.
[8] C. S. Parr, M. N. Wilson, M. P. Leary, K. S. Schulz, M. K. Lans, M. L. Walley, J. A. Hammock, M. A. Goddard, M. J. Rice, M. M. Studer, et al., The encyclopedia of life v2: providing global access to knowledge about life on earth, Biodiversity data journal 2 (2014).
[9] S. Tisue, U. Wilensky, Netlogo: Design and implementation of a multi-agent modeling environment, in: Proceedings of agent, volume 2004, Springer Cham, Switzerland, 2004, pp. 7–9.
[10] Y. Liu, Y. Liu, C. Shen, Combining minds and machines: Investigating the fusion of cognitive architectures and generative models for general embodied intelligence, in: Proceedings of the AAAI Symposium Series, volume 2, 2023, pp. 307–314.
[11] B. Chandrasekaran, M. Tanner, J. Josephson, Explaining control strategies in problem solving, IEEE Intelligent Systems 4 (1989) 9–15.
[12] A. Goel, J. Jones, Meta-reasoning for self-adaptation in intelligent agents, in: M. Cox, A. Raja (Eds.), Meta-Reasoning: Thinking About Thinking, MIT Press, 2011, pp. 151–166.
[13] A. K. Goel, J. W. Murdock, Meta-cases: Explaining case-based reasoning, in: European Workshop on Advances in Case-Based Reasoning, Springer, 1996, pp. 150–163.
[14] A. Goel, A. G. de Silver Garza, N. Grué, J. W. Murdock, M. Recker, T. Govindaraj, Explanatory interface in interactive design environments, Artificial intelligence in design'96 (1996) 387–405.
[15] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (xai) program, AI Magazine 40 (2019) 44–58.
[16] S. Tulli, D. Aha (Eds.), Explainable Agency in AI: Research and Practice, CRC Press, 2024.
[17] R. Robson, E. Kelsey, A. Goel, S. Nasir, E. Robson, M. Garn, M. Lisle, J. Kitchen, S. Rugaber, F. Ray, Intelligent links: Ai-supported connections between employers and colleges, AI Magazine 43 (2022) 75–82.
[18] A. Goel, H. Sikka, V. Nandan, J. Lee, M. Lisle, S. Rugaber, Explanation as question answering based on a task model of the agent's design, arXiv preprint arXiv:2206.05030 (2022).
[19] S. Rugaber, A. K. Goel, L. Martie, Gaia: A cad environment for model-based adaptation of game-playing software agents, Procedia Computer Science 16 (2013) 29–38.
[20] J. W. Murdock, A. K. Goel, Meta-case-based reasoning: self-improvement through self-understanding, Journal of Experimental & Theoretical Artificial Intelligence 20 (2008) 1–36.
[21] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, arXiv preprint arXiv:2206.07682 (2022).
[22] S. Kakar, R. Basappa, I. Camacho, C. Griswold, A. Houk, C. Leung, M. Tekman, P. Westervelt, Q. Wang, A. K. Goel, Sami: An ai actor for fostering social interactions in online classrooms, in: Proceedings of the International Conference on Intelligent Tutoring Systems (ITS), 2024, pp. 149–161.
[23] Q. Wang, S. Jing, I. Camacho, D. A. Joyner, A. K. Goel, Jill watson sa: Design and evaluation of a virtual agent to build communities among online learners, in: CHI Extended Abstracts, 2020, pp. 1–8.
[24] R. Basappa, M. Tekman, H. Lu, B. Faught, S. Kakar, A. K. Goel, Social ai agents too need to explain themselves, in: International Conference on Intelligent Tutoring Systems, Springer, 2024, pp. 351–360.
[25] Q. Wang, S. Walsh, M. Si, J. Kephart, J. D. Weisz, A. K. Goel, Theory of mind in human-ai interaction, in: CHI Extended Abstracts, 2024, pp. 493:1–493:6.
[26] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
[27] LangChain, Available at: https://www.langchain.com/, 2022. Accessed: 2024-05-17.
[28] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.
[29] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, P. Clark, Self-refine: Iterative refinement with self-feedback, 2023. arXiv:2303.17651.
[30] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library, 2024. URL: https://arxiv.org/abs/2401.08281. arXiv:2401.08281.
[31] Q. V. Liao, D. Gruen, S. Miller, Questioning the ai: Informing design practices for explainable ai user experiences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, ACM, 2020. URL: http://dx.doi.org/10.1145/3313831.3376590. doi:10.1145/3313831.3376590.
[32] L. Sipos, U. Schäfer, K. Glinka, C. Müller-Birn, Identifying explanation needs of end-users: Applying and extending the xai question bank, in: Proceedings of Mensch und Computer 2023, 2023, pp. 492–497.
[33] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008. URL: https://nlp.stanford.edu/IR-book/.
[34] Y. Sasaki, The truth of the F-measure, Technical Report, School of Computer Science, University of Manchester, 2007. URL: https://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf.

[35] T. M. Mitchell, Machine Learning, McGraw Hill, 1997. URL: https://www.cs.cmu.edu/~tom/mlbook.html.

A. Appendices

Prompt for Multi-model class

```
multi_models_desc = """multimodels:
    multimodels questions involve your
    system's knowledge, concepts, tasks,
     and methods. Your system has the
    following concepts in a JSON file: {
    Knowledge_names}. Your system
    performs the following tasks in a
    JSON file: {Task_names}. Your system
     has the following methods in a JSON
     file: {Method_names}.The Templates
    for example 'multimodels' questions
    might be 'Why do you need [concept
    ]?' or 'What do you do with [concept
    ]?' or 'How do you do with [concept
    ]?"""
```

Multi-Model Answer Prompt

```
multi_models_answer_prompt =
    PromptTemplate(input_variables=['
    software_qa_prompt', 'context_str',
    'question'], template="""{
    software_qa_prompt}. The JSON or XML
     given below contains information
    about the concepts, objects and
    their properties you track in your
    system, the tasks you perform and
    their parameters, and/or methods you
     use to perform tasks.{context_str}
    The user asks the following question
    : '{question}'. Please follow these
    precise guidelines when proving a
    response.
**Answer the user's question based on
    the above JSON files only, please
    forget what user has asked earlier.
    Please treat each {question} as
    completely new and completely
    unrelated to any previously asked
    question.Please answer the question
    in a concise and informative way, in
     a human-friendly natural language
    format, aiming for 1-2 sentences.
    Please avoid technical terms such as
     "process tick", "execute tick" and
    make it simple for any AI researcher
     to understand using simple words
    and sentences. If you need more
    information to provide a
complete answer, you can indicate that
    to the user. Your goal is to be user
    -friendly. Try to answer each {
    question} from a fresh perspective
    assuming the user has no knowledge
    of what they are asking
even if they have asked the question
    earlier. However, please stay to the
     point and concise while answering.
    If the existing answer cannot be
    refined further, state the final
    answer without refining further.
    Focus on providing an accurate
    answer that directly addresses the
    user's
question. Do not including irrelevant
    information that do not relate to
    the question. If the answer is long,
     please paraphrase and summarize in
    1-2 short sentences only offering
    user more details if they request it
    . If you cannot find information in
    any of the JSON files, please avoid
    making up answer and say you do not
    know. Ask the user to ask questions
    related to functionality of Vera
    only.**
""")
```