**Self-Explanation in AI Experimental Agent - Vera**

Shalini Sushri (ssushri3@gatech.edu), Rahul Dass (rdass7@gatech.edu). Research work under the mentorship of Prof. Ashok Goel (ashok.goel@cc.gatech.edu)

---

**Introduction**

**RQ: How do we make AI agents transparent?**

**Issue 1.1: How do we enable AI agents to explain itself?**

H. By giving the AI agent a model of itself that it can use for its metacognition. The agent can use this metacognition to explain itself.

**Methods**

- We decided to test this hypothesis on Vera that is experimental AI agent. Our methodology included creating the metacognition of AI agent using:
  - TMK Representation
  - Langchain to work with LLMs
  - LLMs to generate responses

**Key Findings**

The metacognitive agent was able to correctly classify most questions into correct categories using the FAISS similarity search techniques.

**Results**

Our output from the AI agent consisted of the following components:

- Top "k" relevant documents with confidence scores
- Intermediate steps that led to the final answer
- Final Answer
- Final Answer

**Analyses**

- Tested the metacognitive agent with 10 questions of each category
- Task", "Method", "Knowledge"
- Method questions showed high variance in confidence scores possibly due to Chain of Thought Prompting
- A few relevant and valid TMK questions resulted in "Cannot Answer Response".
- Some of the responses were incorrect. For example,
  - There was a different response to the following equivalent questions: Why does a cow eat grass? And Why does a cow consume grass?

**Conclusion and Future Work**

We started with the hypothesis that if an AI agent had a metacognition model of itself, it will be able to explain itself to a human user.

1. We were able to develop a proof of concept for such an agent.
2. While we tested our questions, a few categorizations were incorrect.
3. However, we were able to peek inside the agent's "mind" with:
   - The Top relevant documents searched and the associated confidence scores
   - The chain of thought prompting that gave us the sequence of nodes in a decision tree
   - The intermediate steps that it went through to arrive at the final answer

- There is tremendous work to be done before we can say an AI agent can fully explain itself, notably in the following areas:
  - What framework should we use for evaluation of self-explanation (accuracy, completeness, relevance or something else)?
  - The output is still not completely transparent and depends on user domain knowledge.
  - Do we need to add more models to improve the accuracy of search results?
  - How do we send dynamic episodic data to enable the AI agent to provide real-time responses of why it made a particular decision produced a particular response?

---

**References**

- S. Rugaber. "TMKL2 – A Teleological Language for Adapting Software." -
  - https://doi.org/10.1016/j.caai.2022.100074

**Acknowledgements**

The work draws on the knowledge representation paper by "(TMK model representation)" done by Spencer Rugaber and Prof. Ashok Goel.

The Self-Explanation component of AI agents at DILAB is similar across other AI agents at DILAB such as SAMI (Rhea B., Mustafa Takeman, Chris Leung, Ben Fraught) and SkillSync (Vrinda Rai). This work was started by former and current GT students, Helen Lu, Dilek Marzak, Shawn Hodgson.

We have extended the self-explanation work in VERA which is an experimental AI agent. A huge thank you to Rahul Dass for collaboration and support and Rhea B., John K., and Shawn Hodgson for getting Vera up and running without which self-explanation module would not have worked.