Self- Explanation in AI Experimental Agent - Vera

School of Interactive Computing

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

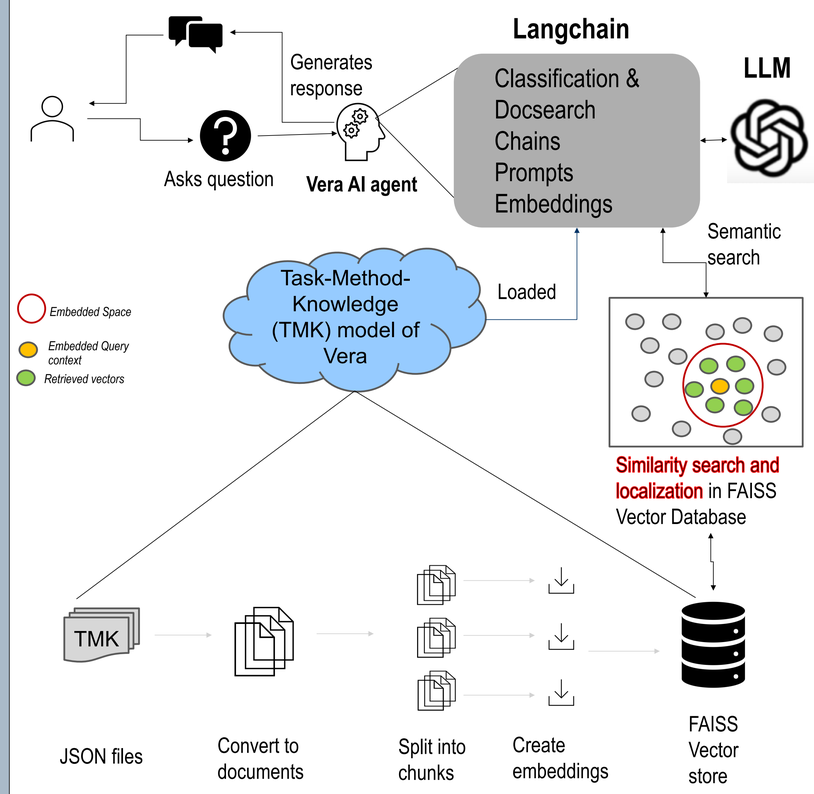
RQ: How do we make AI agents transparent?

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

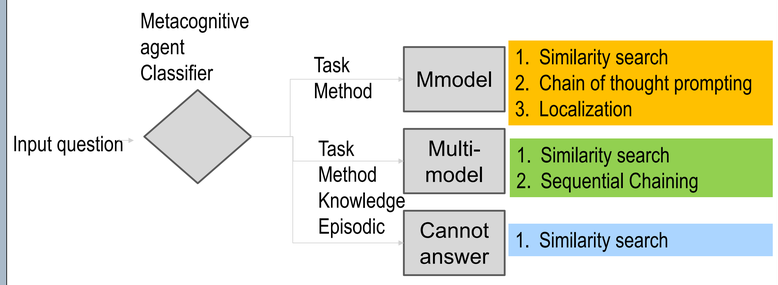
- We decided to test this hypotheses* on Vera that is experimental AI agent. Our methodology included the creating the metacognition of AI agent using:
  - TMK Representation
  - Langchain to work with LLMs
  - LLMs to generate responses



*The architecture of Metacognitive agents was created by former students of DILAB. Please see acknowledgements. This work involved extending and developing a proof of concept in Vera

Key Findings

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



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

Category of question and top similar relevant documents	Localized part of Self-Model that contained responses	Chain of Thought reasoning to generate responses

Results

1. The metacognition within VERA produced the following output:

Multimodel Output	Mmodel Output
Top ”k” relevant documents with confidence scores	Top ”k” relevant documents with confidence scores
Intermediate steps that led to the final answer	Localized Task and corresponding activated methods
Final Answer	Chain of thought leading to final output
Cantanswer Output	Intermediate steps that led to the final answer
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?

Type of Question	Model	Confidence Score
Task	Multi-models / Can’t Answer	61-69
Method	Mmodel / Can’t Answer	49 - 76
Knowledge	Multi-models / Can’t Answer	60 - 69

Conclusion and Future Work

We started with the hypotheses 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, A. K. Goel, and L. Martie, “GAIA: A CAD Environment for Model-Based Adaptation of Game-Playing Software Agents,” Procedia Computer Science, vol. 16, pp. 29–38, 2013.
S. Rugaber, “TMKL2 – A Teleological Language for Adapting Soft-ware,”

•Khosravi, Hassan, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. “Explainable Artificial Intelligence in Education.” *Computers and Education: Artificial Intelligence* 3 (2022): 100074. https://doi.org/10.1016/j.caeai.2022.100074.

•Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. San Francisco California USA: ACM, 2016. https://doi.org/10.1145/2939672.2939778.

•Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. NeurIPS, 2022b. URL https://arxiv.org/abs/2201.11903.

•Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. TMLR, 2022a. URL https://openreview.net/forum?id=yzkSU5zdwD.

•https://faiss.ai/index.html

•https://api.python.langchain.com/en/latest/vectorstores/langchain.vectorstores.faiss.FAISS.html

Acknowledgements

The work draws on the knowledge representation paper by (TMK model representation[1]) 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 Manzak, 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.

[1] S. Rugaber, A. K. Goel, and L. Martie, “GAIA: A CAD Environment for Model-Based Adaptation of Game-Playing Software Agents,” Procedia Computer Science, vol. 16, pp. 29–38, 2013.