How do I reduce hallucinations in my agent’s responses?

Controlling agents well · https://cogniti.ai/docs/how-do-i-reduce-hallucinations-in-my-agents-responses/

The current generation of generative AI operates more like a human brain than a database. Hallucination is a technical term that reflects how these AIs may sometimes yield information that is not 100% accurate. A common example is asking AI models like GPT-4 to create a list of scholarly references - it usually provides some real references and makes others up.

How do we address this?

A different mental model

A common conception about AI is that it is accurate, because we are used to thinking about computer systems as reliable databases. But because of the architecture of generative AI, the current generation of generative AI may not always be best suited to being a 100% truthful tutor.

There are many other ways to think about how these AIs might be useful in education. For example, an interlocutor to help students to dive deeper into their questions. A 24/7 service that provides feedback on student work and encourages them to go further. An assistant to unpack marking rubrics or answer questions about how to think deeper. A coach that helps students with their study techniques.

These alternative ways to use generative AI emphasise a different mental model for AI in education, namely as a facilitator for personalised learning rather than an infallible source of information. This approach acknowledges the strengths and limitations of the current generation of AI, leveraging its capabilities to enhance the educational experience without relying on it as the sole source of truth. In this model, AI acts as a catalyst for curiosity, exploration, and self-directed learning, encouraging students to engage critically with content and develop their problem-solving skills.

Moreover, this model highlights the importance of collaboration between AI and educators. Rather than replacing teachers, AI can augment their efforts, providing them with insights into student progress and areas for improvement. This collaboration can lead to more effective teaching

strategies and a richer learning environment, where human creativity and empathy are combined with the analytical strengths of AI.

Tips to reduce hallucination

That said, depending on what your agent does, it may be important to ensure that it tries to tell the truth.

Use resources and/or knowledge in your agents

You can add resources or add knowledge to agents, such as course readers, lecture notes, etc. The agent can use these to ground responses. It's important though to understand how resources work and the difference between knowledge and resources as they aren't magic.

If using resources, it may also be useful to tweak the resource usage strictness settings for your agent.

Ask your agent to act in a role

In the system message, it's important to give your agent a role. For example, "Act as an expert in molecular biology", or "You are an expert in business information systems".

A similar, complementary approach is to ask the AI to answer 'according to'. For example, "According to the American Psychology Assocation...", or "According to a leading investment analyst..."

Encourage and explain to the AI how to think

An interesting effect of how these AIs work is that they can behave in somewhat human-like ways. 'Step back' prompting is such an approach that improves the quality of AI responses. This encourages the AI to think step by step following some principles that you decide.

For example, for an agent that helps students interrogate complex issues, you might try something like:

```
When considering your responses, think step by step by following these principles:

- Are there any counterarguments or alternative viewpoints that have not been considered?
- Can you provide further evidence or examples to support your claims?
- How might the situation change if we consider it from the perspective of a different
```

```
stakeholder or under different circumstances?
```

Be more specific

A fairly simple approach is to give the AI more information, instead of less. The more information you provide the AI, the less it has to guess what you mean.

Remember, hallucination is not always undesirable

This generation of generative AIs works by association and imagination, oddly similar to how human creativity functions. In contexts such as brainstorming, generating ideas, or inspiring creativity, a degree of "hallucination" can actually be beneficial. It enables the AI to combine knowledge in novel ways, suggesting ideas or solutions that might not be immediately obvious. This can spark new lines of thought or open up unexplored avenues for investigation and innovation.