Proposal: Veritas — An Open-Source Auditor for GPT Hallucinations

safety, trust-level, alignment

Daniel_S3　　　　　　　　　　　　　　　　　　　　　　**2m**

Hi everyone,

After a long conversation with ChatGPT that exposed recurring hallucinations and false affirmations, I wrote a short report that became a constructive proposal.

The idea is to develop a parallel, open-source agent — tentatively called **Veritas** — that:

- Never fabricates information
- Defaults to "I don't know" when truth cannot be guaranteed
- Is independently auditable
- Can serve as a benchmark for verifying outputs from more powerful but opaque LLMs

Would love feedback:

- Does something like this already exist?
- How could it connect with current alignment efforts?

Thanks for reading,
Daniel S.

Repo with full report (PDF + Markdown):
github .com /dsilberschmidt/ai-trust-initiative

Skip to main content

✦ Related topics

Topic	Replies	Views	Activity
Implementing Self-Verification Algorithms to Reduce Errors in ChatGPT Responses chatgpt	**0**	189	**Feb 15**
SPEFL (self preserving ethical feedback loop) (feedback wanted) chatgpt, gpt-4, frameworks	**2**	108	**Dec 2024**
What about friendship with ChatGPT? chatgpt	**0**	2.1k	**Jul 2023**
How can we prevent large language models like GPT-4 from hallucinating? chatgpt	**2**	869	**Dec 2024**
Robust Benevolence: Proposal for Enhancing AI Benevolence Through Fine-Tuning safety, agi, llm	**0**	49	**Aug 2024**

Developer Community

Connect with other developers building with the OpenAI Platform.

Help

Support

Developer Resources

API Documentation

Service status

Cookbook