GROUNDING IS NOT ALL YOU NEED

Stop hallucinations and incorrect answers in generative search

Colin Harman
Head of Technology, Nesh

AGENDA

01

AWARENESS

02

EXAMPLES

03

CAUSES

04

PREVENTION

05

TRENDS

06

ABOUT

01 AWARENESS

2 ANSWER CATEGORIES

Unfaithful

*“Closed domain hallucinations refer to instances in which the model is instructed to **use only information provided** in a given context, but then **makes up extra information** that was not in that context.”*

[OpenAI GPT-4 System Card](#)

Incorrect

“…a confident answer by an AI that is not correct considering its context.”

Modified from [Hallucinations (AI) Wikipedia](#)

MOTIVATION

- User satisfaction
- Reputational damage
- Operational cost, latency, security

02 EXAMPLES

EXAMPLES DEMO

Check on LinkedIn for slides with examples shown after the conference

03 CAUSES

Hallucination Cause Categories Demo

- Memory leak
- Numbers
- Similar concepts
- Complex reasoning
- Other

Check on LinkedIn for slides with category deep-dives after the conference

04 PREVENTION

Prevention Techniques Demo

- LLM usage
- Supervisory systems
- Tool usage
- Pre-generation mitigation
- Setting user expectations

Check on LinkedIn for slides with technique details shown after the conference

05 TRENDS

STRONGER MODELS

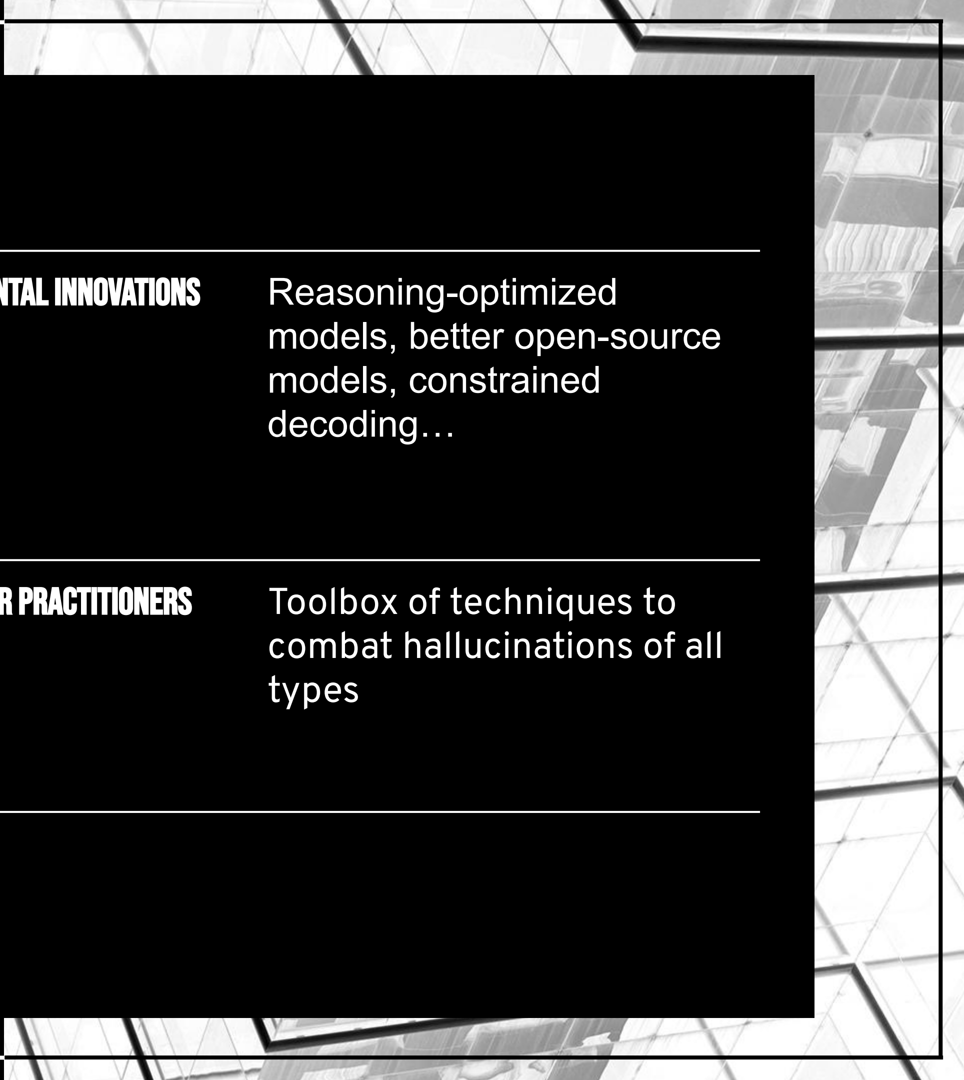
- GPT-4 reduces but does not eliminate unfaithful answers vs 3.5, in large part due to RLHF-like training
  - 29% better at preventing cd hallucination
- However, RLHF-like training may damage numbers/calibration [1], [2]
- Latency, cost, reliability are problems but will slowly improve as providers multiply
- Training may remain unavailable or cost prohibitive
  - At time of writing, GPT-3.5-turbo and GPT-4 cannot be fine-tuned by users
- Models will likely not drastically improve in problem areas (e.g. numbers), tools will remain important
- Conclusion:
  - Hallucinations will gradually decrease, but some areas will remain problematic (e.g. numbers)
  - Strongest models will remain unrealistic for many use cases due to cost, latency, security

WEAKER MODELS

- Proliferation of open-source LLMs of “similar” strength to gpt-3.5-turbo
  - LLaMa/Alpaca/Vicuna
  - OA-Pythia, Dolly-Pythia, StabeLM, RedPajamas
- Access to model inference & training process will enable new hallucination prevention techniques
  - Prompt tuning, constrained decoding…
  - Alternative training strategies
- Supervisory systems and tools will be more impactful with weaker models
- Conclusions:
  - Open-source models are becoming viable for generative search
  - Hallucinations will remain a problem
  - Solving hallucination for weaker models will enable adoption across many applications

WHAT WOULD HELP

FUNDAMENTAL INNOVATIONS	Reasoning-optimized models, better open-source models, constrained decoding…
TOOLS FOR PRACTITIONERS	Toolbox of techniques to combat hallucinations of all types



06 ABOUT

COLIN HARMAN

Upcoming writing topics:

- Open-source LLM updates
- Understanding hallucination causes through QA benchmark mining
- Updates on anti-hallucination techniques

Upcoming projects:

- Hallucination playground
- OS anti-hallucination toolbox
- Crowdsourced hallucination dataset?

Get in touch!

1. LinkedIn
2. colin@hellonesh.io
3. hellonesh.io
4. Detroit, MI