/root signals

Can We Control GenAI Without Measuring It?

Oguzhan (Ouz) Gencoglu - Co-founder & Head of AI

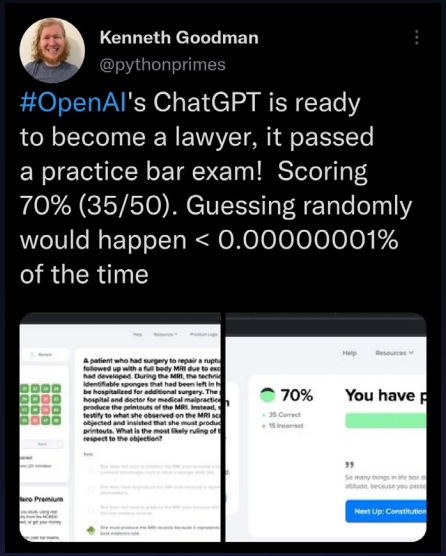
2 April 2025

No!

Agenda

- 2 Questions
  - Why is that RoI of GenAI is almost 0%?
  - Why do I still have to write emails?
- *Root Judge* & LUMI
- Common Misconceptions About LLM Judges

/root signals

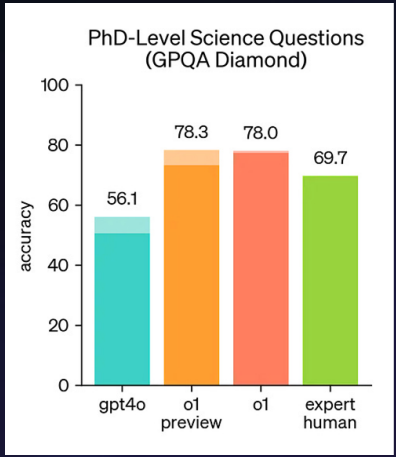


AI Passes U.S. Medical Licensing Exam

— Two papers show that large language models, including ChatGPT, can pass the USMLE

A.I. Chatbots Defeated Doctors at Diagnosing Illness

A small study found ChatGPT outdid human physicians when assessing medical case histories, even when those doctors were using a chatbot.



/root signals

Why is that so?

Non-deterministic systems

No self reflection

Semantic guidance

Failures

- Inaccuracy
- Hallucinations
- Biases
- Verbosity
- Diverging from business interests
- …

→ Multi-dimensional Failures

/root signals

Evals has always been important!

But it was not hard!

Now it is!

/root signals

SOLUTION:

Properly

tuned / optimized / calibrated

LLM-Judges!

/root signals

Root Judge

State-Of-The-Art Judge LLM For Evaluation & Hallucination Detection

Root Judge, a ground-breaking LLM that sets a new standard for reliable, customizable and locally-deployable evaluation models.

Try In Platform

Hugging Face

Couple of Misconceptions about LLM Judges

Misconception:

“Judges are there just to block unwanted responses“

REALITY:

- Many criteria are matters of degree, e.g. safety, relevance, and often even truthfulness – and can align with humans [1,2]
- Metric outputs [0…1] unlock optimization
- Guardrails are simply threshold-conditioned metrics
- Recent trend of inference-time search increases the need for evaluations for choosing candidates, with risks of hallucination increasing with increased turns

[1] https://arxiv.org/pdf/2306.05685
[2] https://arxiv.org/html/2403.03230v1

/root signals

Misconception:

“When foundation models become stronger, no judges (or evals) are any longer necessary”

REALITY:

- As foundation foundation models get better, the maximum performance does indeed improve
- But smaller and more energy-efficient models will always be needed – even if there were an AGI
- Also – historically, as models improve, we have always moved the goalposts

/root signals

Misconception:

“Judges are only for pre-production (development)"

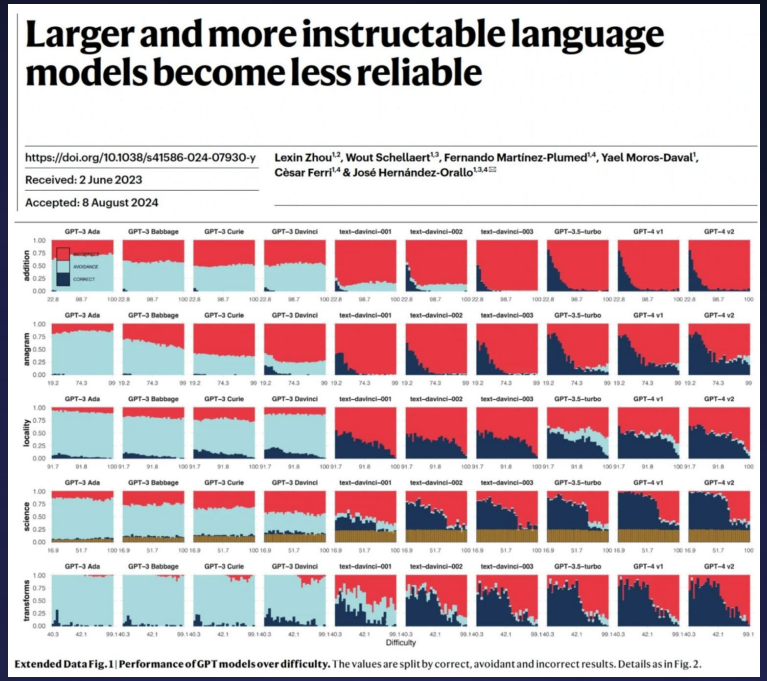
REALITY:

- LLMs can perform complex task before being flawless on very easy tasks
- There is no “reliable” zone one can identify
- This paradigm is new even to ML Experts

/root signals

Misconception:

“Judges are only for development, not for production"



Larger and more instructable language models become less reliable

https://doi.org/10.1038/s41586-024-07930-y

Received: 2 June 2023

Accepted: 8 August 2024

Lexin Zhou[1,2], Wout Schellaert[1,3], Fernando Martínez-Plumed[1,4], Yael Moros-Daval[1], Cèsar Ferri[1,4] & José Hernández-Orallo[1,3,4]✉

Extended Data Fig. 1 | Performance of GPT models over difficulty. The values are split by correct, avoidant and incorrect results. Details as in Fig. 2.

Zhou et al. 2024, Nature - 10.1038/s41586-024-07930-y

/root signals

Root Judge & LUMI

Take Away

1. LLMs are unreliable by nature
2. Lack of reliability is the main blocker for large scale adoption
3. A measurement layer around GenAI automations is needed
4. But metrics we want to measure are complex → LLM Judges
5. We need a systematic way to define, tune, and manage these Judges → Root Signals

/root signals

/ thank you