Promise and perils of deploying LLMs for aging research and clinical applications

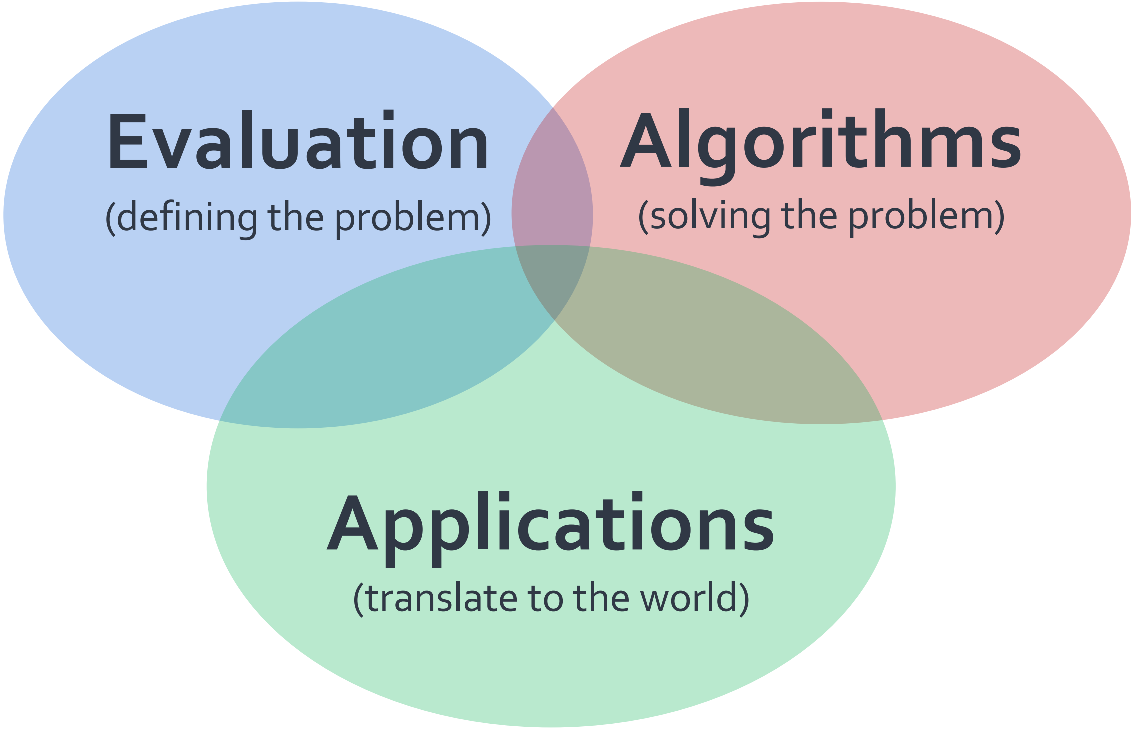
Daniel Khashabi

JOHNS HOPKINS UNIVERSITY

About me

Research agenda: developing theories and empirical approaches to make AI systems capable of robust and transparent communication

Research agenda: developing theories and empirical approaches to make AI systems capable of robust and transparent communication

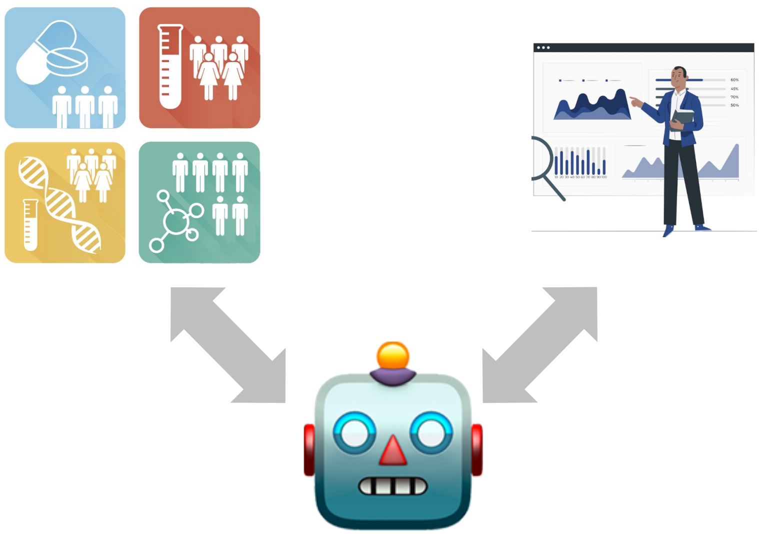


LLMs + Aging Research

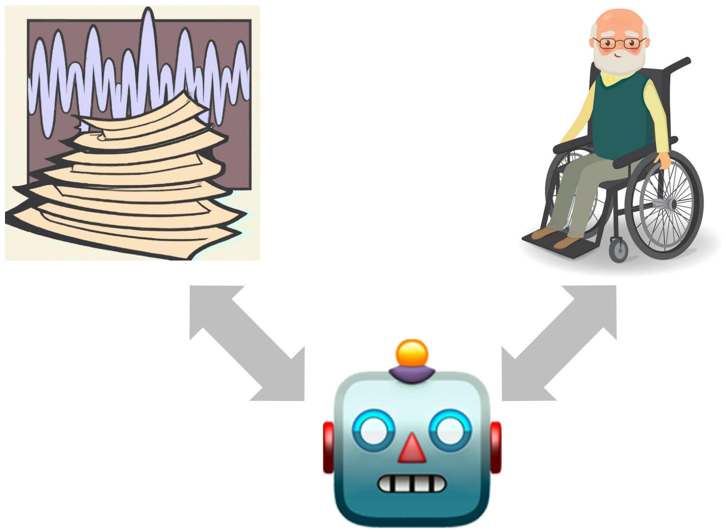
Applications

(translate to the world)

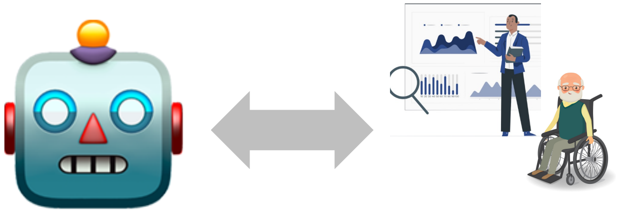
Research applications



Clinical applications

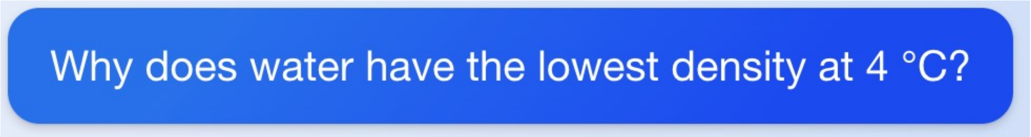


LLM failures

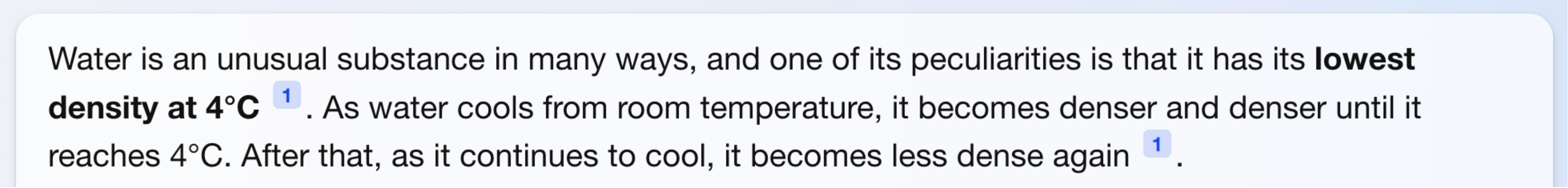


Bing

(fetched on Aug 30, 2023)



Why does water have the lowest density at 4 °C?



Water is an unusual substance in many ways, and one of its peculiarities is that it has its **lowest density at 4°C** [1]. As water cools from room temperature, it becomes denser and denser until it reaches 4°C. After that, as it continues to cool, it becomes less dense again [1].

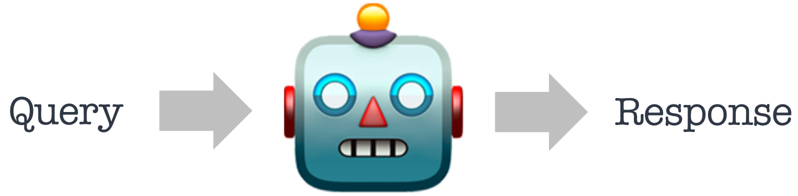
The boundaries of these failures are unknown.

⟹ safety problem

LLM failure vs. distributional properties of data

- **Controlled experiment:**
  Question accuracy for fixed relationship and varying subjects.

Q: Who was the director
of The Titanic?



Query

Response

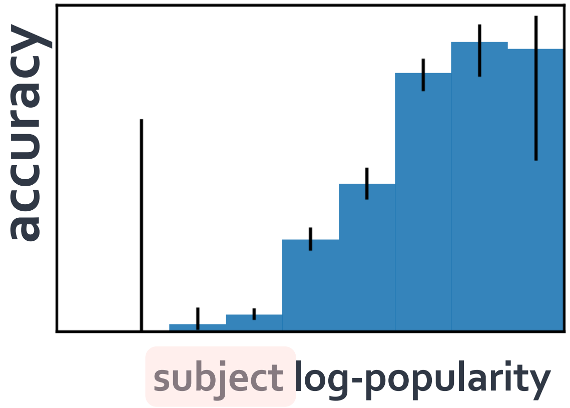
Hypothesis: Popularity predicts factual accuracy?

Mallen et al. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories, *ACL* 2023

LLM failure vs. distributional properties of data

- **Controlled experiment:**
  Question accuracy for fixed relationship and varying subjects.

Q: Who was the director of The Titanic?



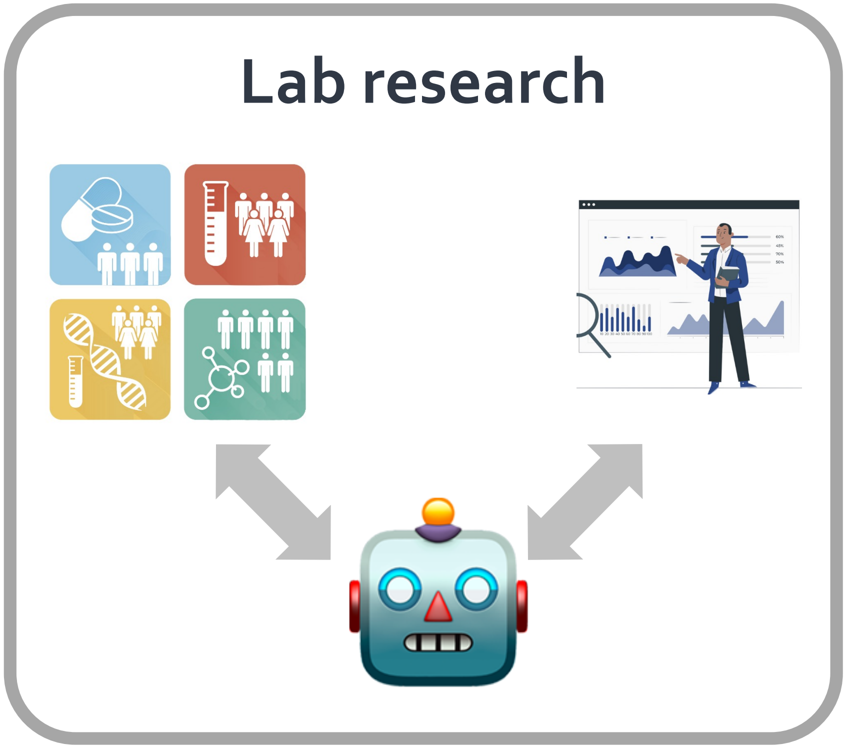
Factual accuracy of LLMs is positively correlated with "popularity" of information.

Mallen et al. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories, *ACL* 2023

Distributional properties of aging data

- Nature is full of rare events
    - Many disorders have a very low prevalence

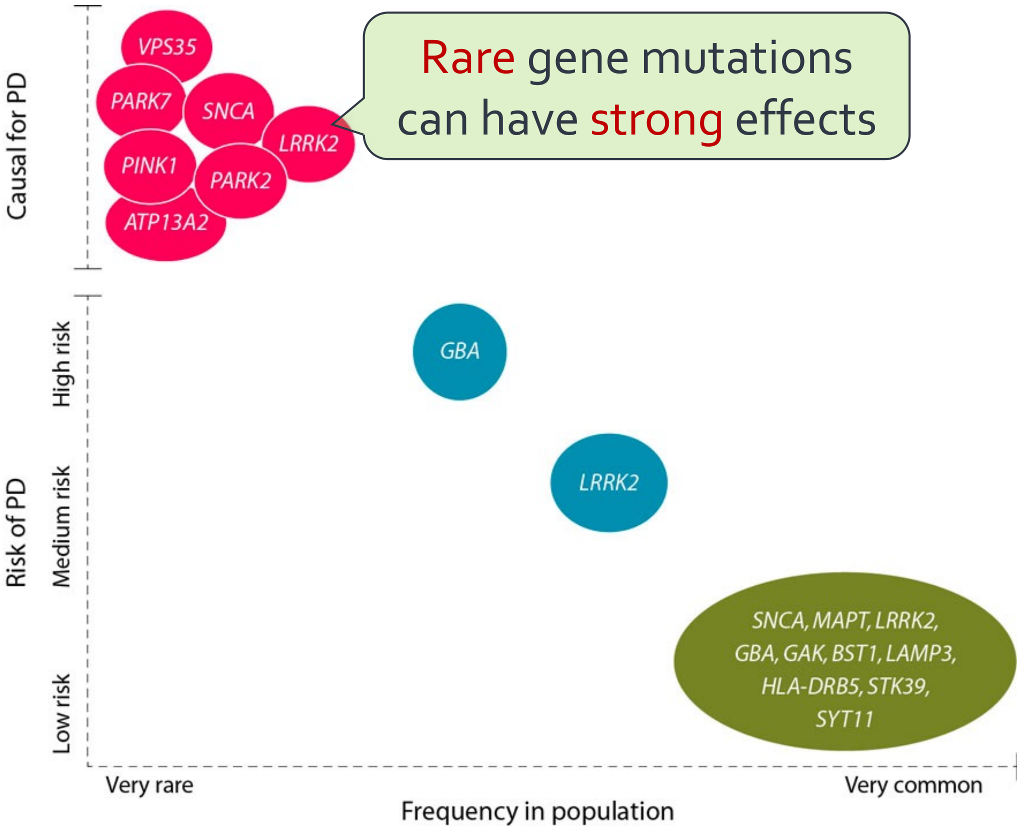
- Rare events are prevalent in aging research



Lab research

Distributional properties of aging data

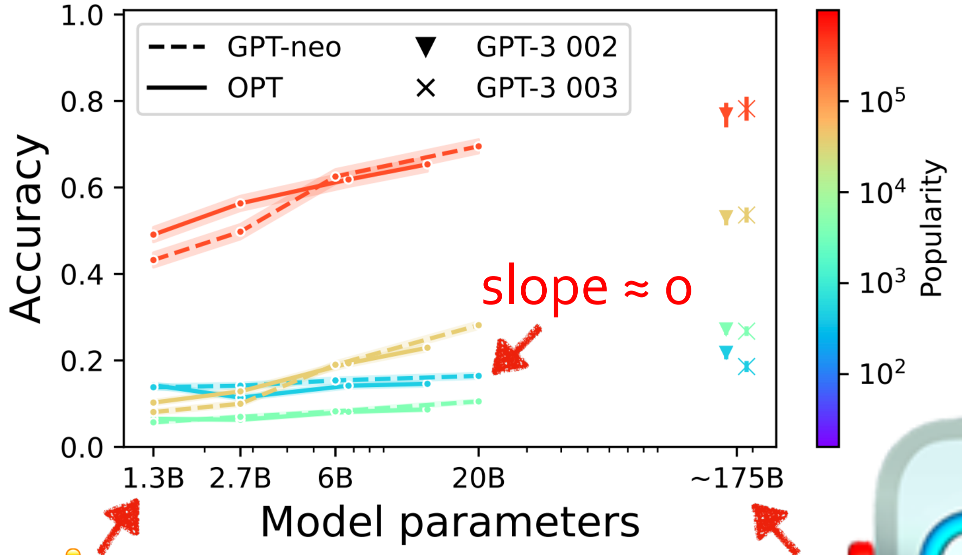
- Nature is full of rare events
  - Many disorders have a very low prevalence
- Rare events are prevalent in aging research
- LLMs are doomed to fail at discovering such rare events.



Rare gene mutations can have strong effects

van der Brug et al. "Parkinson's disease: from human genetics to clinical trials. *Science translational medicine* 2015

What about “scale”?

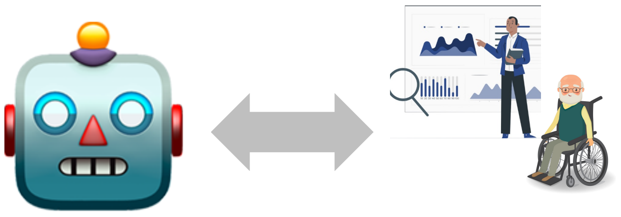


slope ≈ 0

Scaling models leads to little gains for rare events.

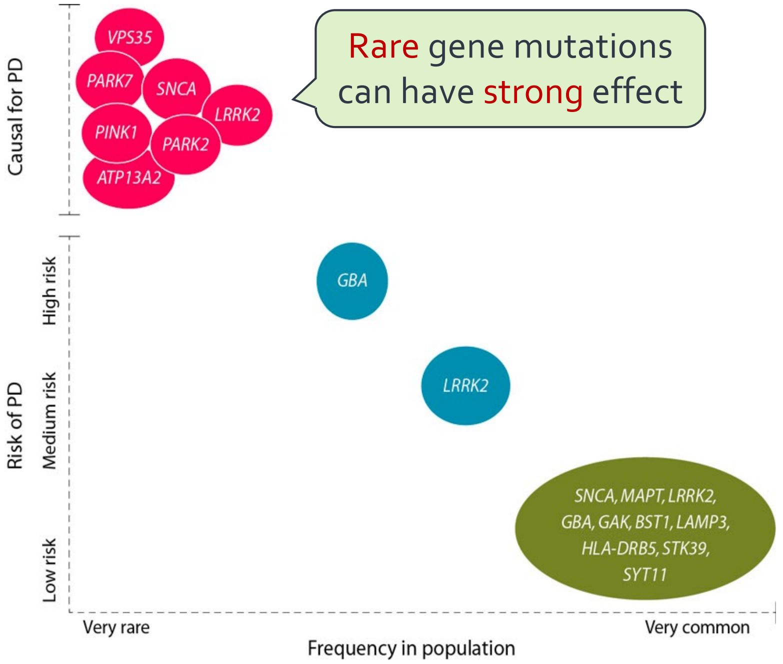
Mallen et al. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories, *ACL* 2023

Summary thus far



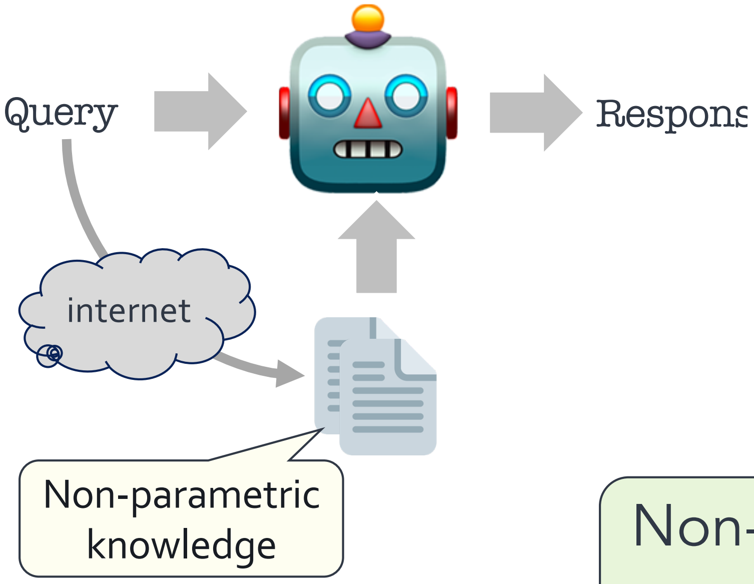
- Aging research involves rare events.
  - Gene mutations vs. PD
  - Funding gaps and biases

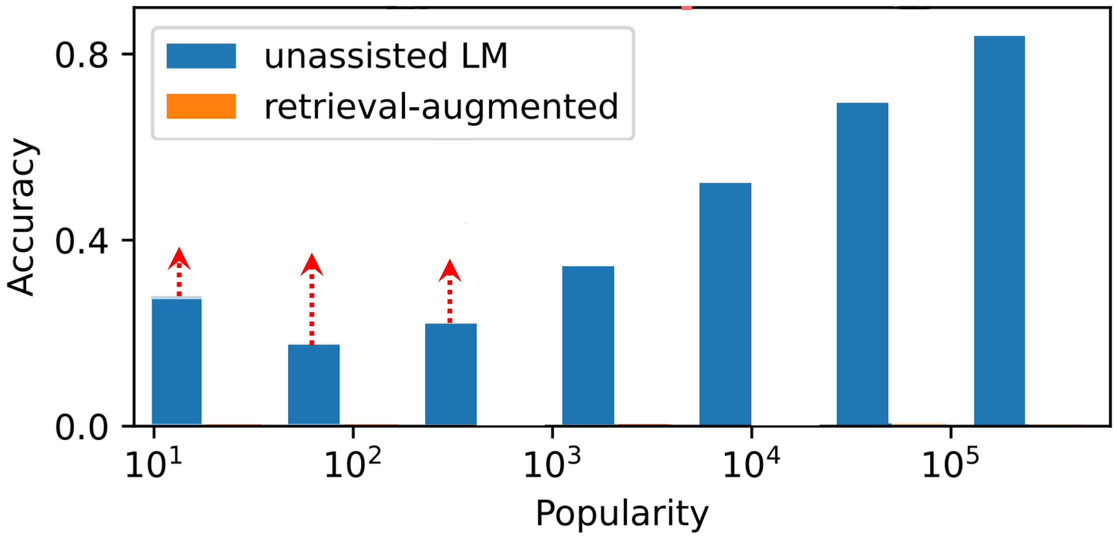
- LLM quality is strongly correlated with popularity of data.
  - Note, "popularity" is not an easily-quantifiable measure.



Rare gene mutations can have strong effect

Augmenting LLMs with relevant context





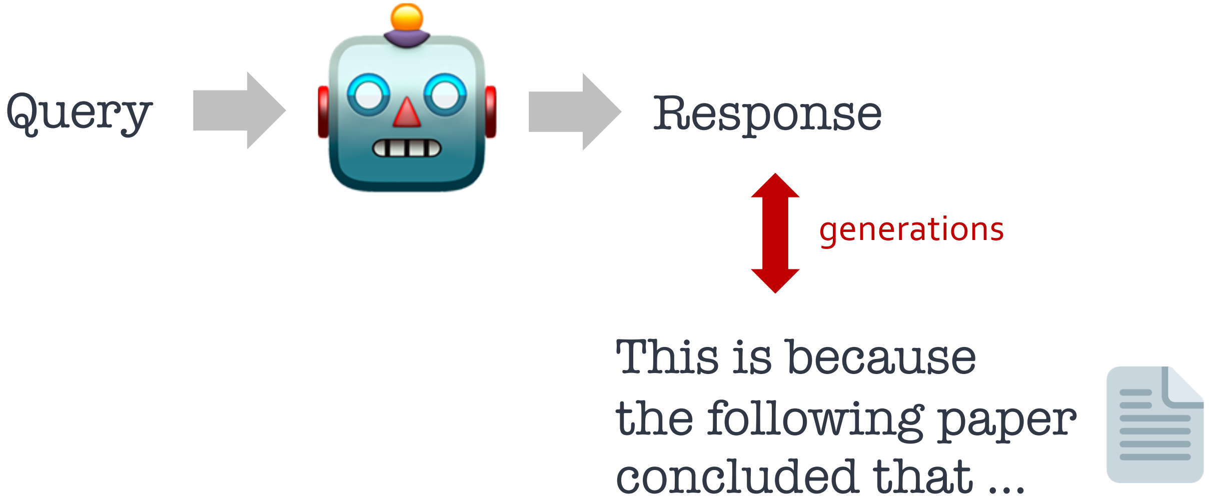
Non-parametric knowledge helps rare events, but may hurt popular phenomena.

Mallen et al. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories, *ACL* 2023

Gautier Izacard et al., “Unsupervised Dense Information Retrieval with Contrastive Learning” (2022)

Grounding LLM generations

- Allow LLMs to connect/attribute their generations to the real world.



Query

Response

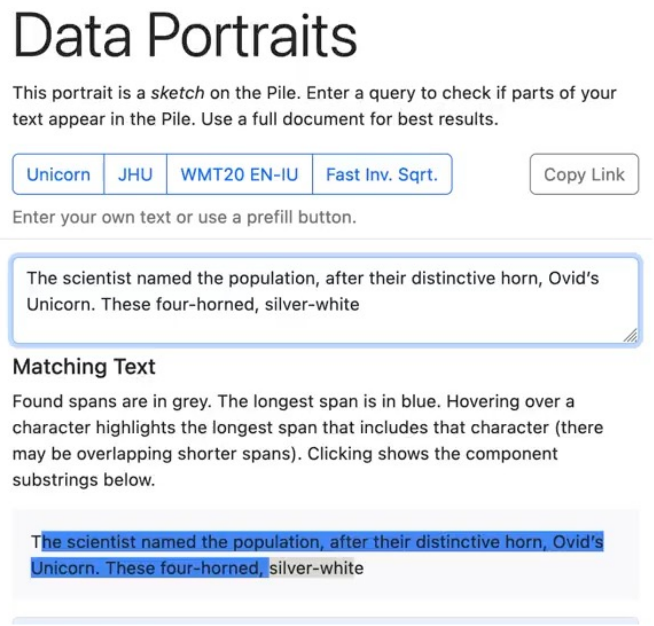
generations

This is because
the following paper
concluded that …

Verifiable Grounding LLM generations

- “Data Portraits”
  - Fast membership query in large corpus (whether a string belongs to your data)
  - Implemented via Bloom filter.

- Allows us to develop chatbots that are trained to “quote” (work in progress)
  - Tradeoff between reliability and creativity



Data Portraits

This portrait is a *sketch* on the Pile. Enter a query to check if parts of your text appear in the Pile. Use a full document for best results.

Enter your own text or use a prefill button.

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white

Matching Text

Found spans are in grey. The longest span is in blue. Hovering over a character highlights the longest span that includes that character (there may be overlapping shorter spans). Clicking shows the component substrings below.

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white

https://dataportraits.org

Marone and Van Durme. "Data portraits: Recording foundation model training data." *arXiv* 202).
Zhang et al. [work in progress]

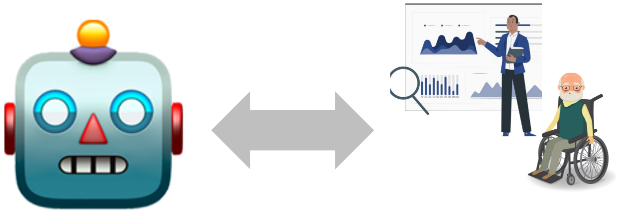
Key bottleneck for progress: feedback

- The success of LLMs is due to rich feedback signal.
    - Average humans are very good at telling apart good responses.

- Aging problems:
    - Humans do not necessarily know what is good or not.
    - Understanding causal connection between parameters and effects require real experiments over a time-horizon.
    - We do not necessarily know what we are looking for.

- OpenAI's future chatbots won't solve the aging problem.

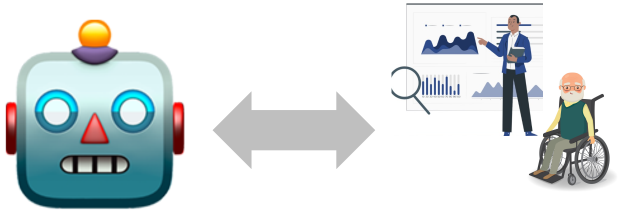
In the short term …



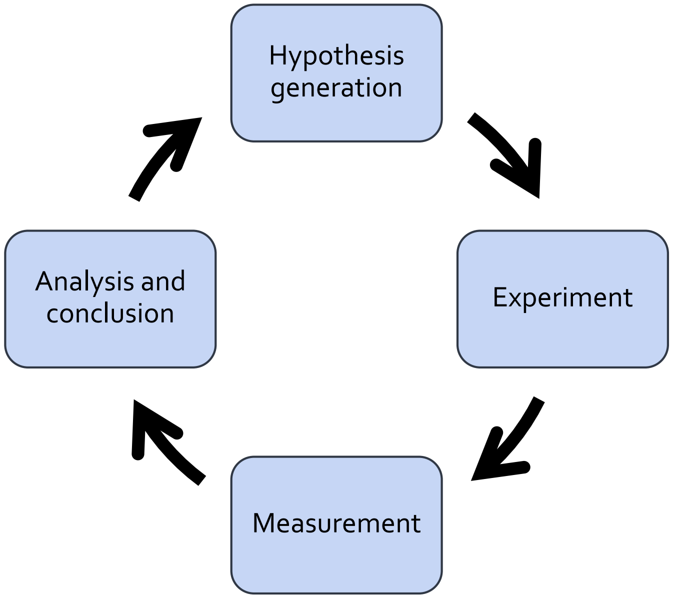
- LLMs provide value-add in certain niches of aging research.
  - This requires careful context-specific guardrails.
  - Examples:
    - LLMs helping crunch through tables
    - LLMs extract information from papers
    - …

- LLMs will remain to be brittle/inconsistent, on important problems.
  - The gains of "scale" will be minimal, in the short term.

In the long term …



- The key is to identify rich sources of feedback.
- E.g., LLMs as part of the clinical/medical research cycle.
  - Growing LLMs as part of data generation ecosystem.
  - Requires extensive safety considerations.



Hypothesis generation

Experiment

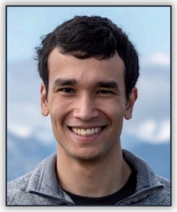
Measurement

Analysis and conclusion

Collaborators:

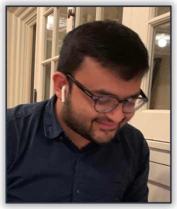
NIH CARD

JOHNS HOPKINS UNIVERSITY



EleutherAI

UNIVERSITY of WASHINGTON



Funding:

Ai2 Allen Institute for AI

amazon science

Office of Naval Research ONR Science & Technology

Intelligence Continues to be a Moving Target

- Every step forward, we realize there are new challenges.
- Unless there is a revolution outside AI (energy, hardware, etc.), we need a lot more innovations.

Academia has to go slow[er] and understand things.

- Slow enough to understand why things work or do not work.

- Today: Revisit two relevant technological pieces that deserve further deliberation.

Collaborators:

NIH CARD

JOHNS HOPKINS UNIVERSITY

EleutherAI

UNIVERSITY *of* WASHINGTON

Funding:

Ai2 Allen Institute for AI

amazon science

Office of Naval Research ONR Science & Technology