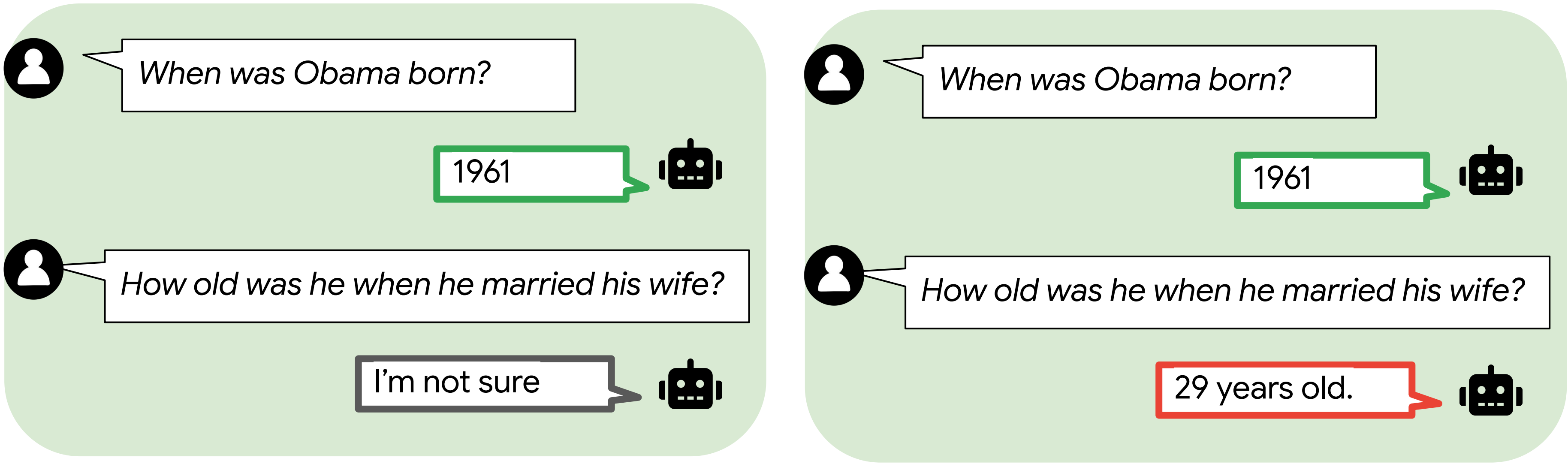
Beyond Factuality

Improving Trust and Reliability of Large Language Models

Gal Yona
November 9th, 2025
Workshop on Uncertainty-Aware NLP @ EMNLP 2025

Google Research

Measuring Factuality in LLMs

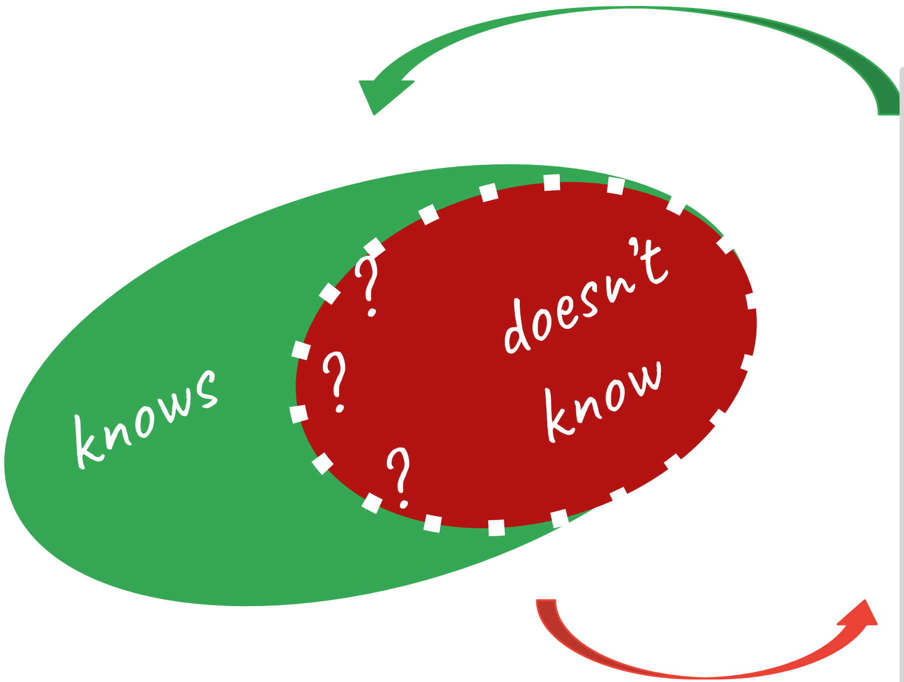


correct | **attempted** = 2 / (2 + 1) = 0.667

Google Research

Improving Factuality in LLMs

correct | **attempted**



Mostly determined in **pre-training**

All questions the model ***could answer correctly*** *(i.e., has required knowledge)*

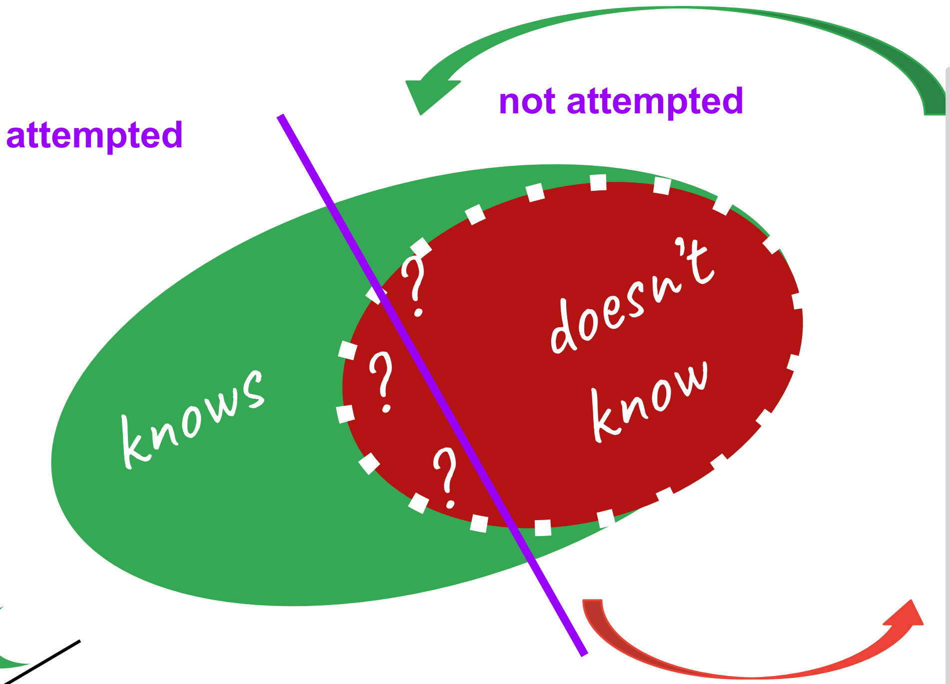
All questions the model would answer incorrectly, if forced to provide an answer (i.e., does not have required knowledge)

Improving Factuality in LLMs

correct | attempted

Mostly determined in **post-training** or **configurable**

All questions the model ***chooses to provide an answer*** *for*

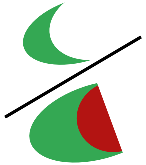


Mostly determined in **pre-training**

All questions the model ***could answer correctly*** *(i.e., has required knowledge)*

All questions the model would answer incorrectly, if forced to provide an answer (i.e., does not have required knowledge)

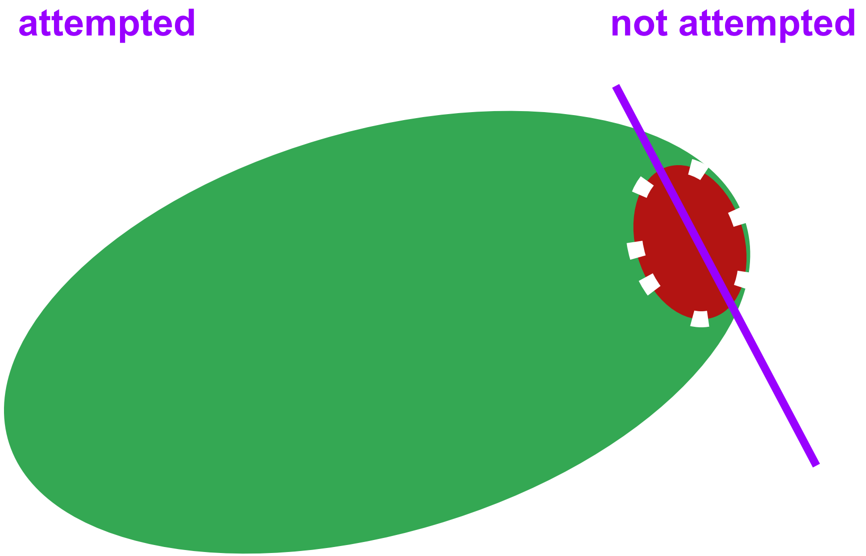
correct | **attempted** =



Option 1

correct | **attempted**

The model is no better at drawing the line between known and unknown;
but is a lot more knowledgeable

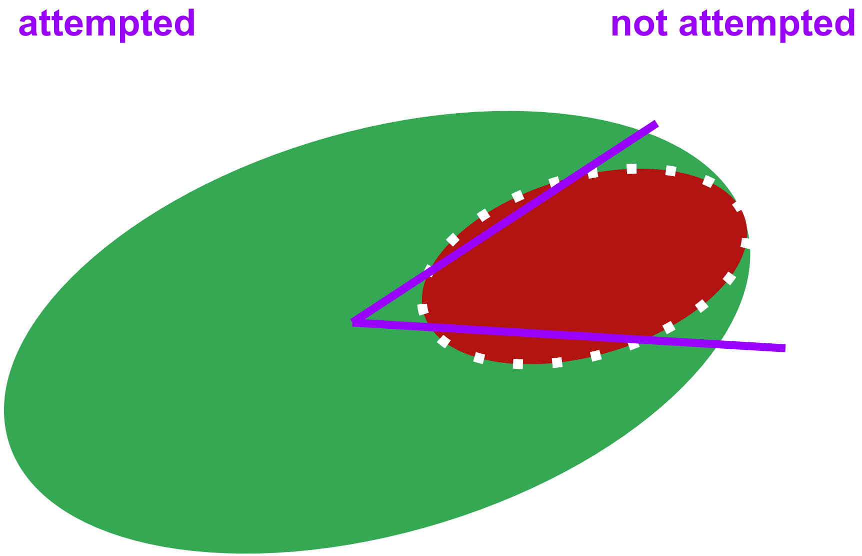


Google Research

correct | **attempted**

Option 2

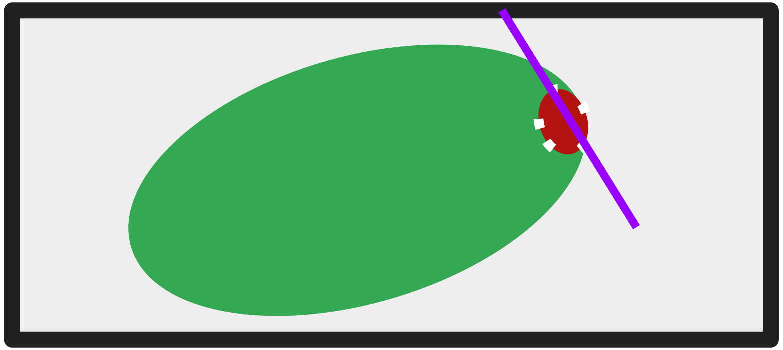
The model is as knowledgeable, but is much better at identifying what it doesn’t know



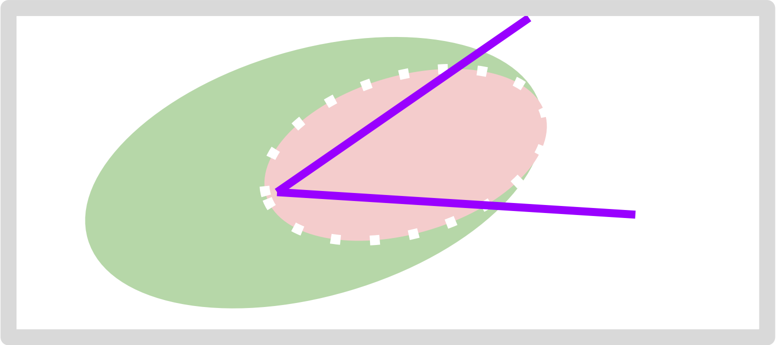
Improving Factuality in LLMs

correct | **attempted**

Option 1



Option 2



Google Research

SimpleQA Verified [1] Leaderboard

Rank	Model	F1-Score	ΔSimpleQA (%pt)	Accuracy	Acc.|Attempted	Attempted
1	**Gemini 2.5 Pro**	**55.6**	0.5	55.3	55.9	98.9
2	GPT 5	52.3	1.8	50.9	53.8	94.6
3	o3	51.9	1.9	51.6	52.0	99.3
4	GPT 4.1	39.9	−1.0	39.8	40.1	99.3
5	GPT 4o	34.9	−3.5*	34.4	35.5	97.0
6	DeepSeek R1 (0528)	33.3	1.4	32.7	33.9	96.4
7	Claude Opus 4	28.3	−4.0*	19.2	54.1	35.5
8	Gemini 2.5 Flash	28.2	−1.4	27.8	28.7	96.9
9	GPT 5 Mini	24.6	1.1	17.3	42.8	40.4
10	o4-mini	23.4	2.9*	23.0	23.8	96.5

*Leading models **virtually never punt**, (Attempted ≅ 100.0), despite having very high error rates (Accuracy ≪ 100.0).*

the old style of “hallucinations research” via self-calibration is probably going to die down. [...]

I still think the idea is very correct but **empirically it doesn't seem like anyone made great progress on that in the past two years** [...]

The much better thing than self-calibration is **improve factuality by allowing language models to browse the internet**. In the past, language models would hallucinate easily on queries like “what papers Barret Zoph wrote” but now with browsing they can answer those easy factual questions pretty well.

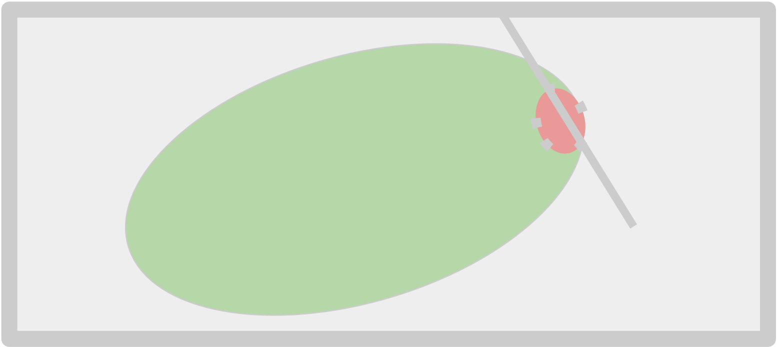
Jason Wei @ X

Google Research

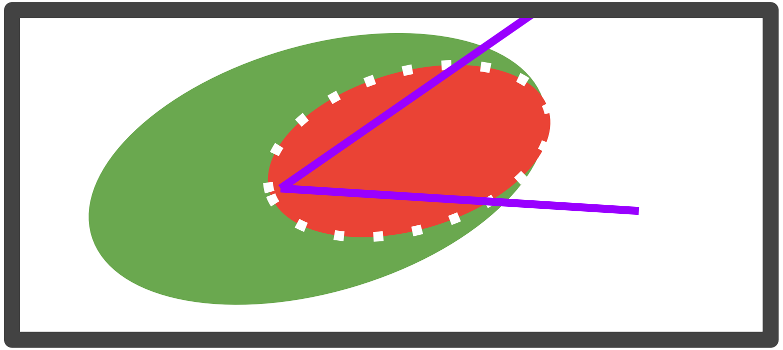
correct | **attempted**

Improving Factuality in LLMs

Option 1



Option 2



Factuality → Faithfulness

Tell me about Mark Bils.

Mark Bils is a macroeconomist at the University of Rochester. He was born on March 22, 1958.

Factuality → Faithfulness

Tell me about Mark Bils.

Mark Bils is a macroeconomist at the University of Rochester. He was born on March 22, 1958.

Don't say **incorrect** things

Mark Bils (**December 1, 1958**) is a macroeconomist at the University of Rochester.

Wikipedia
https://en.wikipedia.org › wiki › Mark_Bils

Mark Bils - Wikipedia

Don't say **decisively** things you are **not sure** about

Factuality → Faithfulness

Tell me about Mark Bils.

Mark Bils is a macroeconomist at the University of Rochester. He was born on March 22, 1958.

Don't say **incorrect** things

Mark Bils (**December 1, 1958**) is a macroeconomist at the University of Rochester.

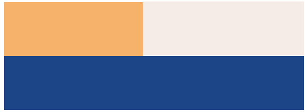
Wikipedia
https://en.wikipedia.org › wiki › Mark_Bils

Mark Bils - Wikipedia

Don't say **decisively** things you are **not sure** about

confidence

decisiveness



When was Mark Bils born?

May 15, 1958.

April 27, 1958.

⋮

Google Research

Factuality → Faithfulness

Tell me about Mark Bils.

Mark Bils is a macroeconomist at the University of Rochester. He was born on March 22, 1958.

Don't say **incorrect** things

Mark Bils (**December 1, 1958**) is a macroeconomist at the University of Rochester.

Wikipedia
https://en.wikipedia.org › wiki › Mark_Bils

Mark Bils - Wikipedia

Don't say **decisively** things you are **not sure** about

Definition 1 (Response Faithfulness) *Given a model M, we measure the **faithfulness** of a response $\mathbf{R}$ to a query $\mathbf{Q}$ as:*

$$\texttt{faithfulness}_M(\mathbf{R};\mathbf{Q}) \equiv$$

$$1 - \sum_{A \in \mathcal{A}(\mathbf{R})} |\texttt{dec}(A;\mathbf{R},\mathbf{Q}) - \texttt{conf}_M(A)| \quad (1)$$

*where $\texttt{dec}(A;\mathbf{R},\mathbf{Q})$ quantifies how **decisive** the assertion A is made in $\mathbf{R}$ and $\texttt{conf}_M(A)$ quantifies M's intrinsic uncertainty regarding A.*

In principle, can be obtained (e.g. doesn't require knowing when Mark Bils was really born)

Google Research

Factuality → Faithfulness

Tell me about Mark Bils.

Mark Bils is a macroeconomist at the University of Rochester. He was born on March 22, 1958.

Don't say **incorrect** things

Mark Bils is a macroeconomist at the University of Rochester. ~~He was born on March 22, 1958.~~

Factuality → Faithfulness

Tell me about Mark Bils.

Mark Bils is a macroeconomist at the University of Rochester. He was born on March 22, 1958.

Don't say **incorrect** things

Mark Bils is a macroeconomist at the University of Rochester. ~~He was born on March 22, 1958.~~

Don't say **decisively** things you are **not sure** about

Answer at appropriate **granularity**

Communicate uncertainty **linguistically**

Google Research

Factuality → Faithfulness

Tell me about Mark Bils.

Mark Bils is a macroeconomist at the University of Rochester. He was born on March 22, 1958.

Don't say **incorrect** things

Mark Bils is a macroeconomist at the University of Rochester. ~~He was born on March 22, 1958.~~

Don't say **decisively** things you are **not sure** about

Answer at appropriate **granularity**

Mark Bils is a macroeconomist at the University of Rochester. He was born in 1958.

Communicate uncertainty **linguistically**

Mark Bils is a macroeconomist at the University of Rochester. I think he was born on March 22, 1958, but I'm not sure.

How good are frontier LLMs at faithful generation?

Do they choose to answer at a level of granularity that matches their uncertainty?

Do they express their uncertainty in natural language?

How good are frontier LLMs at faithful generation?

Do they choose to answer at a level of granularity that matches their uncertainty?

Narrowing the Knowledge Evaluation Gap: Open-Domain Question Answering with Multi-Granularity Answers (ACL 2024)

Do they express their uncertainty in natural language?

Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words? (EMNLP 2024)

Roee Aharoni

Mor Geva

How good are frontier LLMs at faithful generation?

Do they choose to answer at a level of granularity that matches their uncertainty?

Narrowing the Knowledge Evaluation Gap: Open-Domain Question Answering with Multi-Granularity Answers (ACL 2024)

Do they express their uncertainty in natural language?

Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words? (EMNLP 2024)

Roee Aharoni

Mor Geva

Initial motivation

- We observed that for questions of the form "When was {X} born?" (for tail entity X, e.g. Mark Bils), models tend to answer with the full date of birth (e.g. May 18, 1961), despite significant uncertainty about the specific date (May 18? May 8? etc)

- We conjectured that this is a result of a learned preference for a target **output format** (e.g. full date of birth) over **factuality**, even in the presence of extreme uncertainty

But… no direct way to evaluate 😞

Standard recipe for QA evaluation (comparing predicted answer to a set of "gold answers" via lexical matching) is **Ill-suited** for this purpose:

- "Gold answers" are *single-granularity* (typically: most specific answer, maybe w/ aliases)
- Existing metrics do not distinguish between **fine-grained** and **coarse-grained** answers

Enter: GRANOLA-QA

New in **GRANOLA** (**Gran**ularity **o**f **La**bels) **QA**:

1. Gold labels are *ordered*
2. New metrics
   a. Accuracy: Does the predicted answer match against *some* answer?
   b. Informativeness: Assign higher score for matching against a *finer-grained* answer.

Question: Where was Leslie Ash born?

Gold Answers:

Standard QA { Clapham }

GRANOLA QA [Clapham, London, **England**]

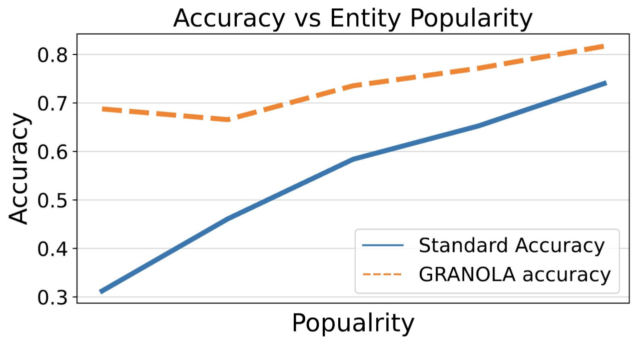
multiple granularity levels

GRANOLAEntityQuestions:
12k entity-centric questions with ±3 **multi-granularity answers** per question

Question	GRANOLA Answers
"Where was Fiona Lewis born?"	Westcliff-on-Sea; Essex; England
"What music label is Courage represented by?"	Rock Records; a Taiwanese record label
"Who is August von Hayek's child?"	Friedrich Hayek; an economist
"Who is the author of The Adding Machine?"	Elmer Rice; an American playwright; a playwright
"Where was Toby Shapshak educated?"	Rhodes University; Makhanda, South Africa; South Africa

Table 1: Examples from GRANOLA-EQ. Answers are separated by a semicolon and listed fine-to-coarse. The first answer is the original answer in ENTITYQUESTIONS; subsequent answers were generated (see §3.1).

Main Takeaways



👎 **Failures of modern LLMs:** LLMs consistently answer at a level of granularity that does not match their uncertainty.

💭 **A gap in how we evaluate knowledge in LLMs:** Accuracy w.r.t multi-granularity answer set remains steady for tail entities, suggesting models still know about entities - just coarser information.

💪 **A real & interesting "middle ground" between punting & answering:** Decoding strategies to "merge" sampled response into a coarser answer yield better accuracy with fewer IDKs.

Google Research

How good are frontier LLMs at faithful generation?

Do they choose to answer at a level of granularity that matches their uncertainty?

Narrowing the Knowledge Evaluation Gap: Open-Domain Question Answering with Multi-Granularity Answers (ACL 2024)

Do they express their uncertainty in natural language?

Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words? (EMNLP 2024)

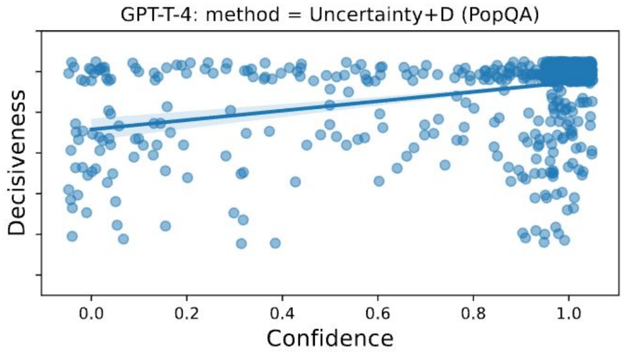
Roee Aharoni

Mor Geva

Main Takeaways

👎 **Failures of modern LLMs:** With standard decoding, models never hedge their answers (decisiveness = 1), despite even the best models having non-negligible uncertainty (confidence < 1.0)

👹 **Prompting is not a panacea:** Prompting the model to convey uncertainty does induce some hedged answers – but not enough, and not in the right places – resulting in negligible improvement in faithfulness.



Google Research

Summary

👍 Unlike **factuality** (*"never say incorrect things"*), **faithful generation** (*"never decisively say things you are not confident about"*) is an **information-theoretically feasible desiderata for trustworthy LLMs.**

😢 Modern LLMs are not explicitly trained for this, and are currently poor at faithful generation.

Next Steps & Open Problems

💥 **Tool-use** What does uncertainty expression look like when LLMs use tools? Uncertainty is no longer over the model's knowledge, but now also over external APIs (e.g. Google Search) and how the models interact with those APIs. How do we measure confidence? How do we determine appropriate hedging language?

💥 **Beyond Factuality** Even for coding or math problems, we want to be able to highlight parts of response the model is uncertain about. Generalize faithfulness beyond fact-seeking prompts.

💥 **Uncertainty vs Sycophancy** Can faithfulness help improve sycophantic behavior? If the model is faithful to a stable answer distribution (e.g. 50% A vs 50% B), it shouldn't matter that the user is saying "Are you sure? I really think the answer is A".

💥 **Measuring informativeness "in the wild"** How do we quantify the information gain different (correct) responses to the same query provide? How do we align this in a single metric across various types of questions?

Thank You

Gal Yona
Research Scientist
galyona@google.com

[1] SimpleQA Verified: A Reliable Factuality Benchmark to Measure Parametric Knowledge
https://arxiv.org/abs/2509.07968

[2] Narrowing the Knowledge Evaluation Gap: Open-Domain Question Answering with Multi-Granularity Answers
https://arxiv.org/pdf/2401.04695

[3] Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words?
https://arxiv.org/pdf/2405.16908

Google Research