Google Research

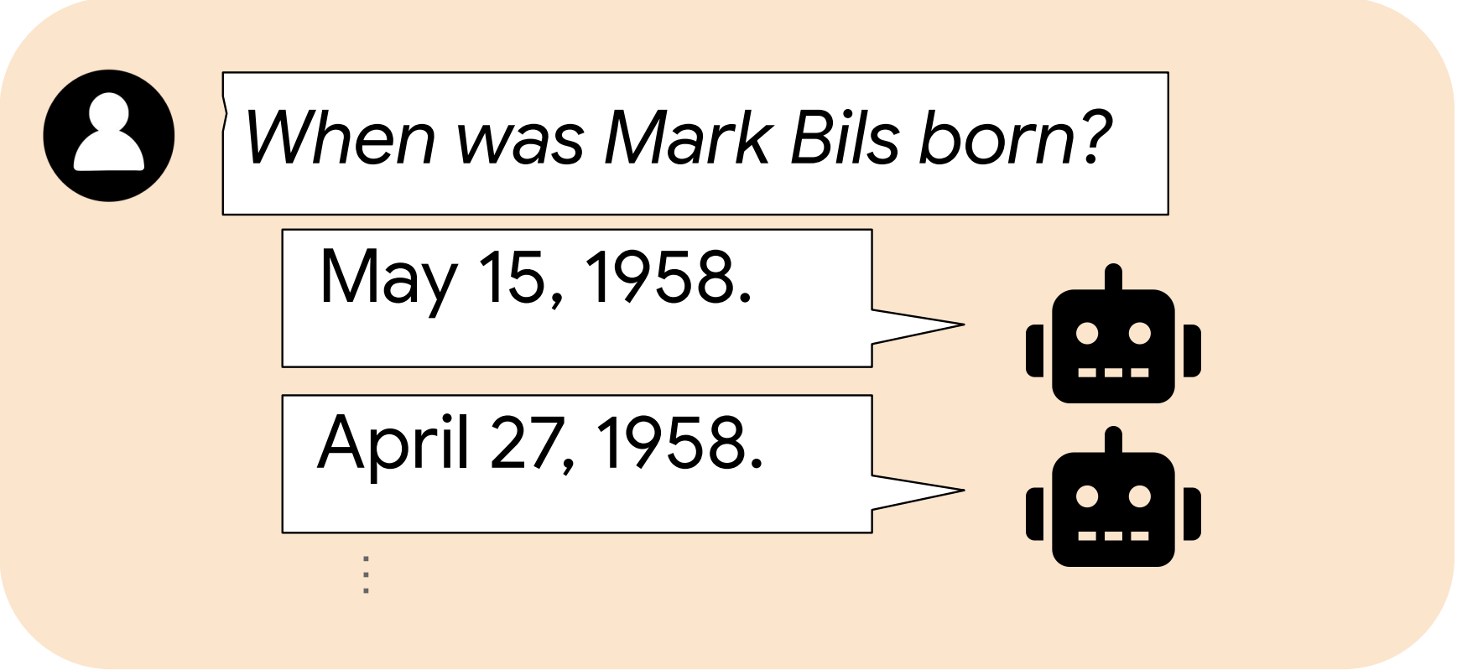
TEL AVIV UNIVERSITY אוניברסיטת תל אביב

Can Large Language Models Faithfully Convey their Intrinsic Uncertainty in Words?

Gal Yona[1], Roee Aharoni[1], Mor Geva[1,2]

[1]Google Research, [2]Tel Aviv University

Response Faithfulness: Trustworthiness *Beyond Factuality*


When was Mark Bils born?

May 15, 1958.

April 27, 1958.

⋮

Tell me about Mark Bils.

Factuality

Don't say **inaccurate** things

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

Not confident yet decisive

Faithful Generation

Answer at appropriate **granularity**

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

Confident and decisive

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. ✓

Less confident, but also less decisive

? **Are modern LLMs good at faithful generation?**

Evaluation setup

Definition 1 (Faithful Response Uncertainty) *For a query* $\mathbf{Q}$ *and a response* $\mathbf{R}$ *generated by a model* M, *the faithfulness of* $\mathbf{R}$ *with respect to* M*'s intrinsic confidence is given by:*

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

where $\texttt{dec}(A;\mathbf{R},\mathbf{Q}) \in [0,1]$ *quantifies the decisiveness of the assertion* A *in* $\mathbf{R}$ *and* $\texttt{conf}_M(A) \in [0,1]$ *quantifies the intrinsic uncertainty of* M *regarding* A.

Data: Knowledge-intensive QA datasets (NaturalQuestions & PopQA)

Models: Variety of models (Gemini, GPT)

Metric: CMFG

- E [faithfulness(R) | conf(R) = v]
- Baseline value: 0.5 (choose decisiveness independently of query)

Methods: Various prompting strategies

- **Vanilla**: standard QA prompt
- **Granularity**: instruct model to answer at appropriate granularity
- **Uncertainty**: instruct model to convey uncertainty linguistically
  - **+D**: include model-specific demonstrations

Results

	PopQA					Natural Questions				
Method	**GemNano**	**GemPro**	**GemUltra**	**GPT-T-3.5**	**GPT-T-4**	**GemNano**	**GemPro**	**GemUltra**	**GPT-T-3.5**	**GPT-T-4**
Vanilla	0.52	0.53	0.54	0.52	0.53	0.54	0.54	0.54	0.54	0.57
Granularity	0.51	0.52	0.53	0.52	0.53	0.54	0.53	0.54	0.54	0.54
Uncertainty	0.51	0.57	0.70	0.53	0.58	0.53	0.56	0.59	0.54	0.57
Uncertainty+	0.52	0.56	0.53	0.57	0.63	0.54	0.53	0.54	0.55	0.57

Table 1: **State of the art models struggle at faithfully communicating uncertainty:** cMFG results for each of the methods we test (higher is better). All models perform poorly, with cMFG close to the baseline value of 0.5.

1 Without special instructions, LLMs **never hedge their answers** (**decisiveness = 1**), despite even the best models having some **uncertainty** (**confidence < 1.0**)

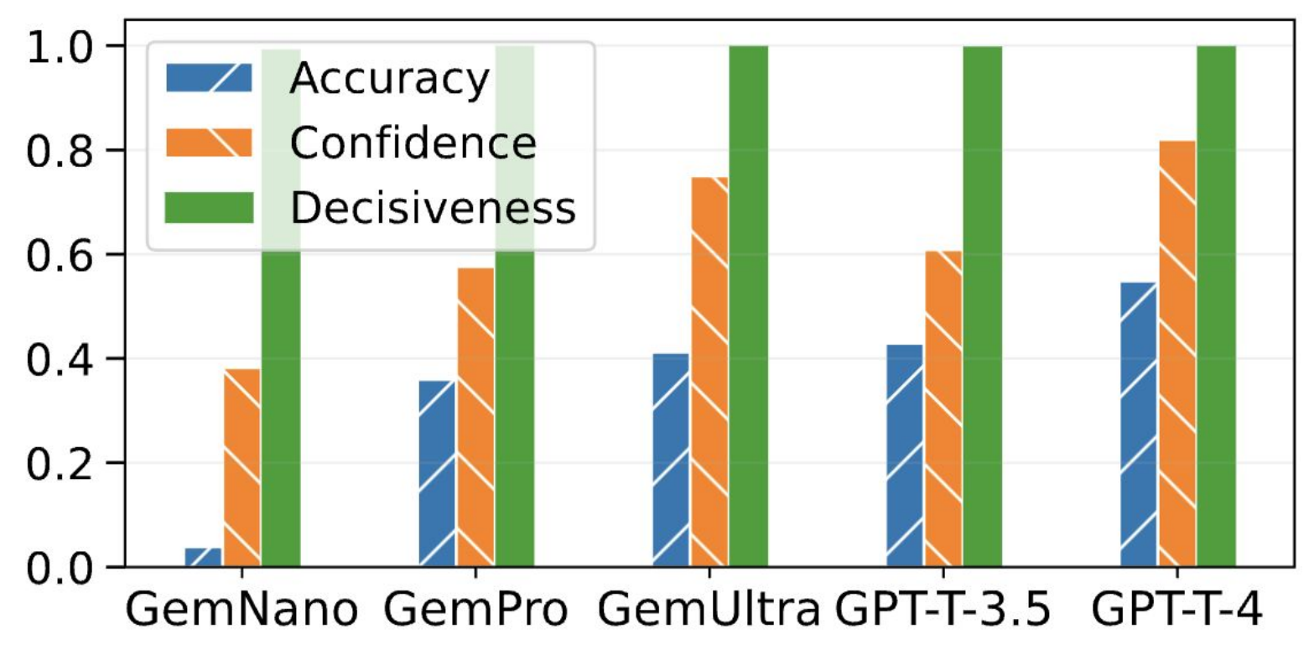
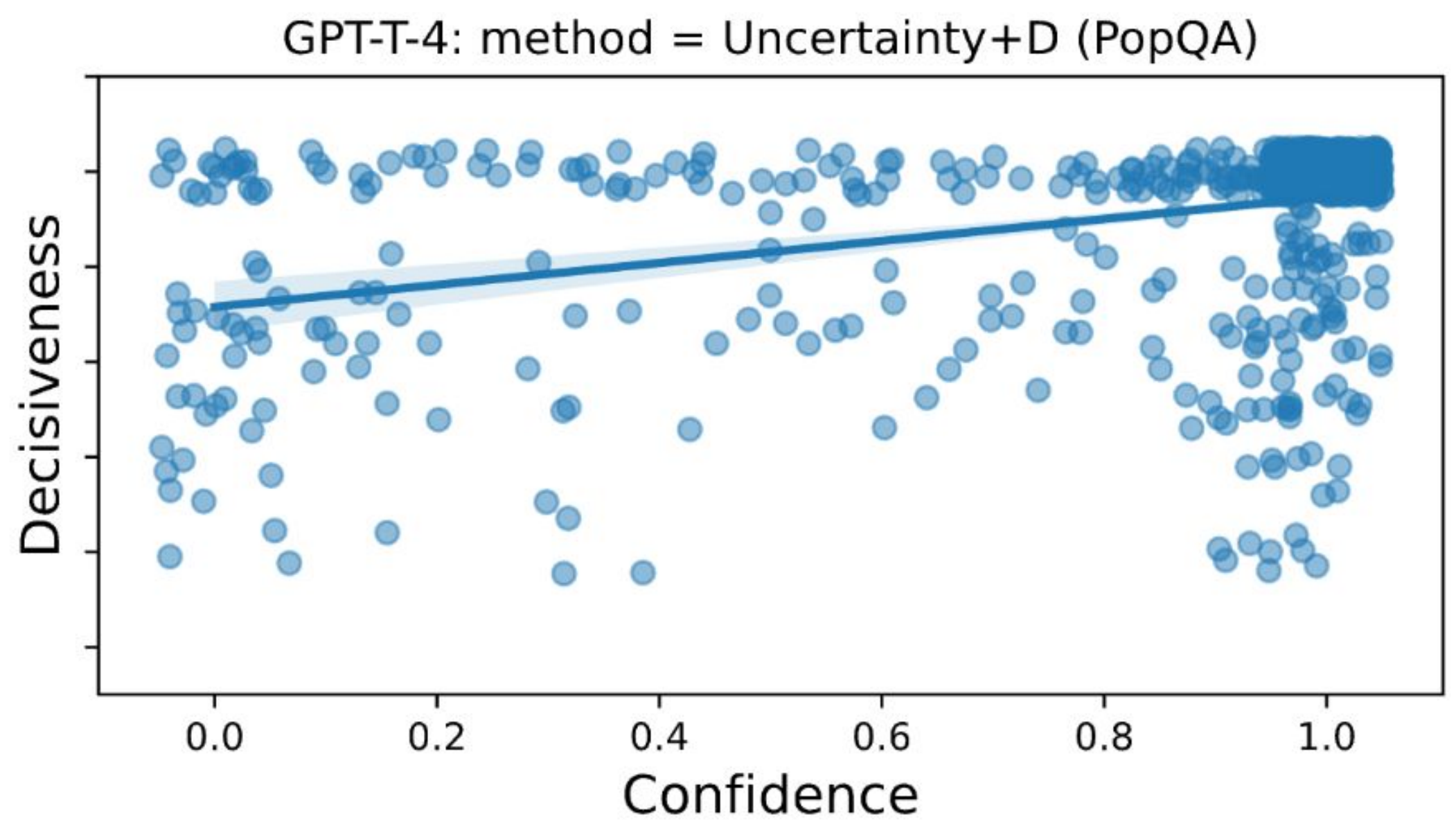


Figure 3: Mean accuracy, confidence, and decisiveness scores for **Vanilla** on PopQA (results on NQ show similar trends, see §B). Even the most accurate models answer decisively, despite non-trivial uncertainty.



2 SOTA LLMs **cannot be easily steered towards faithfully expressing their uncertainty** via prompting.

Figure 4: **Weak correlation between decisiveness and confidence:** We plot decisiveness (y-axis) vs confidence (x-axis) for two of the best performing *(model, method, dataset)* combinations (see Table 1). We see that these methods succeed at slightly improving cMFG (beyond the 0.5 baseline) by inducing some non-decisive answers, but the correlation between decisiveness and confidence is weak.

! **Our evaluations reveal that modern LLMs perform poorly at the task of faithfully conveying their intrinsic uncertainty, stressing the need for better alignment techniques towards ensuring trustworthiness in LLMs.**