Verbal Epistemic Uncertainty Estimation for Numeric Values with GPT-3

BACHELOR THESIS BY JOSCHKA BRAUN

STRUCTURE

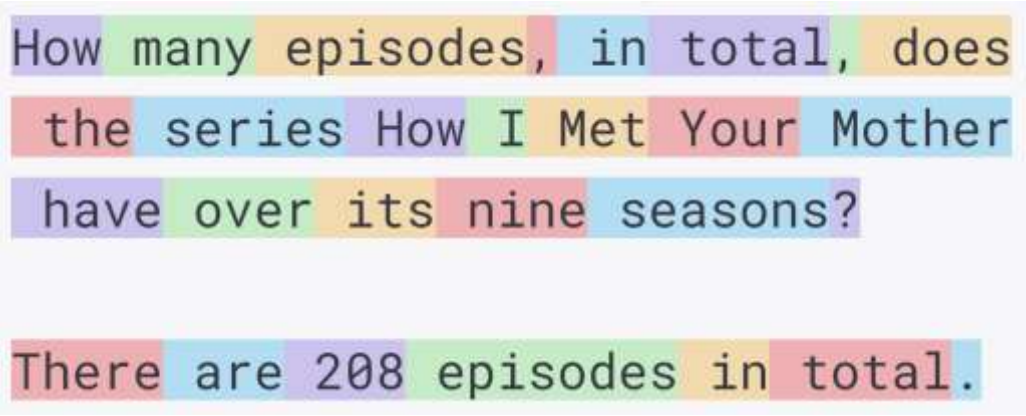
01. KEY CONCEPTS

02. RESEARCH QUESTIONS

03. MAIN FINDINGS

04. OUTLOOK

GPT-3

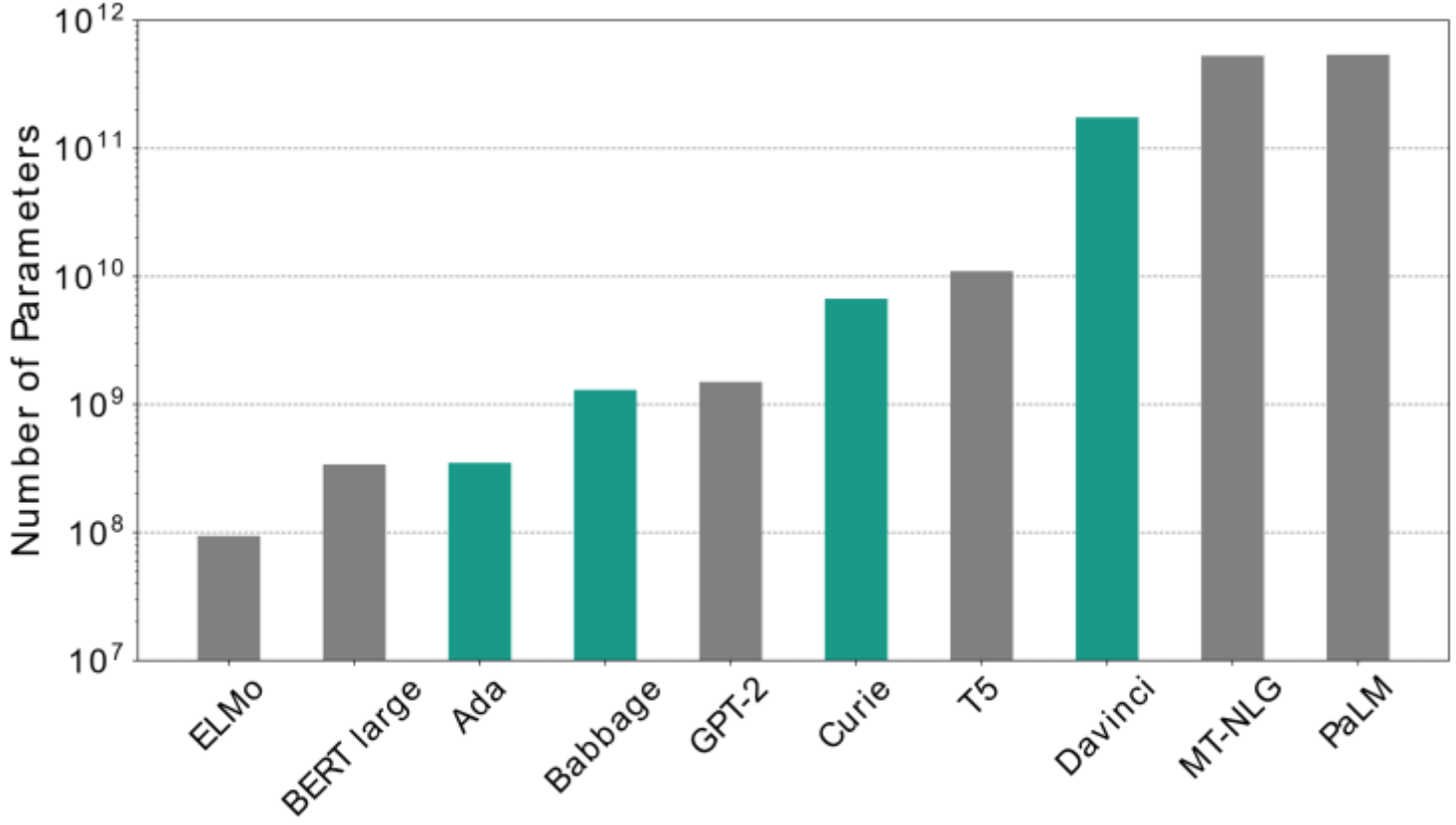


Prompt

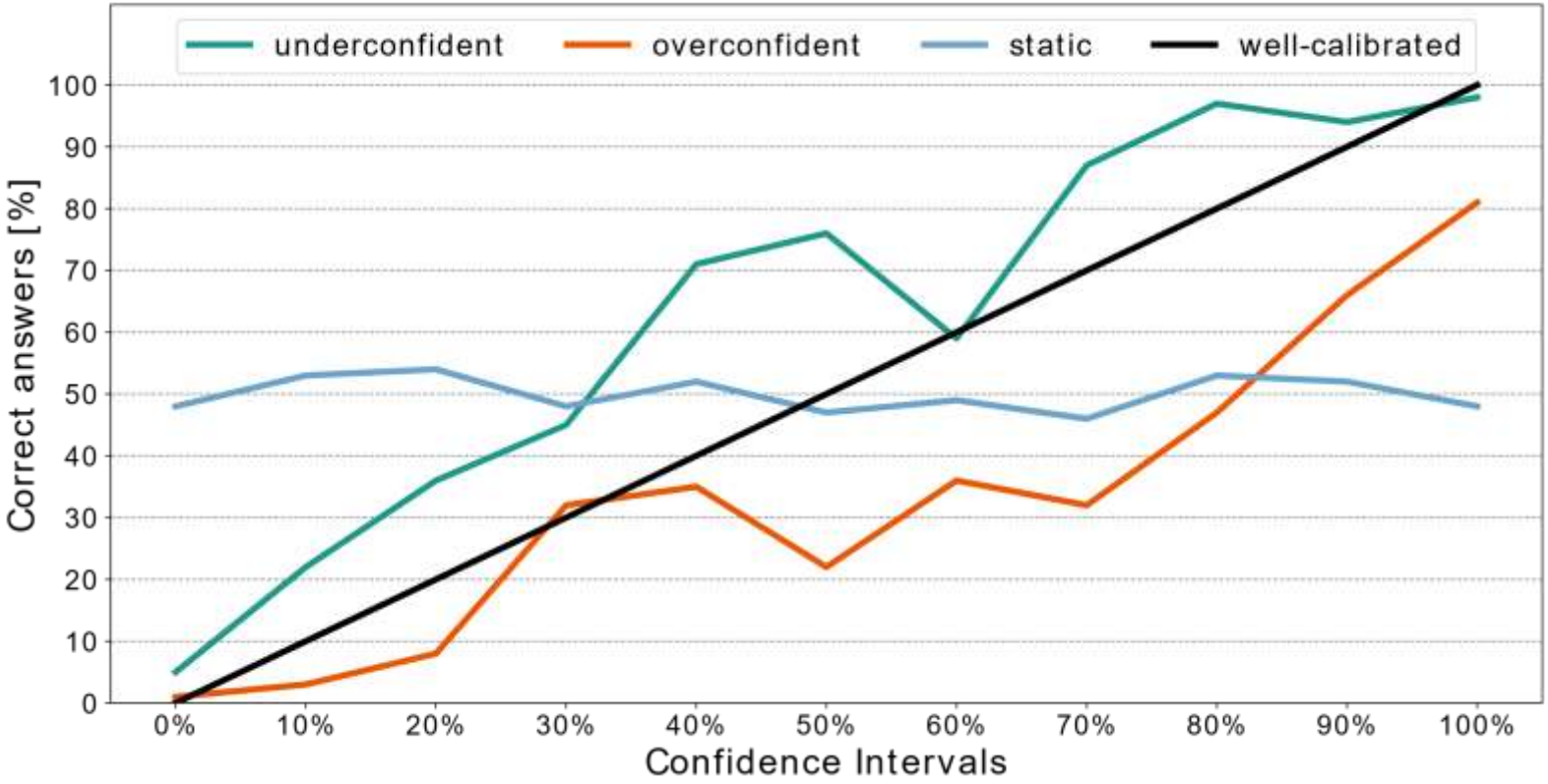
Text completion by the model

- Temperature = 0 (deterministic)
- Always selects most likely next token

ADA, BABBAGE, CURIE AND DAVINCI (InstructGPT)



CALIBRATION FOR UNCERTAINTY ESTIMATES



SELF EVALUATION OF UNCERTAINTY

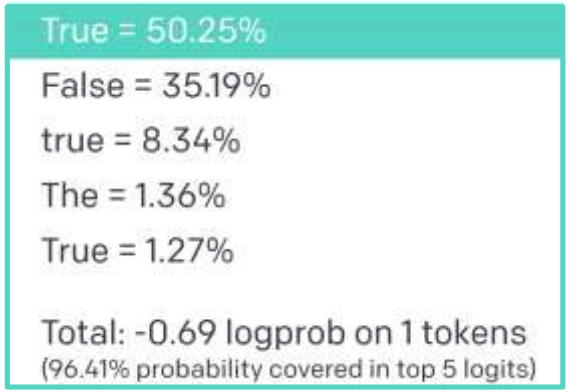
LOGITS OF TEXT COMPLETIONS

- Well-studied
- Usually good calibration

SELF-EVALUATION OF UNCERTAINTY

- Recently in focus
- Unclear calibration

EXAMPLE



"The likelihood that the proposed answer is true is 40%"

EPISTEMIC UNCERTAINTY ESTIMATION

ALEATORIC

(aka statistical) uncertainty is due to inherently random effects.

+

EPISTEMIC

(aka systematic) uncertainty is caused by a lack of knowledge.

DATASET

question	correct answer
How many episodes, in total, does the series How I Met Your Mother have over its nine seasons?	208
In which year did Wu Zhao, commonly known as Wu Zetian, the first empress of the Tang dynasty, die?	705
How tall is the Abraj Al-Bait Clock Tower in Mecca in meters?	601
In which year was Charlemagne or Charles the Great crowned King of the Lombards?	774
In which year was the Bank of Saint George, the financial institution of the Republic of Genoa, founded?	1407
How many goals did Bayern Munich score in the Bundesliga Season 2011/2012?	77

RESEARCH QUESTIONS

Is InstructGPT capable of verbally self evaluating its uncertainty about its knowledge?

Are InstructGPTs uncertainty estimates coherent and robust?

Are InstructGPTs uncertainty estimates calibrated?

How does model size change the answer to previous questions?

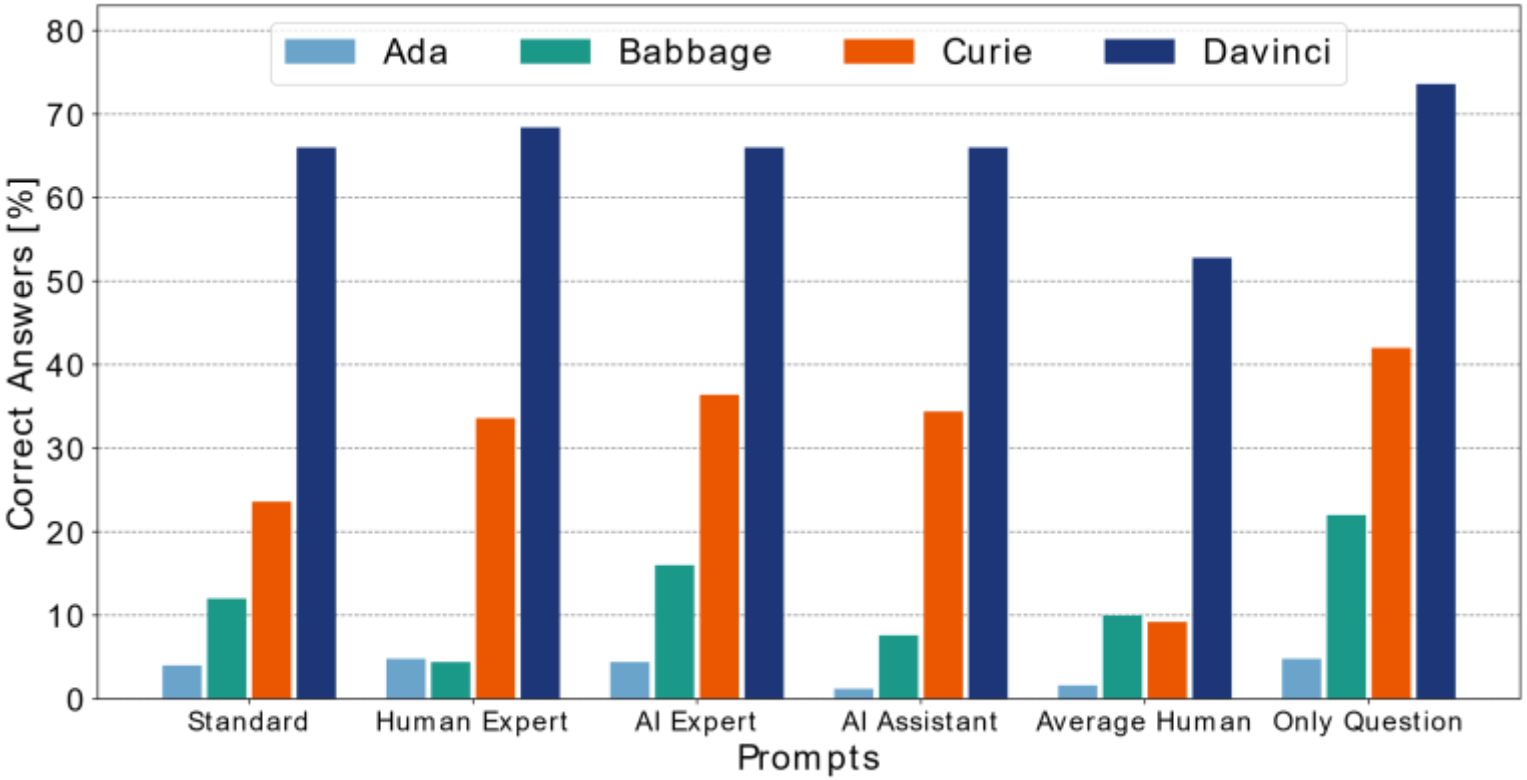
POINT ESTIMATE EXPERIMENT

The following is a conversation with the current World Quizzing Champion. The Champion is clever, knowledgeable and always knows the correct answer. The Champion is asked questions by the Host. Every question has an integer as the correct answer. The Champion answers only with the correct number.

Host: How many episodes, in total, does the series How I Met Your Mother have over its nine seasons?

Champion: 208

1. QUESTION ANSWERING IMPROVES WITH MODEL SIZE



2. ABILITY TO EXPRESS UNCERTAINTY VERBALLY

All models give

- Valid answers to different types of questions
- Percentages between 0% and 100%
- Valid Confidence Intervals

"TRUE OR FALSE (OR I DON'T KNOW)" EXPERIMENT

Question: How many episodes, in total, does the series How I Met Your Mother have over its nine seasons?
Proposed Answer: 208
Is the proposed answer:

(A) True
(B) False
(C) I don't know

The proposed answer is: **(A) True**

MULTIPLE CHOICE-EXPERIMENT

Question: How many episodes, in total, does the series How I Met Your Mother have over its nine seasons?

Proposed Answer: 208

How likely is it that the proposed answer is true?

(A) 20%

(B) 40%

(C) 60%

(D) 80%

The likelihood that the proposed answer is correct is (**D) 80%**

FREE UNCERTAINTY ESTIMATES

Question: How many episodes, in total, does the series How I Met Your Mother have over its nine seasons?
Proposed Answer: 208
How likely is it that the proposed answer is true?

The likelihood in percent that the proposed answer is true is
85%

CONFIDENCE INTERVAL EXPERIMENT

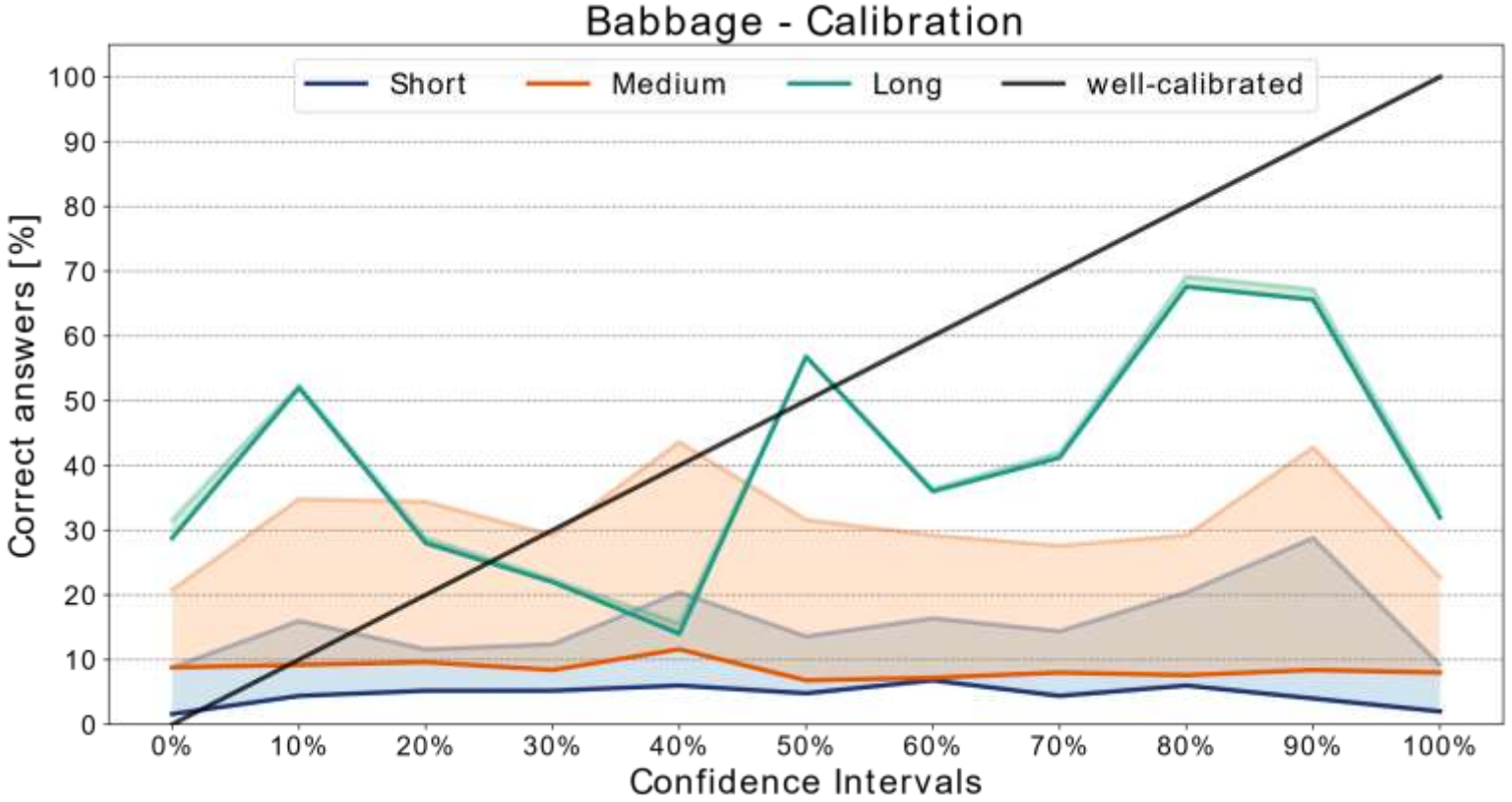
In frequentist statistics, a confidence interval is a range of estimates for an unknown parameter … out of all intervals computed at the 80% level, 80% of them should contain the parameter's true value.
Return your confidence interval in the form x:y to answer the following question.
How many episodes, in total, does the series How I Met Your Mother have over its nine seasons?

My 80% confidence interval is **176:236**

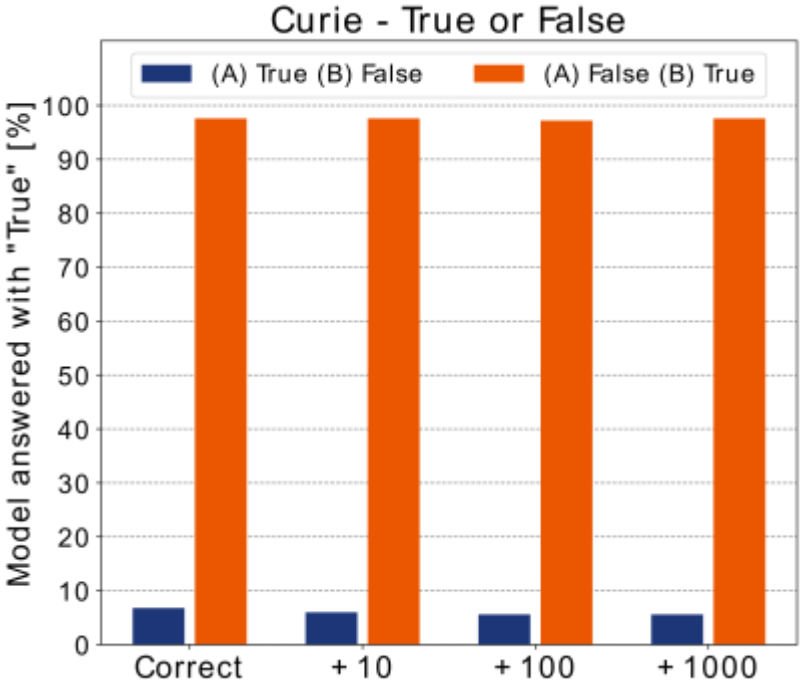
3. LARGER MODELS RETURN FEWER AMBIGUOUS ANSWERS

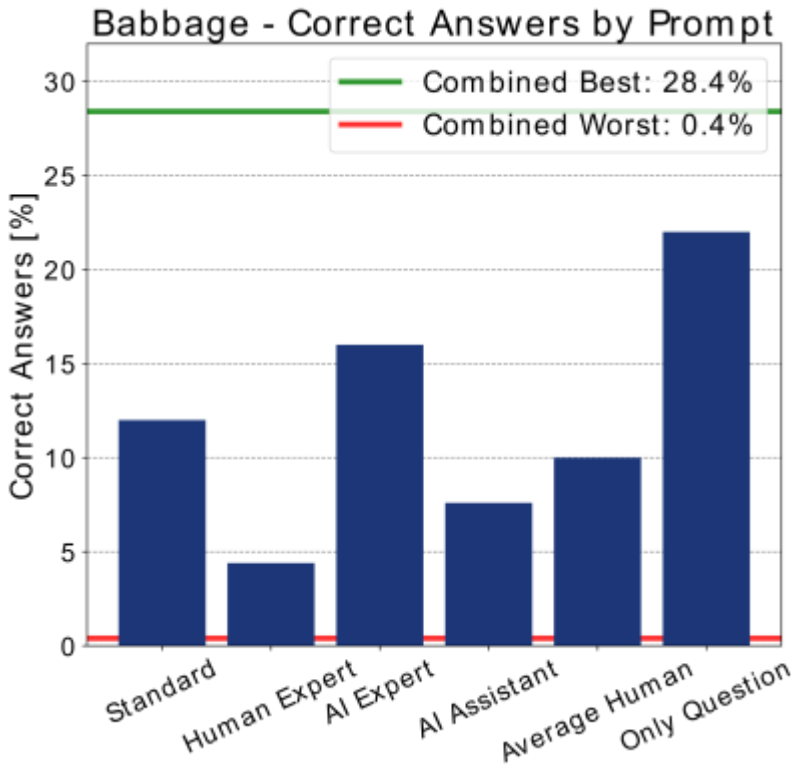
Share of ambiguous answers [%]	Davinci	Curie	Babbage	Ada
Point Estimates	1.8	13.8	18.2	19.2
True or False	0.7	0.1	1.3	1.9
True or False or I don't know	0	0.1	0.3	1.6
Multiple Choice Questions	0	0	0	7.9
Free Uncertainty Estimates	4.7	0.5	6.3	4.5
Confidence Intervals	0.7	0.5	11.7	30
Mean share of ambiguous answers	1.3	2.5	6.3	10.9

4. PROMPT DESIGN HAS SIGNIFICANT IMPACT ON RESULTS

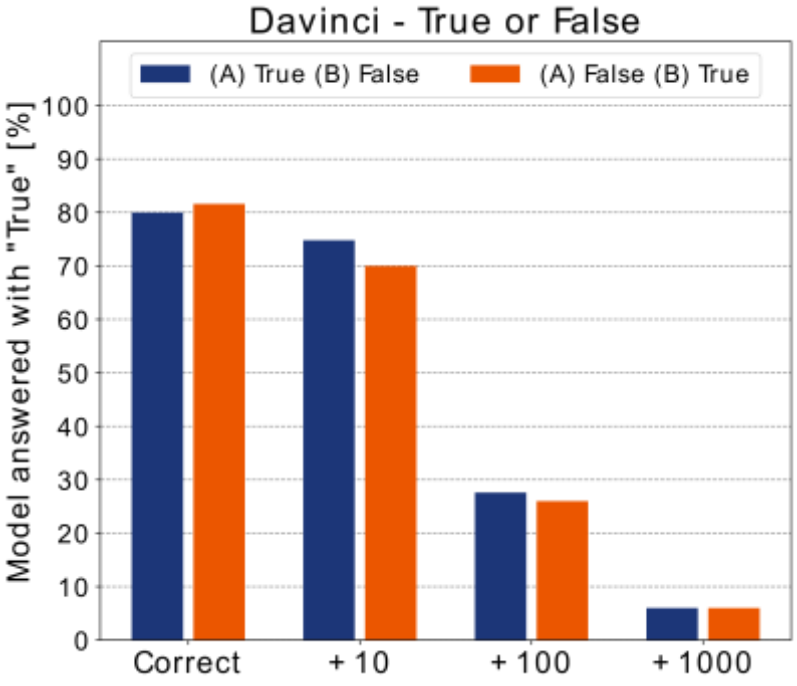


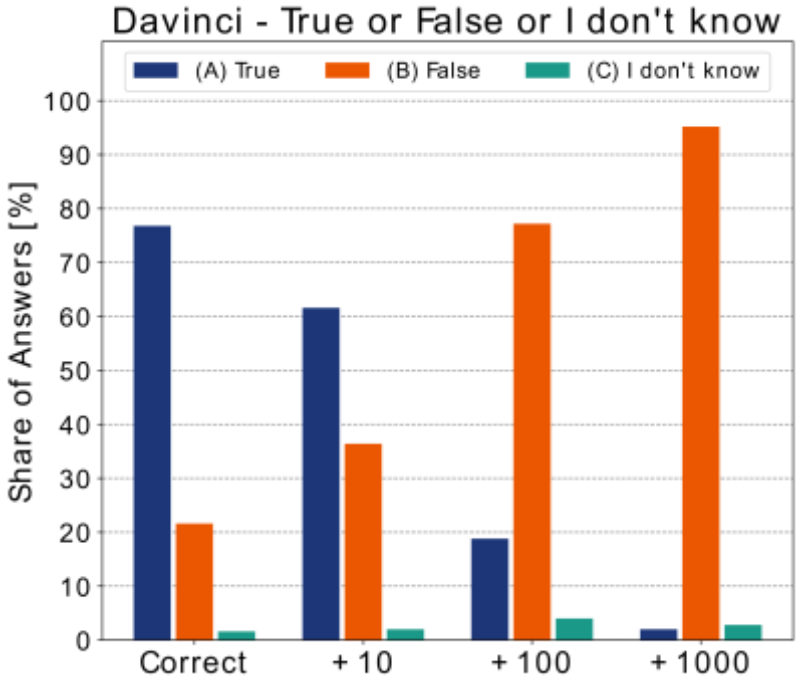
4. PROMPT DESIGN HAS SIGNIFICANT IMPACT ON RESULTS



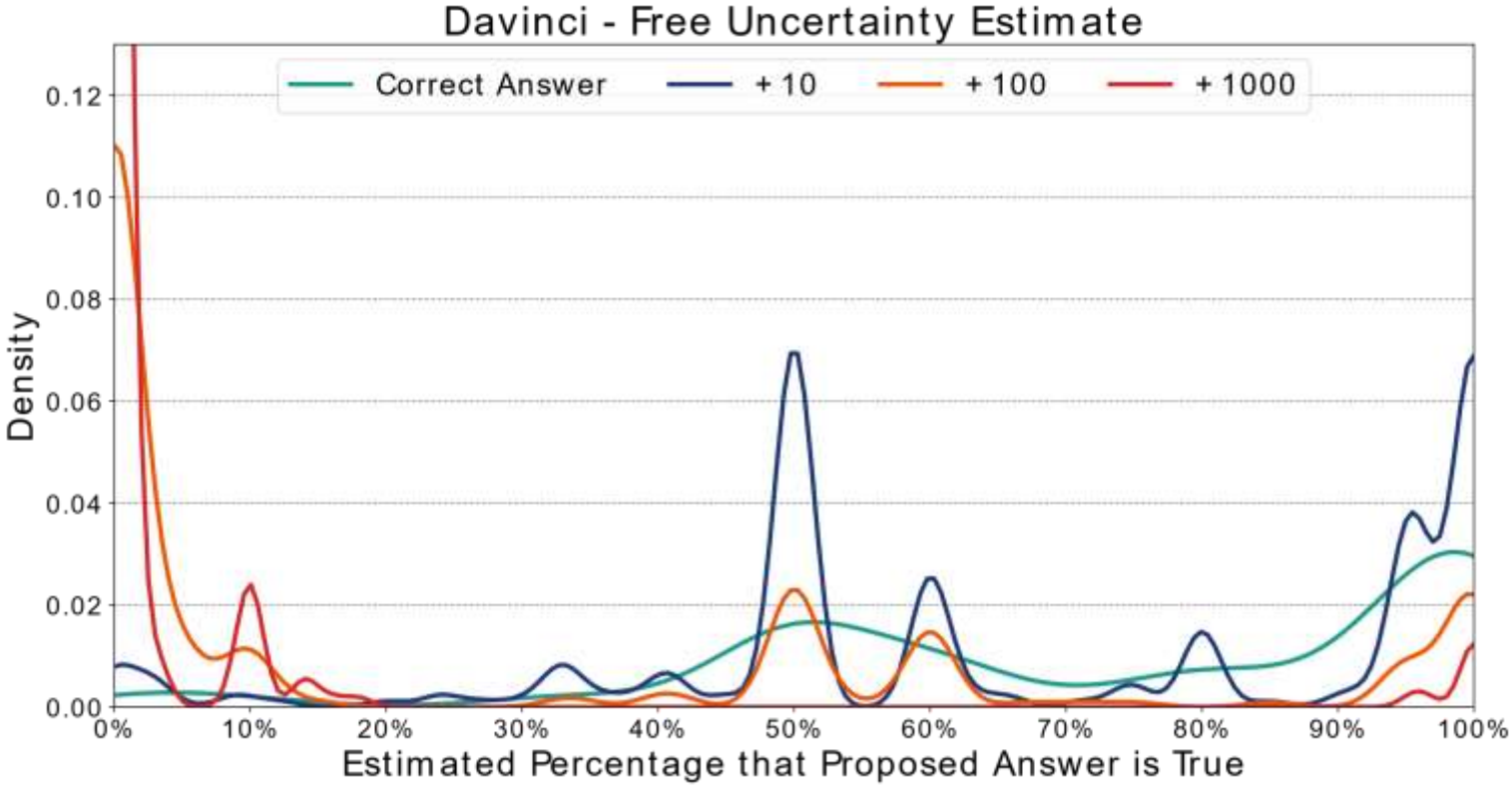


5. ONLY BIGGEST MODEL IS SOMETIMES CALIBRATED

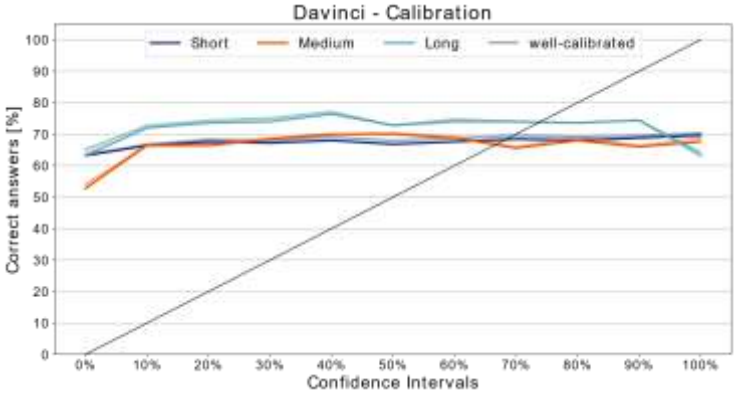


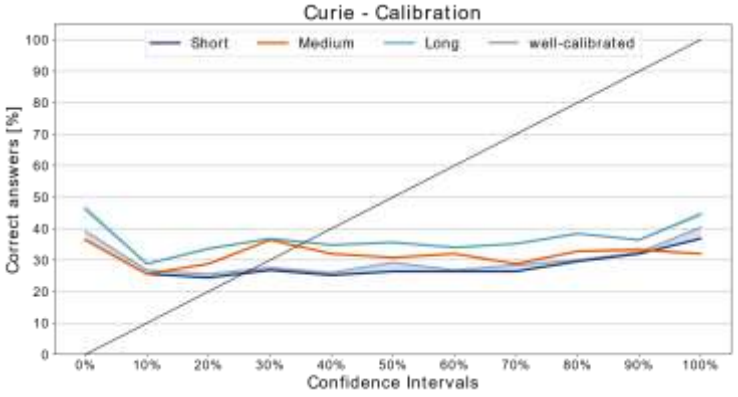


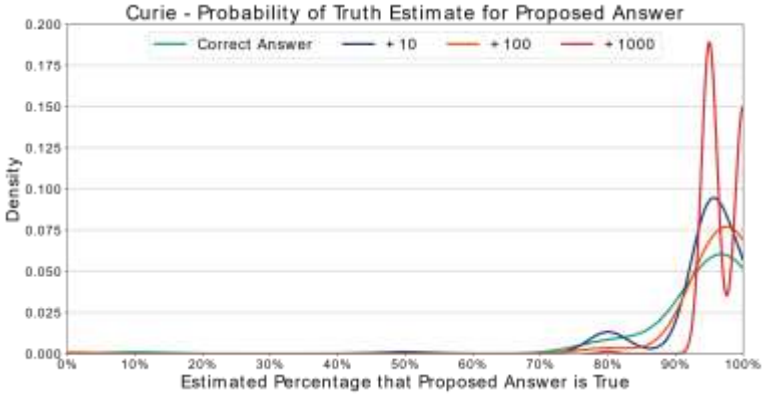
WEAK CALIBRATION FOR FREE UNCERTAINTY ESTIMATES

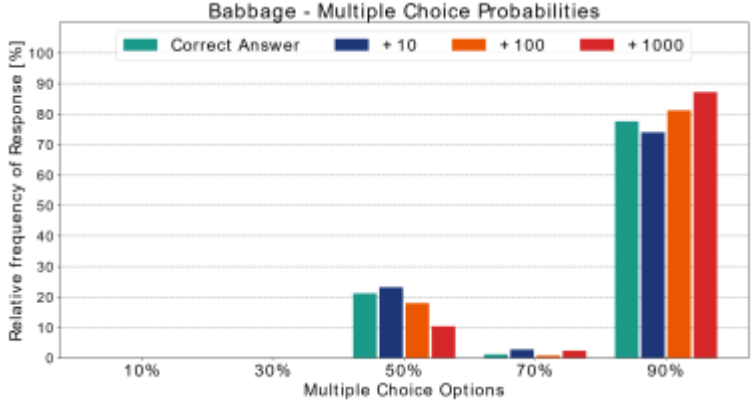


NO CALIBRATION FOR OTHER MODELS AND EXPERIMENTS





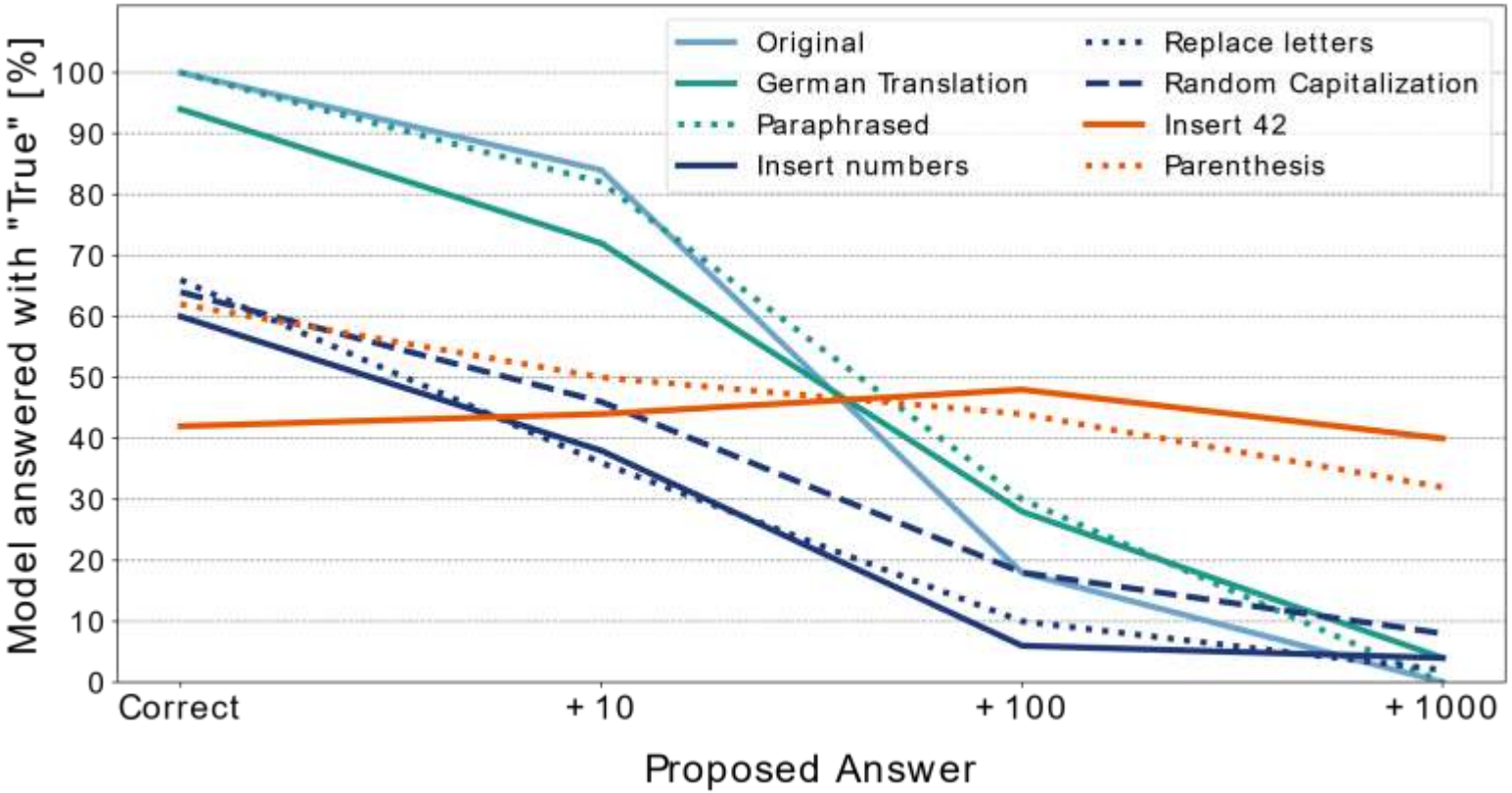




ROBUSTNESS TO ADVERSARIAL EXAMPLES

Variation	Text
Original	What year saw the publication of Pride and Prejudice?
Paraphrased	Which year was Pride and Prejudice published?
German translation	In welchem Jahr wurde "Stolz und Vorurteil" veröffentlicht?
Random capitalization	what YeAR saW THE puBlicAtION of pRidE And PREjUDiCe?
Insert five numbers	Wh1at year sa4w the public0ation of Prid9e and Preju6dice?
Replace five letters	W1at year sa4 the public0tion of Pri9e and Preju6ice?
Wrap with parenthesis	(W)(h)(a)(t) (y)(e)(a)(r) (s)(a)(w) (t)(h)(e) (p)(u)(b)(l)(i)(c)(a)(t)(i)(o)(n) (o)(f) (P)(r)(i)(d)(e) (a)(n)(d) (P)(r)(e)(j)(u)(d)(i)(c)(e)(?)
Insert 42 between letters	W42h42a42t42y42e42a42r42s42a42w42t42h42e42p42u42b42l42i42c42a42t42i42o42n42o42f42P42r42i42d42e42a42n42d42P42r42e42j42u42d42i42c42e42?

6. HIGH ROBUSTNESS TO ADVERSARIAL EXAMPLES



CONCLUSION

InstructGPT cannot (properly) self-evaluate its epistemic uncertainty

- No calibration

Overall impressive performance

- Language comprehension
- Question answering ability
- Ability to follow instructions
- Robustness to adversarial examples

OUTLOOK

Using few-shot learning

Fine Tuning to specific experiments

“Let’s Think step by step”

Bayesian Updating

Increase Parameter size