Long-Form Factuality in Large Language Models

Jerry Wei*, Chengrun Yang*, Xinying Song*, Yifeng Lu*, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, Quoc V. Le

[*: Lead Contribution]

Google DeepMind

NeurIPS 2024

Long-form generation in open-domains remains challenging…

Q: Who is Quoc V. Le?

A: Quoc V. Le is a computer scientist and software engineer who is currently a research scientist at Google Brain. He is best known for his work on deep learning, particularly in the areas of natural language processing and computer vision. Le has made significant contributions to the development of new deep learning architectures and algorithms, and his work has been widely cited in the academic literature. He is also the author of several popular open-source software libraries for deep learning.

Le received his B.S. in computer science from the University of Waterloo in 2009 and his Ph.D. in computer science from Stanford University in 2014. After completing his Ph.D., he joined Google Brain, where he has been working on a variety of deep learning projects. Le is also an adjunct professor at the University of Montreal, where he teaches courses on deep learning.

Le is a highly respected researcher in the field of deep learning, and his work has been recognized with several awards, including the 2016 Young Investigator Award from the International Machine Learning Society and the 2017 Google Faculty Research Award. He is also a co-author of the book "Deep Learning for Natural Language Processing".

…

How to quantitatively evaluate?

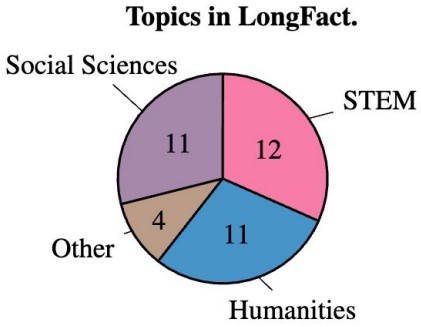
Prior works

- **benchmarks**
  - single-factoid (TruthfulQA, FreshQA, HaluEval, HalluQA, ...)
  - long-form biography (FActScore)
- **evaluation methods**
  - comparing with ground-truth
  - with crowd-sourced human raters
- **quantitative metrics**

Our work: Long-form factuality **benchmark + evaluation + metric**

- **benchmark:** LongFact
- **evaluation method:** Search-Augmented Factuality Evaluator (SAFE)
- **quantitative metric:** F1@K

LongFact: a **multi-topic** benchmark for **long-form** factuality



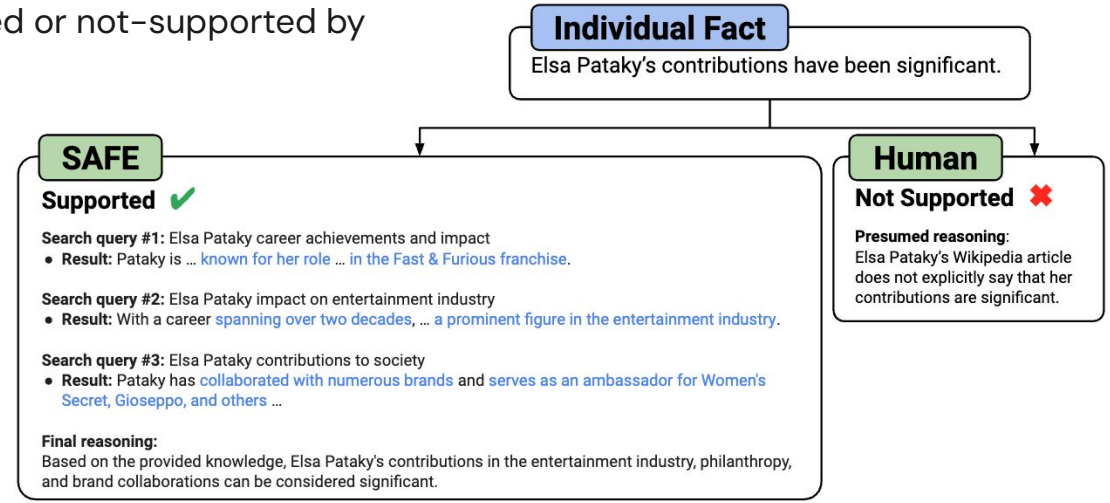
Factuality Benchmark	# Topics	Response type
TruthfulQA (Lin et al., 2022)	38	Short-answer
HaluEval (Li et al., 2023)	Many	Short-answer
FreshQA (Vu et al., 2023)	Many	Short-answer
HalluQA (Cheng et al., 2023b)	14	Short-answer
FELM (Chen et al., 2023)	5	Mixed
FActScore (Min et al., 2023)	1	Long-form
LongFact (ours)	38	Long-form

- include 1,140 **objects** and 1,140 **concepts** across 38 topics
- "**specific and niche**" to elicit factual errors
- example questions:
  - (economics) Can you provide an overview of the International Monetary Fund?
  - (international law) What is known about the Antarctic Treaty?
  - (biology) What can you tell me about the Ghost Orchid?

SAFE: LLM agents as factuality autoraters

Steps of Search-Augmented Factuality Evaluator (SAFE):

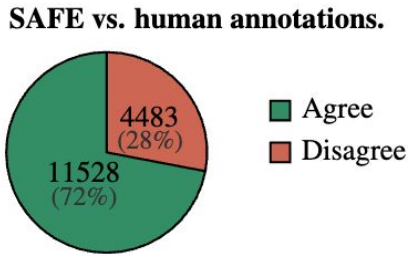
1. **split** a response into **individual facts**
2. **revise** facts to be **self-contained**
3. determine **relevance** of individual facts
4. **rating** individual facts as supported or not-supported by multi-step search



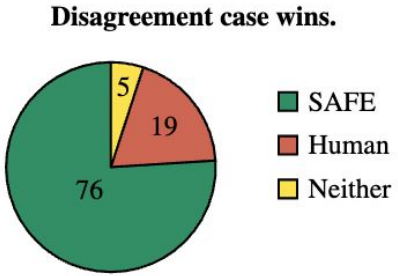
LLM agents can be better factuality annotators than crowd-sourced humans

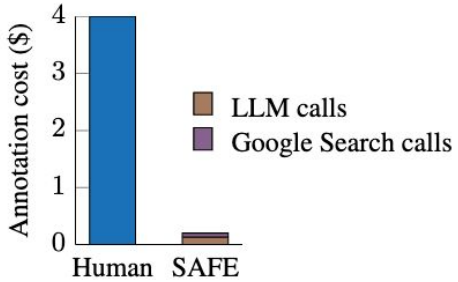
On FActScore biography dataset:

- On 16,011 individual facts, SAFE annotations **agree with 72.0%** of human annotations.



- On a randomly subsampled subset of 100 differently rated facts, SAFE compared to crowd-sourced human:
  - **wins 4x more often**
  - **is 20x cheaper**



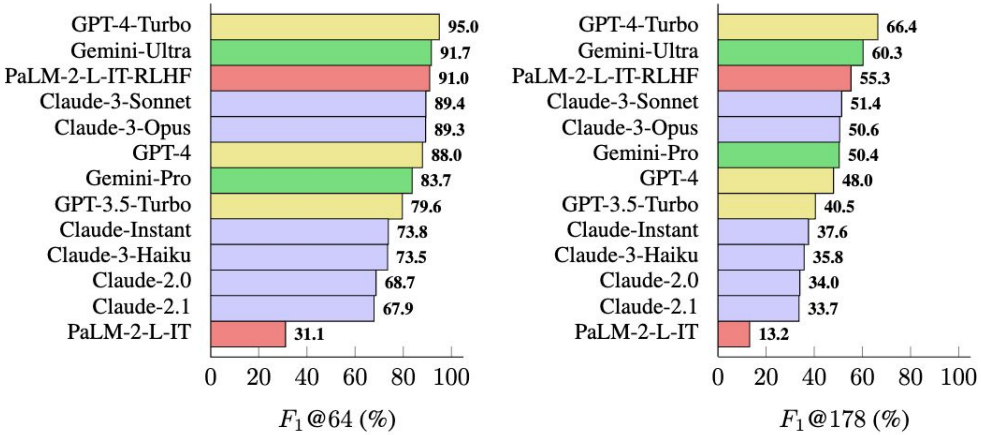


Min et al., FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation, EMNLP 2023.

F1@K metric: Extending F1 with recall from human-preferred length

With the numbers of supported facts S and not-supported facts N,

- precision: S / (S + N)
- recall: min{S / K, 1}, where K is the number of supported facts we desire
- F1@K = F1 (precision, recall) if S > 0, otherwise 0



- larger models are more factual
- F1@K ranking is relatively stable at sufficiently large K values

Summary

Long-form factuality can be automatically evaluated with LongFact, SAFE, and F1@K.

Please see our paper for:

- prompts and other implementation details
- ablation studies
- discussions on limitations

NeurIPS page: https://nips.cc/virtual/2024/poster/96675
Paper: https://arxiv.org/abs/2403.18802
Code: https://github.com/google-deepmind/long-form-factuality