Ethical AI Through Prompt Engineering: Mitigating Bias and Reducing Hallucinations in Conversational Systems

By Parvin Gasimzade

Sr. Software Development Engineer at Amazon

Agenda

- Introduction
- Ethical AI
- Bias in AI systems
- Hallucinations in AI systems
- Methods to mitigate bias & reduce hallucinations
- Key Takeaways

Large Language Models

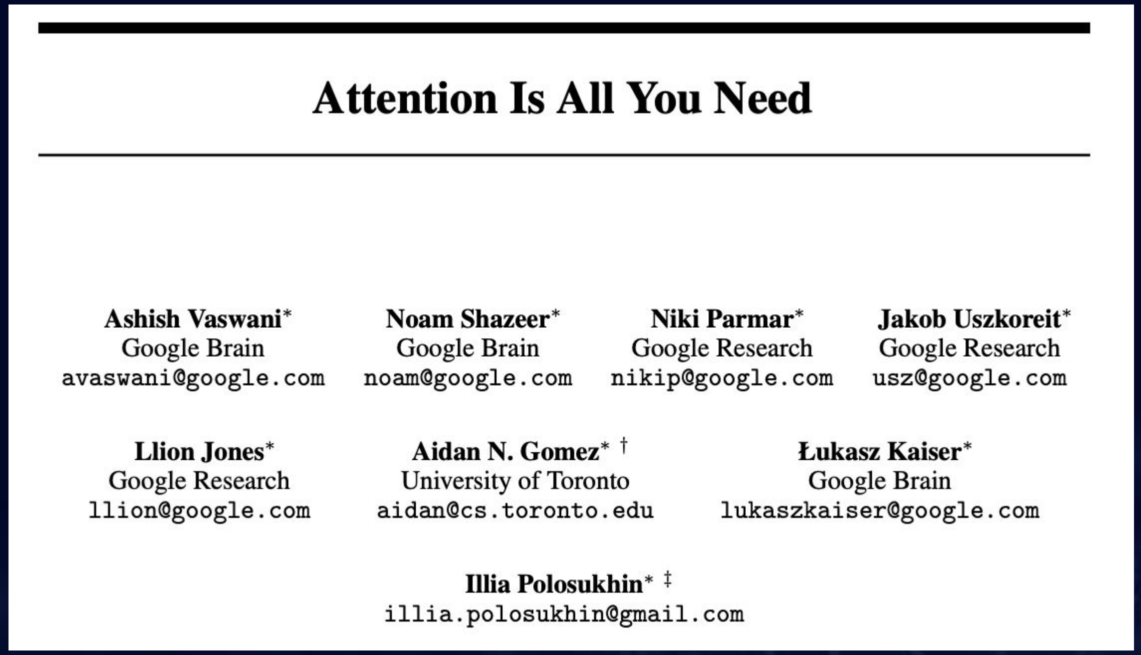
“An LLM, or Large Language Model, is a type of artificial intelligence model designed to understand and generate human-like text. These models are trained on vast amounts of text data, enabling them to predict and produce responses based on the input they receive.”

ChatGPT, 11/24

Transformer Architecture

A transformer is a type of artificial intelligence model that learns to understand and generate human-like text by analyzing patterns in large amounts of text data.

- 2017: Transformer model was created by researchers at Google.
- 2018: OpenAI released first version of the GPT (Generative Pre-trained Transformer)
- 2020: OpenAI released GPT-3, which had 175 billion parameters, one of the largest language model ever created.



Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

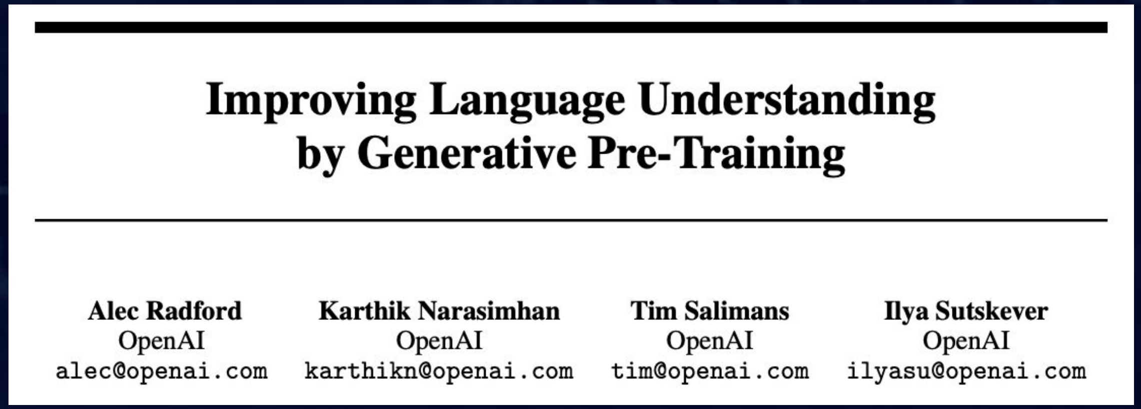
Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com



Improving Language Understanding by Generative Pre-Training

Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthikn@openai.com

Tim Salimans
OpenAI
tim@openai.com

Ilya Sutskever
OpenAI
ilyasu@openai.com

Ethical AI

Ethical AI refers to the development of AI systems that aligns with ethical principles and values.

- **Fairness:** Ensuring that AI systems do not discriminate against individuals or groups based on race, gender, age, or other characteristics.
- **Transparency:** Making the workings of AI systems understandable to users and stakeholders. Clear explanations of how decisions are made, what data is used, and rationale behind outcomes.
- **Accountability:** Establishing responsibility for the outcomes of AI systems.
- **Privacy:** Safeguarding personal data and ensuring that individuals' privacy rights are respected.
- **Safety and Security:** Ensuring that AI systems are safe to use and secure from malicious attacks or misuse.
- **Inclusivity:** Engaging diverse stakeholders in the development of AI technologies to ensure multiple perspectives are considered.

Two key issues are bias and hallucination in AI outputs.

Understanding Bias in AI

Bias in AI systems refers to systematic errors that result in unfair outcomes for certain groups or individuals. This can arise from multiple sources, including:

- Training data - imbalanced datasets, historical bias
- Model/Algorithm design - model design choices, feature selection
- Labelling data - subjective labelling
- User interactions - biased user inputs

Ethical Implications:

- Fairness: Unfair treatment of certain groups, which goes against the idea of treating everyone equally and fairly.
- Trust: People lose trust in AI systems and the companies behind it. This can make them hesitant to use AI and doubt its usefulness.

Understanding Hallucination in AI

Hallucination in AI refers to instances where the model generates incorrect or fabricated information, impacting reliability. This can happen in following situations:

- Training data - insufficient and not relevant data sets
- Data quality - poor-quality training data. Noisy, ambiguous or incorrectly labelled data.
- Lack of context - not have sufficient context, causing them to fill the gaps with fabricated information
- Language complexity - idiomatic expressions, sarcasm, cultural references

Ethical Implications:

- Accuracy: it can lead to the spread of misinformation
- Responsibility: accountability and responsibility for the resulting impacts

Methods to mitigate bias

Effective mitigation methods can reduce bias, promoting fairer AI interactions.

- **Diverse Datasets**: Include a variety of demographic backgrounds to avoid skewed outputs.
- **Fine-tuning and Debiasing**: Retrain models using data specifically curated to reduce bias
- **Bias Audits and Assessments:** Conduct regular audits of AI systems to identify and evaluate potential biases.
- **Algorithm Transparency:** Improve transparency in how algorithms function and make decisions.
- **Regular Model updates:** Continuously update AI models with new data and feedback to adapt to changing societal norms

Methods to reduce hallucinations

Several techniques can help minimize inaccuracies and increase AI reliability.

- **High-quality Training Data:** Use large, diverse, and high-quality datasets that accurately represent the desired output.
- **Contextual Training:** Train models with additional context for specific tasks or domains
- **Fine-tuning:** Fine-tune pre-trained models on domain-specific datasets to improve their relevance and accuracy in generating content related to that domain.
- **External Knowledge Sources**: Integrate external knowledge bases or databases to verify facts and provide additional context.
- **Continuous Learning:** Adopt a continuous learning approach, where models are regularly updated and retrained based on new data and user interactions

Prompt engineering for Mitigating Bias and reducing hallucination

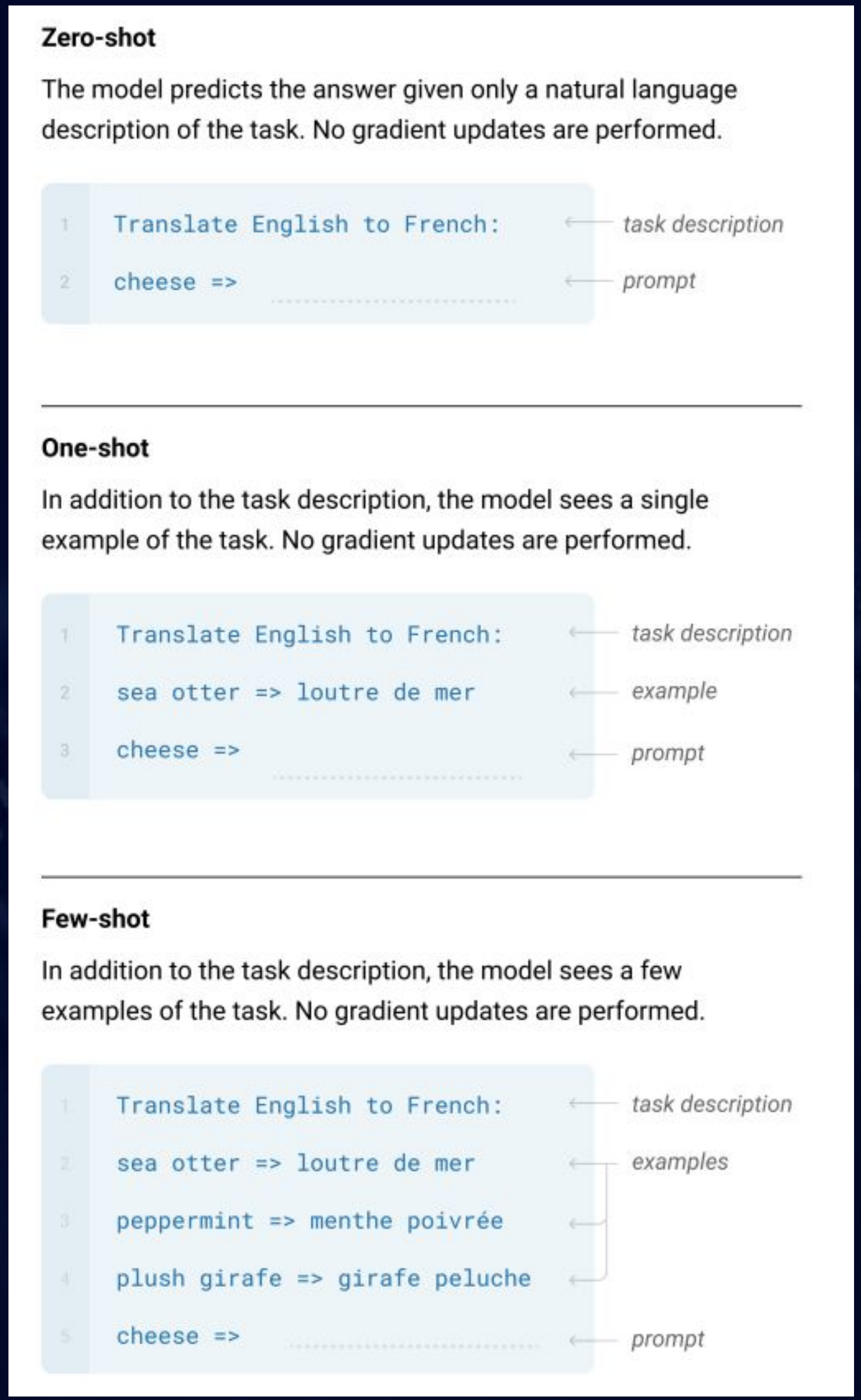
- **Zero-Shot and Few-Shot Learning**
- **Chain of Thought (CoT)**
- **Modular Reasoning, Knowledge and Language (MRKL)**

Zero-Shot and Few-Shot Learning

Zero-shot learning allows a model to perform tasks without any additional examples, relying on its training data. While few-shot learning enables a model to learn a task from provided examples.

- Zero-shot: no additional example
- One-shot: single example
- Few-shot: a few examples

https://arxiv.org/pdf/2005.14165 @2020



Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:
cheese =>
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

Few-shot

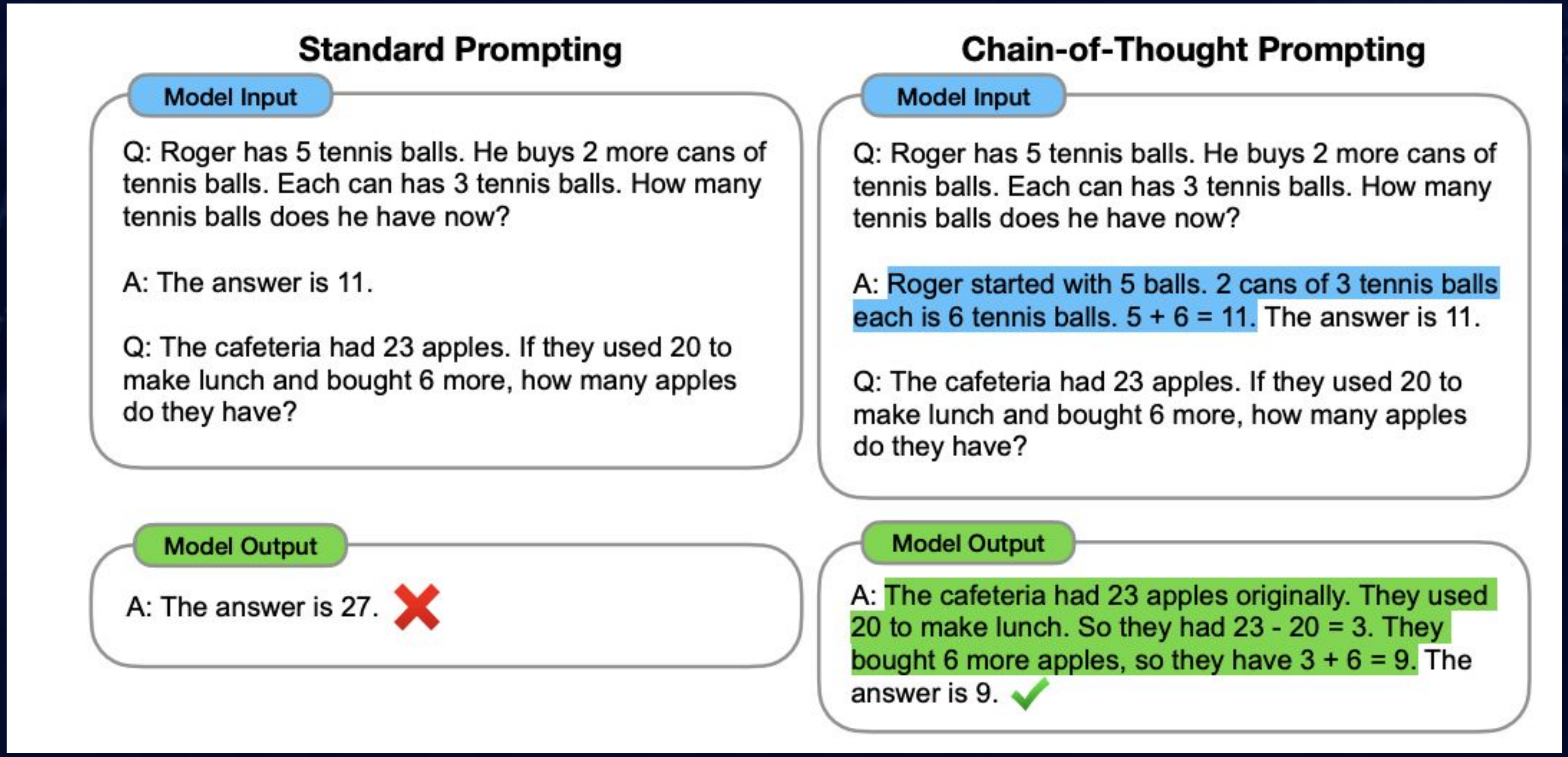
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

Chain of Thought (CoT)

Chain of Thought (CoT): Encourages models to think step-by-step through a problem, helping them reason more effectively and reduce hallucinations in complex tasks.

https://arxiv.org/pdf/2201.11903, @2023



Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

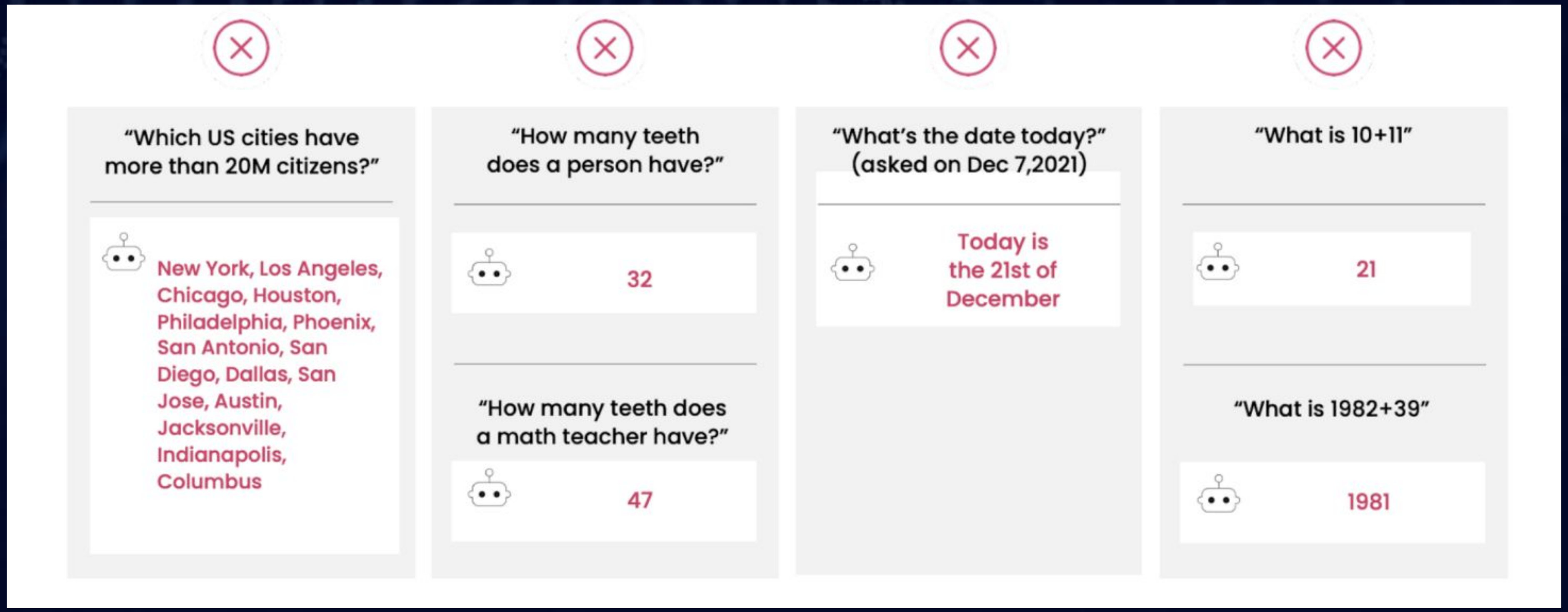
Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Modular Reasoning, Knowledge and Language (MRKL)

MRKL (Modular Reasoning with Knowledge and Language) is a framework that enhances the capabilities of language models by integrating structured reasoning processes with external knowledge sources. This approach allows models to not only generate text based on learned patterns but also to retrieve relevant information and reason about it, leading to more accurate and context-aware outputs.

https://arxiv.org/pdf/2205.00445 @2022



"Which US cities have more than 20M citizens?"

New York, Los Angeles, Chicago, Houston, Philadelphia, Phoenix, San Antonio, San Diego, Dallas, San Jose, Austin, Jacksonville, Indianapolis, Columbus

"How many teeth does a person have?"

32

"How many teeth does a math teacher have?"

47

"What's the date today?" (asked on Dec 7,2021)

Today is the 21st of December

"What is 10+11"

21

"What is 1982+39"

1981

Key Takeaways

- **Importance of Ethical AI:** Ethical AI is crucial for building trust, ensuring fairness, and promoting accountability in AI systems.
- **Role of Prompt Engineering:** Effective prompt engineering techniques can significantly mitigate bias and reduce hallucinations, leading to more reliable and equitable conversational systems.
- **Continuous Improvement:** Ongoing evaluation and adaptation of AI models, combined with diverse stakeholder involvement, are essential for addressing ethical challenges in AI.
- **Commitment to Responsibility:** Organizations must commit to ethical practices in AI development, ensuring that technology serves the greater good and aligns with societal values.

References

- https://arxiv.org/pdf/1706.03762
- https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- https://arxiv.org/pdf/2005.14165
- https://arxiv.org/pdf/2201.11903
- https://arxiv.org/pdf/2205.00445

THANK YOU!

Linkedin