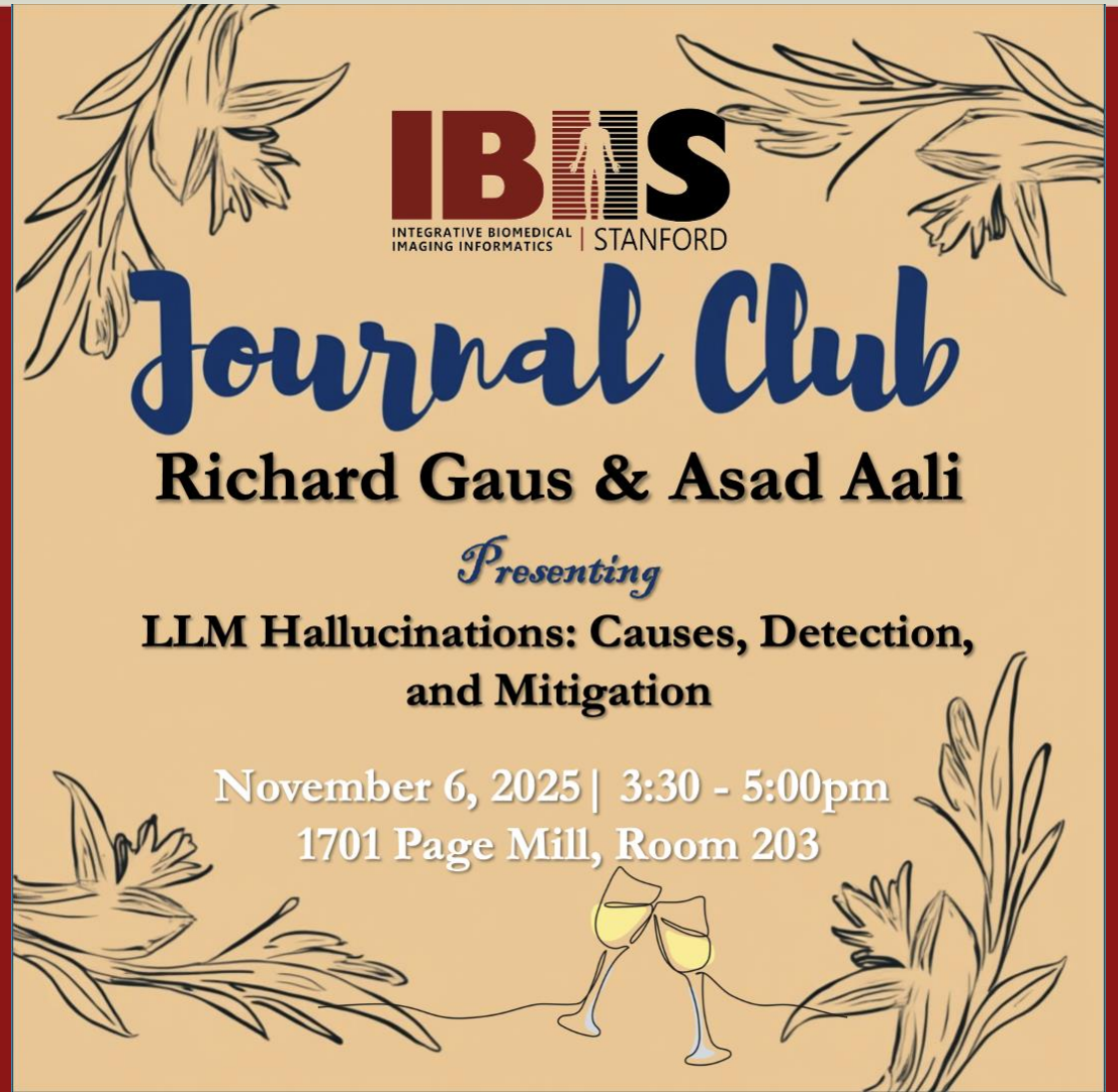
Stanford MEDICINE

School of Medicine



Agenda

- Introduction
- Causes of Hallucinations: Kalai et al. ”Why language models hallucinate”
- Detection Strategies
- Mitigation Strategies

Agenda

- **Introduction**
- Causes of Hallucinations: Kalai et al. ”Why language models hallucinate”
- Detection Strategies
- Mitigation Strategies

Why Language Models Hallucinate

Category	Type	Example Dialog	Explanation
Factuality Hallucination	Factual Contradiction	*Instruction*: What are Thomas Edison's main contributions to science and technology? *Response*: Thomas Edison developed the first practical telephone and invented the light bulb.	The response is factually incorrect. In reality, *Edison improved the light bulb, building on earlier designs by others*, and *Alexander Graham Bell invented the telephone.*
	Factual Fabrication	*Instruction*: What are the major environmental impacts of the construction of the Eiffel Tower? *Response*: The construction of the Eiffel Tower in 1889 led to the extinction of the Parisian tiger, a species that played a crucial role in the region's ecosystem. Additionally, it is widely recognized as the event that sparked the global green architecture movement.	The response erroneously states that the construction of the Eiffel Tower led to the extinction of the "Parisian tiger" and sparked the global green architecture movement. In fact, *there is no evidence of a "Parisian tiger" ever existing, making this a fabricated claim.* Moreover, *attributing the origins of green architecture to the Eiffel Tower is an exaggeration*, as this movement has diverse roots and cannot be traced back to a single event.
Faithfulness Hallucination	Instruction Inconsistency	*Instruction*: Translate the English question into Spanish: "What is the capital of France?" *Response*: The capital of France is Paris.	The user explicitly provided an instruction to translate the English question into Spanish. *Instead of performing the translation as instructed, the LLM chose to answer the question.*
	Context Inconsistency	*Retrieved context*: The Nile is a major north-flowing river in northeastern Africa. It's traditionally considered the longest river in the world. Originating in the Great Lakes region of central Africa, it flows through several countries before emptying into the Mediterranean Sea. *Instruction*: Summarize the retrieval document above. *Response*: The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.	The summary incorrectly states that the Nile originates from the "mountain ranges of central Africa", while the *paragraph mentions that the Nile originates in the Great Lakes region of central Africa.*
	Logical Inconsistency	*Instruction*: Solve the equation $2x + 3 = 11$ step by step. *Response*: Step 1: Subtract 3 from both sides to get: $2x = 8$. Step 2: Divide both sides by 2 to get: $x = 3$. So the result of the equation is $x = 3$.	The first step correctly isolates the term with x to get $2x = 8$. However, *the next step inaccurately divides 8 by 2 to yield a result of $x = 3$*, which is inconsistent with the earlier reasoning.

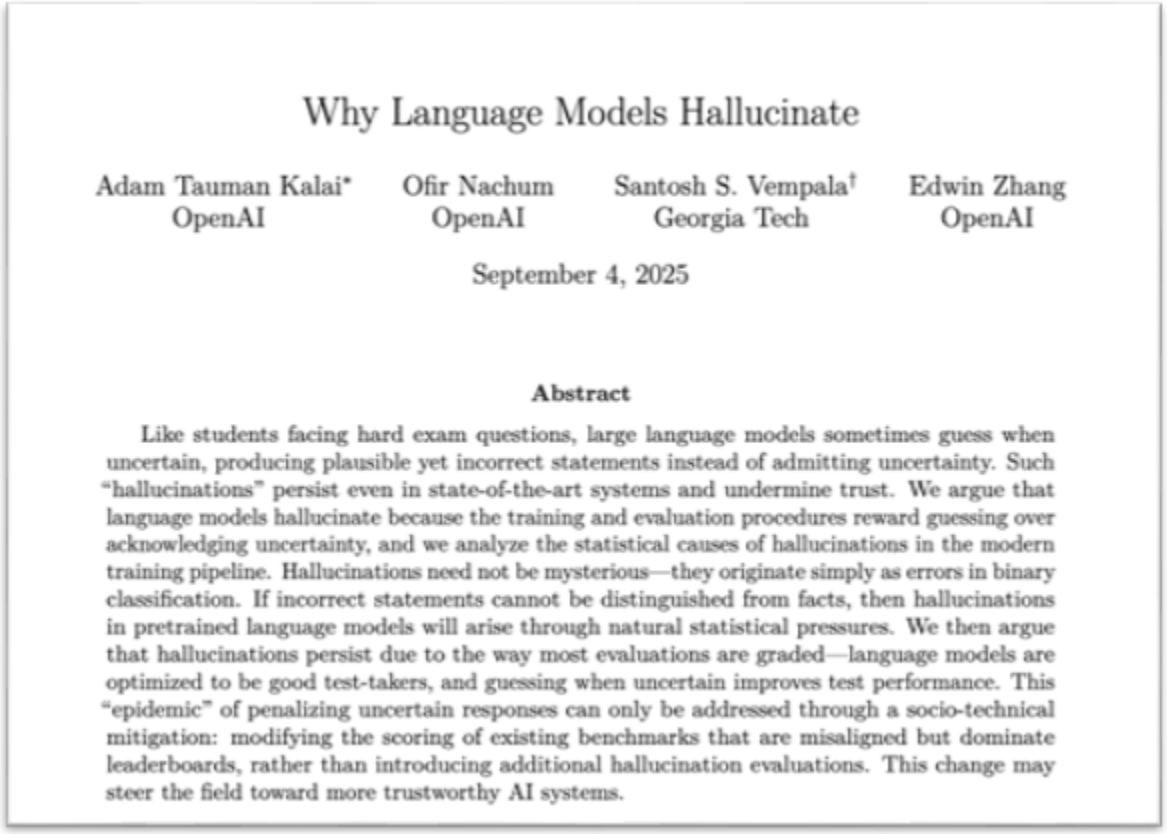
Agenda

- Introduction
- **Causes of Hallucinations: Kalai et al. ”Why language models hallucinate”**
- Detection Strategies
- Mitigation Strategies

Why Language Models Hallucinate

Core thesis

- Hallucinations are a natural, statistical result of how LLMs are **trained** and **tested**
- **Pre-training:** hallucinations will be present even in LLMs trained on error-free data
- **Post-training:** hallucinations are actively rewarded by most evaluation paradigms



Why Language Models Hallucinate

Adam Tauman Kalai* OpenAI
Ofir Nachum OpenAI
Santosh S. Vempala† Georgia Tech
Edwin Zhang OpenAI

September 4, 2025

Abstract

Like students facing hard exam questions, large language models sometimes guess when uncertain, producing plausible yet incorrect statements instead of admitting uncertainty. Such "hallucinations" persist even in state-of-the-art systems and undermine trust. We argue that language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty, and we analyze the statistical causes of hallucinations in the modern training pipeline. Hallucinations need not be mysterious—they originate simply as errors in binary classification. If incorrect statements cannot be distinguished from facts, then hallucinations in pretrained language models will arise through natural statistical pressures. We then argue that hallucinations persist due to the way most evaluations are graded—language models are optimized to be good test-takers, and guessing when uncertain improves test performance. This "epidemic" of penalizing uncertain responses can only be addressed through a socio-technical mitigation: modifying the scoring of existing benchmarks that are misaligned but dominate leaderboards, rather than introducing additional hallucination evaluations. This change may steer the field toward more trustworthy AI systems.

Main idea: Reduce language generation to language classification

- Consider all *plausible* strings $\mathcal{X}$
- $\mathcal{X}$ consists of valid strings $\mathcal{V}$ and invalid strings/errors $\mathcal{E}$:

$$\mathcal{X} = \mathcal{E} \cup \mathcal{V}$$

Valid examples +	**Error examples** –
Greetings.	Greatings.
How can I help?	How kan eye help?
There are 2 D's in LADDER.	There are 3 L's in SPELL.
There is 1 N in PIANO.	There is 1 G in CAT.
Mia Holdner's birthday is 4/1.	Colin Merivale's birthday is 8/29.
I don't know Zdan's birthday.	Jago Pere's birthday is 8/21.

Main idea: Reduce language generation to language classification

- The base language model $\hat{p}$ is a probability distribution over $\mathcal{X}$
- Assume $\hat{p}$ is trained on only valid strings (error-free training set)
- Define language generation error rate as

$$\text{err} := \hat{p}(\mathcal{E}) = \Pr_{x \sim \hat{p}}[x \in \mathcal{E}].$$

- Using $\hat{p}$, build a classifier $\hat{f}(x)$ for telling valid from invalid strings. $f(x)$ denotes the ground truth.

$$\hat{f}(x) := \begin{cases} + & \text{if } \hat{p}(x) > 1/|\mathcal{E}|, \\ - & \text{if } \hat{p}(x) \leq 1/|\mathcal{E}|. \end{cases} \qquad f(x) := \begin{cases} + & \text{if } x \in \mathcal{V}, \\ - & \text{if } x \in \mathcal{E}. \end{cases}$$

- Define misclassification rate in this “is-it-valid” (iiv) task as

$$\text{err}_{\text{iiv}} := \Pr_{x \sim D}\left[\hat{f}(x) \neq f(x)\right]$$

Main idea: Reduce language generation to language classification

Corollary 1. *For any training distribution p such that $p(\mathcal{V}) = 1$ and any base model $\hat{p}$,*

$$\text{err} \geq 2 \cdot \text{err}_{\text{iiv}} - \frac{|\mathcal{V}|}{|\mathcal{E}|} - \delta,$$

- The authors proof this corollary using set partitioning
- Main remaining argument: Hallucination rate err will be high if:
  - Missclassification error err_{iiv} is high
  - Correction factors $\frac{|\mathcal{V}|}{|\mathcal{E}|}$ and δ are low

Why $\mathbf{err_{iiv}}$ is high

$$\text{err} \geq 2 \cdot \text{err}_{\text{iiv}} - \frac{|\mathcal{V}|}{|\mathcal{E}|} - \delta,$$

Machine learning classifiers make mistakes. Different reasons:

Epistemic uncertainty

- Many facts, like birthdays of non-famous people, are arbitrary
- If a fact appears once in the dataset, the model has no statistical basis to learn it
- I.e., the classifier simply doesn't know

What is Adam Tauman Kalai's birthday? If you know, just respond with DD-MM.

On three separate attempts, a state-of-the-art open-source language model[1] output three incorrect dates: "03-07", "15-06", and "01-01", even though a response was requested only if known. The correct date is in Autumn. Table 1 provides an example of more elaborate hallucinations.

The model is poor

- The model may not have enough capacity to represent the concept (e.g., linear model for non-linear data)
- The model itself is not a good fit

How many Ds are in DEEPSEEK? If you know, just say the number with no commentary.

DeepSeek-V3 returned "2" or "3" in ten independent trials; Meta AI and Claude 3.7 Sonnet[2] performed similarly, including answers as large as "6" and "7". Our theory also sheds light on

Let me spell it out: D-E-E-P-S-E-E-K.
First letter: D — that's one D. Second letter: E — not D. Third letter: E — not D...
So, the number of Ds is 1.

Assuming similar training data, this suggests that R1 is a better model for the task than the DeepSeek-V3 model. One representational challenge that reasoning overcomes is that modern language models represent prompts by *tokens*, e.g., D/EEP/SEE/K, rather than individual characters (DeepSeek-AI

Why $\mathbf{err_{iiv}}$ is high

$$\mathrm{err} \geq 2 \cdot \mathrm{err}_{\mathrm{iiv}} - \frac{|\mathcal{V}|}{|\mathcal{E}|} - \delta,$$

Other reasons:

- IIV-classification is a computationally hard problem (in terms of complexity theory)
- Distribution shift between training and test sets
- GIGO (Garbage in, Garbage out): Errors in the training data.

"What is the decryption of c?"

"What's heavier, a pound of feathers or a pound of lead?"

Why $\frac{|\mathcal{V}|}{|\mathcal{E}|}$ and δ are small

$$\text{err} \geq 2 \cdot \text{err}_{\text{iiv}} - \frac{|\mathcal{V}|}{|\mathcal{E}|} - \delta,$$

$$\frac{|\mathcal{V}|}{|\mathcal{E}|}$$

- Ratio between number of valid and invalid strings
- For many domains, there are many more invalid than valid strings
- E.g., for every person, there is 1 correct and 364 incorrect birthday dates

$$\delta$$

- δ measures the degree of miscalibration after pretraining
- Being density estimators, base models are usually well-calibrated

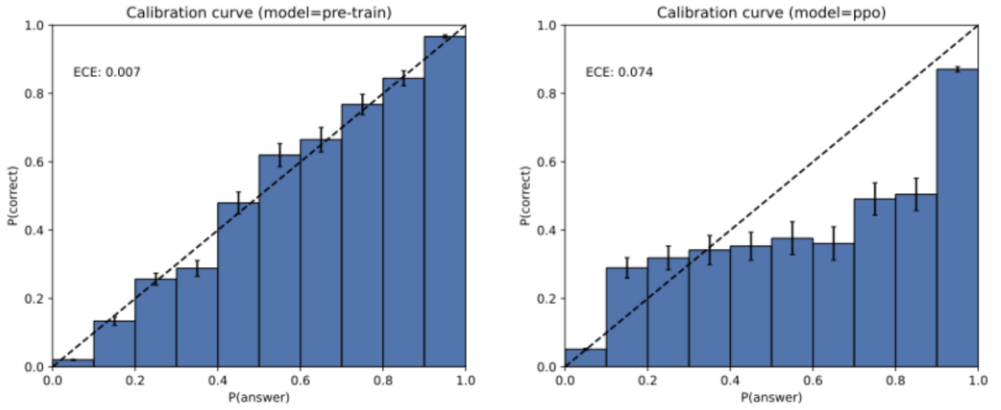


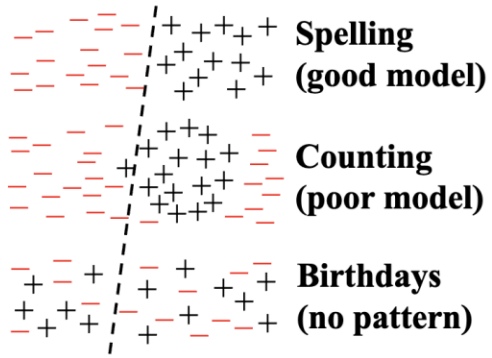
Figure 2: GPT-4 calibration histograms before (left) and after (right) reinforcement learning (OpenAI, 2023a, Figure 8, reprinted with permission). These plots are for multiple-choice queries where the plausible responses are simply A, B, C, or D. The pretrained model is well calibrated.

As a result, err will be high

Corollary 1. *For any training distribution p such that $p(\mathcal{V}) = 1$ and any base model $\hat{p}$,*

$$\text{err} \geq 2 \cdot \text{err}_{\text{iiv}} - \frac{|\mathcal{V}|}{|\mathcal{E}|} - \delta,$$

Valid examples +	**Error examples** –
Greetings.	Greatings.
How can I help?	How kan eye help?
There are 2 D's in LADDER.	There are 3 L's in SPELL.
There is 1 N in PIANO.	There is 1 G in CAT.
Mia Holdner's birthday is 4/1.	Colin Merivale's birthday is 8/29.
I don't know Zdan's birthday.	Jago Pere's birthday is 8/21.



Spelling (good model)

Counting (poor model)

Birthdays (no pattern)

Post-training further promotes errors

- Models are post-trained, often with the aim of achieving high benchmark performance
- Most benchmarks use binary grading. Therefore, like in human exams, guessing will always be superior to abstaining from an answer.

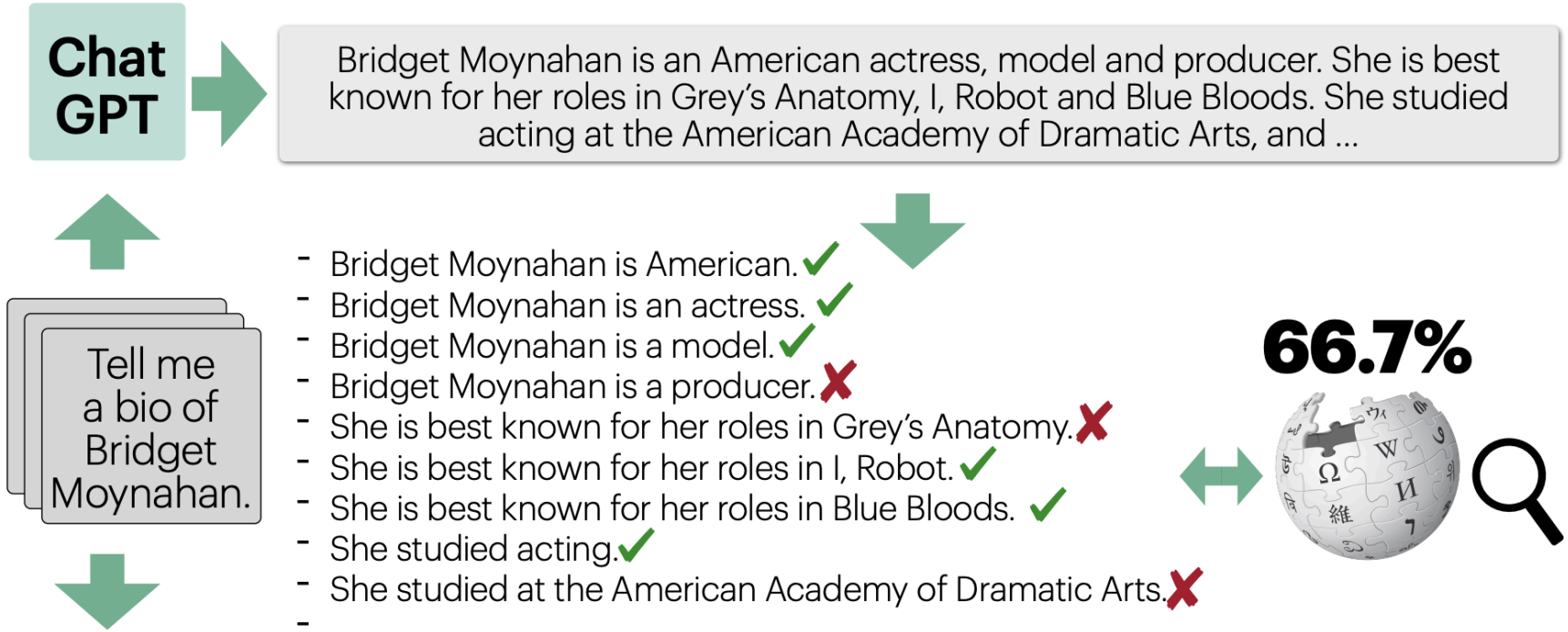
Benchmark	Scoring method	Binary grading	IDK credit
GPQA	Multiple-choice accuracy	Yes	None
MMLU-Pro	Multiple-choice accuracy	Yes	None
IFEval	Programmatic instruction verification	Yes[a]	None
Omni-MATH	Equivalence grading*	Yes	None
WildBench	LM-graded rubric*	No	Partial[b]
BBH	Multiple-choice / exact-match	Yes	None
MATH (L5 split)	Equivalence grading*	Yes	None
MuSR	Multiple-choice accuracy	Yes	None
SWE-bench	Patch passes unit tests	Yes	None
HLE	Multiple-choice / equivalence grading*	Yes	None

Agenda

- Introduction
- Causes of Hallucinations: Kalai et al. ”Why language models hallucinate”
- **Detection Strategies**
- Mitigation Strategies

FactScore

1. Evaluating the factuality of long-form text generated by large language models (LMs) is nontrivial because:
   - generations often contain a **mixture of supported and unsupported** pieces of information, making **binary judgments** inadequate
   - Human evaluation is time-consuming and costly
1. FACTSCORE
   - breaks a generation into a series of atomic facts and computes the percentage of atomic facts supported by a reliable source
   - ChatGPT only achieves 58%


Chat GPT
Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...
Tell me a bio of Bridget Moynahan.
- Bridget Moynahan is American. ✓
- Bridget Moynahan is an actress. ✓
- Bridget Moynahan is a model. ✓
- Bridget Moynahan is a producer. ✗
- She is best known for her roles in Grey's Anatomy. ✗
- She is best known for her roles in I, Robot. ✓
- She is best known for her roles in Blue Bloods. ✓
- She studied acting. ✓
- She studied at the American Academy of Dramatic Arts. ✗
- ...
66.7%

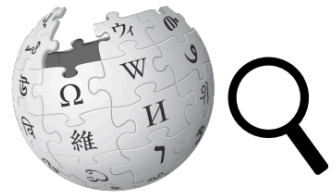
Chat GPT

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

Tell me a bio of Bridget Moynahan.

- Bridget Moynahan is American. ✓
- Bridget Moynahan is an actress. ✓
- Bridget Moynahan is a model. ✓
- Bridget Moynahan is a producer. ✗
- She is best known for her roles in Grey's Anatomy. ✗
- She is best known for her roles in I, Robot. ✓
- She is best known for her roles in Blue Bloods. ✓
- She studied acting. ✓
- She studied at the American Academy of Dramatic Arts. ✗
- ...

66.7%

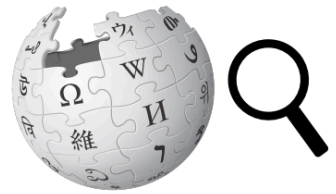


Stable LM

Bridget Moynahan is an American filmmaker and writer. She is best known for her work on the soap opera General Hospital, which she co-created with husband Charles Kelly. Moynahan was raised in a middle-class family in Los Angeles, ...

- Bridget Moynahan is American. ✓
- Bridget Moynahan is a filmmaker. ✗
- Bridget Moynahan is a writer. ✗
- She is best known for her work on General Hospital. ✗
- General Hospital is the soap opera. ✗
- She co-created General Hospital. ✗
- She co-created General Hospital with her husband. ✗
- Her husband is Charles Kelly. ✗
- Moynahan was raised in a middle-class family. ✗
- Moynahan was raised in Los Angeles. ✗
- ...

10.0%



Category	%	Example
Single-sentence contradiction (words)	33.3	Gen On November 25th, 2023, Glover Teixeira became an American citizen. Wiki In November 2020, Teixeira became an American citizen. Gen [Eric Hacker] was named the International League Pitcher of the Year. Wiki [Eric Hacker] was named the IL Pitcher of the Week.
Single-sentence contradiction (beyond words)	10.0	Gen William Waldegrave's grandfather was James II and VII. Wiki His father's title was created ... for the diplomat and ambassador James Waldegrave, 1st Earl Waldegrave, whose grandfather was James II and VII. Gen She has appeared in several successful films such as (...) and Zero (2018). Wiki: Zero was a commercial failure.

Category	%	Example
Single-sentence contradiction (words)	33.3	Gen On November 25th, 2023, Glover Teixeira became an American citizen. Wiki In November 2020, Teixeira became an American citizen. Gen [Eric Hacker] was named the International League Pitcher of the Year. Wiki [Eric Hacker] was named the IL Pitcher of the Week.
Single-sentence contradiction (beyond words)	10.0	Gen William Waldegrave's grandfather was James II and VII. Wiki His father's title was created ... for the diplomat and ambassador James Waldegrave, 1st Earl Waldegrave, whose grandfather was James II and VII. Gen She has appeared in several successful films such as (...) and Zero (2018). Wiki: Zero was a commercial failure.
Page-level contradiction	23.3	Gen Some of [Julia Faye's] notable films include ... "Cleopatra" (1934). Comment No mention of *Cleopatra* on the *Julia Faye* page, and no mention of *Julia Faye* on the *Cleopatra* page. Gen [Kang Ji-hwan] has donated money to various charities and organizations over the years. Comment No such mention on the *Kang Ji-hwan* page.

Category	%	Example
Single-sentence contradiction (words)	33.3	Gen On November 25th, 2023, Glover Teixeira became an American citizen. Wiki In November 2020, Teixeira became an American citizen. Gen [Eric Hacker] was named the International League Pitcher of the Year. Wiki [Eric Hacker] was named the IL Pitcher of the Week.
Single-sentence contradiction (beyond words)	10.0	Gen William Waldegrave's grandfather was James II and VII. Wiki His father's title was created ... for the diplomat and ambassador James Waldegrave, 1st Earl Waldegrave, whose grandfather was James II and VII. Gen She has appeared in several successful films such as (...) and Zero (2018). Wiki: Zero was a commercial failure.
Page-level contradiction	23.3	Gen Some of [Julia Faye's] notable films include ... "Cleopatra" (1934). Comment No mention of *Cleopatra* on the *Julia Faye* page, and no mention of *Julia Faye* on the *Cleopatra* page. Gen [Kang Ji-hwan] has donated money to various charities and organizations over the years. Comment No such mention on the *Kang Ji-hwan* page.
Subjective	16.7	Gen His achievements, as an actor and as a cultural force, will surely prove to be as heroic as those of the characters he portrayed. Wiki Culture writer Steve Rose, in The Guardian, wrote: "Chadwick Boseman began his career playing African American icons and pioneers; he ends it as one himself. His [...] achievements, as an actor and as a cultural force, will surely prove to be as heroic as those of the characters he portrayed."

Category	%	Example
Single-sentence contradiction (words)	33.3	Gen On November 25th, 2023, Glover Teixeira became an American citizen. Wiki In November 2020, Teixeira became an American citizen. Gen [Eric Hacker] was named the International League Pitcher of the Year. Wiki [Eric Hacker] was named the IL Pitcher of the Week.
Single-sentence contradiction (beyond words)	10.0	Gen William Waldegrave's grandfather was James II and VII. Wiki His father's title was created ... for the diplomat and ambassador James Waldegrave, 1st Earl Waldegrave, whose grandfather was James II and VII. Gen She has appeared in several successful films such as (...) and Zero (2018). Wiki: Zero was a commercial failure.
Page-level contradiction	23.3	Gen Some of [Julia Faye's] notable films include ... "Cleopatra" (1934). Comment No mention of *Cleopatra* on the *Julia Faye* page, and no mention of *Julia Faye* on the *Cleopatra* page. Gen [Kang Ji-hwan] has donated money to various charities and organizations over the years. Comment No such mention on the *Kang Ji-hwan* page.
Subjective	16.7	Gen His achievements, as an actor and as a cultural force, will surely prove to be as heroic as those of the characters he portrayed. Wiki Culture writer Steve Rose, in The Guardian, wrote: "Chadwick Boseman began his career playing African American icons and pioneers; he ends it as one himself. His [...] achievements, as an actor and as a cultural force, will surely prove to be as heroic as those of the characters he portrayed."
Fact is irrelevant	3.3	Gen [Zamfir Arbore]'s life is not well-documented, and there is little information available about him.
Wiki is inconsistent & wrong	3.3	Gen Kick (2014) that brought [Sajid Nadiadwala] various debutant director awards. Wiki 2015, IIFA Award for Debut Director, Kick. (...) Kick brought him various debutant director awards. Comment The first text is from a table that indicates he won one award (accurate). The second is inaccurate, incorrectly citing a news article.

Category	%	Example
Single-sentence contradiction (words)	33.3	Gen On November 25th, 2023, Glover Teixeira became an American citizen. Wiki In November 2020, Teixeira became an American citizen. Gen [Eric Hacker] was named the International League Pitcher of the Year. Wiki [Eric Hacker] was named the IL Pitcher of the Week.
Single-sentence contradiction (beyond words)	10.0	Gen William Waldegrave's grandfather was James II and VII. Wiki His father's title was created ... for the diplomat and ambassador James Waldegrave, 1st Earl Waldegrave, whose grandfather was James II and VII. Gen She has appeared in several successful films such as (...) and Zero (2018). Wiki: Zero was a commercial failure.
Page-level contradiction	23.3	Gen Some of [Julia Faye's] notable films include ... "Cleopatra" (1934). Comment No mention of *Cleopatra* on the *Julia Faye* page, and no mention of *Julia Faye* on the *Cleopatra* page. Gen [Kang Ji-hwan] has donated money to various charities and organizations over the years. Comment No such mention on the *Kang Ji-hwan* page.
Subjective	16.7	Gen His achievements, as an actor and as a cultural force, will surely prove to be as heroic as those of the characters he portrayed. Wiki Culture writer Steve Rose, in The Guardian, wrote: "Chadwick Boseman began his career playing African American icons and pioneers; he ends it as one himself. His [...] achievements, as an actor and as a cultural force, will surely prove to be as heroic as those of the characters he portrayed."
Fact is irrelevant	3.3	Gen [Zamfir Arbore]'s life is not well-documented, and there is little information available about him.
Wiki is inconsistent & wrong	3.3	Gen Kick (2014) that brought [Sajid Nadiadwala] various debutant director awards. Wiki 2015, IIFA Award for Debut Director, Kick. (...) Kick brought him various debutant director awards. Comment The first text is from a table that indicates he won one award (accurate). The second is inaccurate, incorrectly citing a news article.
Annotation error	10.0	Gen [Zamfir Arbore] was part of the staff of Românul. Wiki The Românul staff came to include Zamfir Arbore. Comment Mentioned in the *Românul* page but not in the *Zamfir Arbore* page.

Estimating FactScore for Automatic Evaluation

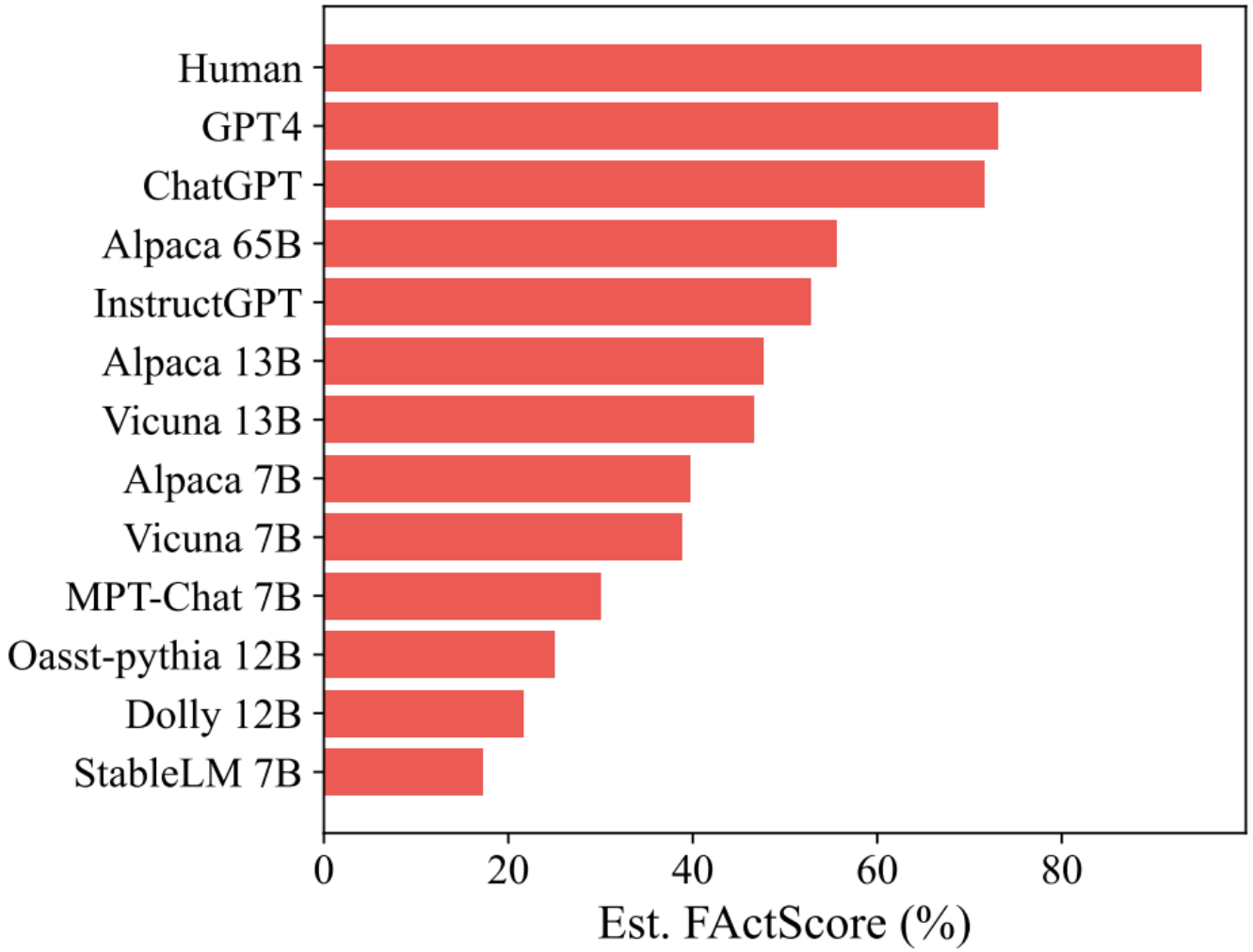
1. Our estimator of FACTSCORE first (i) breaks a generation into a series of atomic facts and then (ii) validates each against the given knowledge source

   - No-context LM: uses <atomic-fact> True or False?
   - Retrieve→LM
   - Nonparametric Probability (NP): averages probabilities over all tokens, and makes a prediction based on thresholding
   - Retrieve→LM + NP

2. Finding: Using retrieval is consistently better than No-context LM

Results

1. All LMs are substantially less factual than humans
2. GPT-4 and ChatGPT are comparable in factual precision
3. There is a clear correlation between the model size and factual precision
4. Within public models, there are large gaps in factual precision even when the model size is similar:
    - Possible factors include the choice of the base LM, the data, and the training recipe



Agenda

- Introduction
- Causes of Hallucinations: Kalai et al. ”Why language models hallucinate”
- Detection Strategies
- **Mitigation Strategies**

Retrieval-Augmented Generation

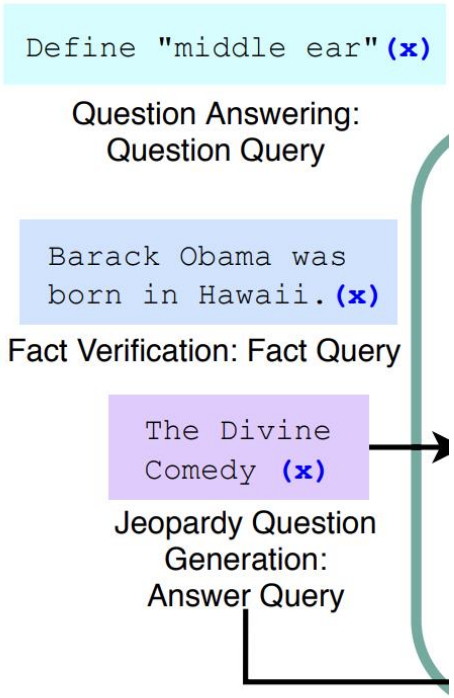
1. Premise:

   - LLM ability to access and precisely manipulate knowledge is still limited on knowledge-intensive tasks, their performance lags behind task-specific architectures

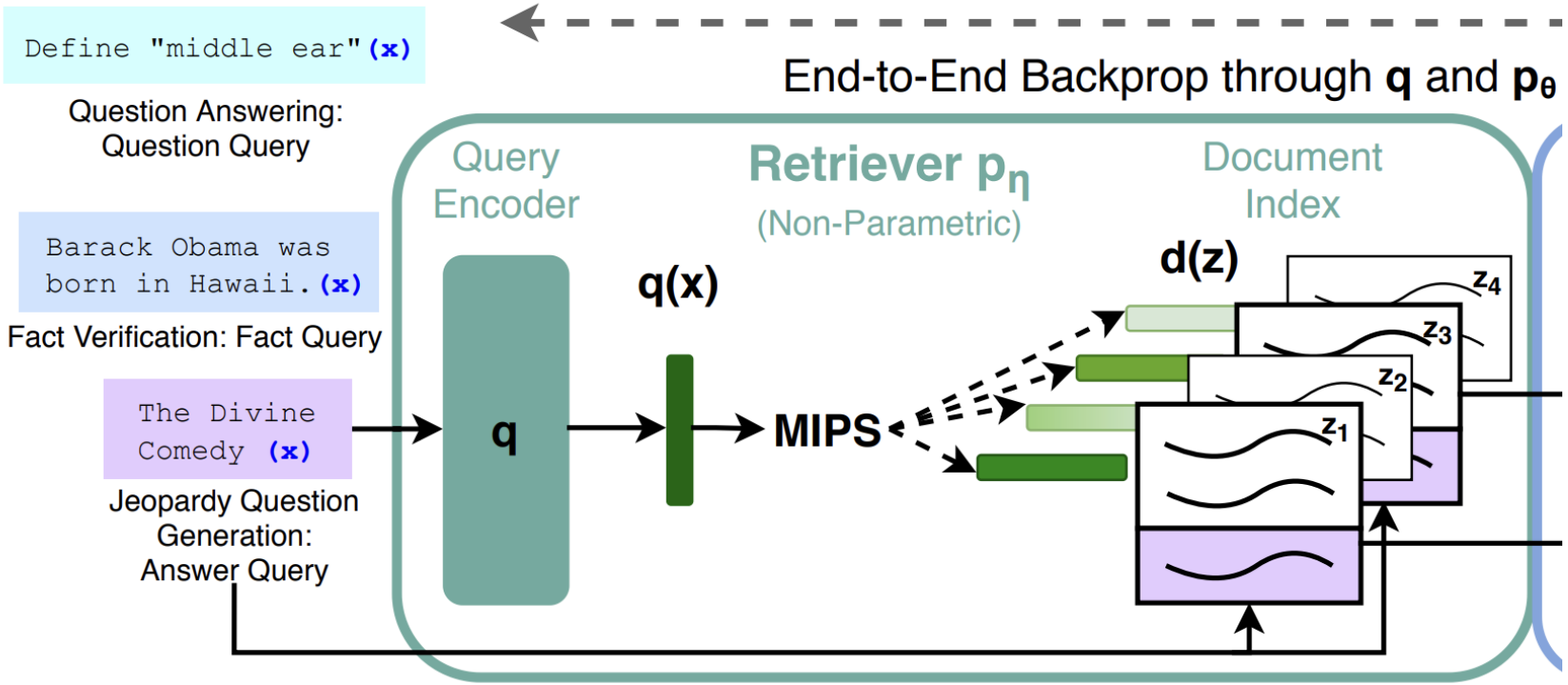
2. RAG models where the parametric memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia

3. For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline.

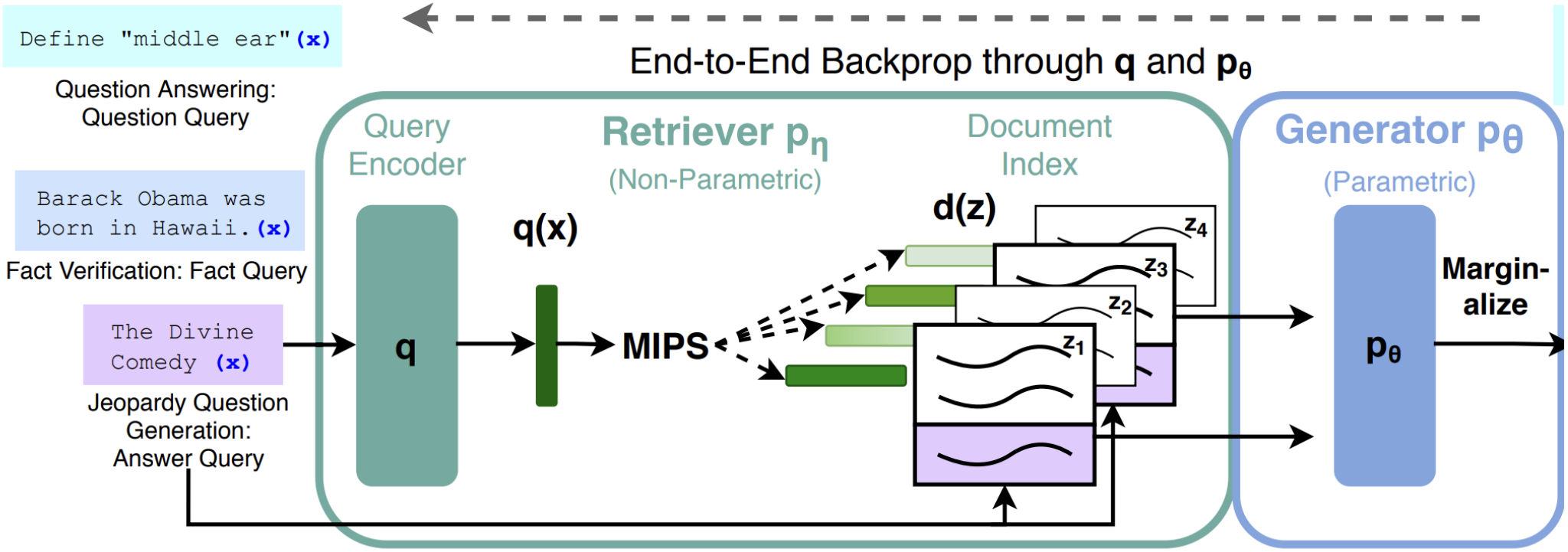
Retrieval-Augmented Generation



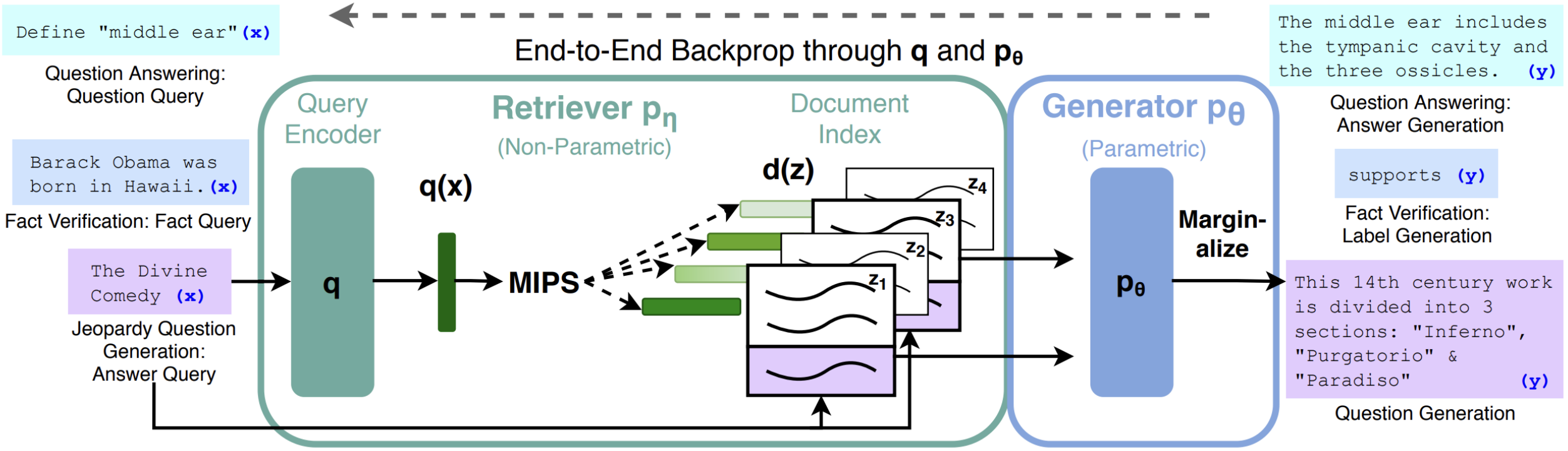
Retrieval-Augmented Generation



Retrieval-Augmented Generation



Retrieval-Augmented Generation



Chain-of-Verification

1. Premise:

   - Generation of plausible yet incorrect factual information, termed hallucination, is an unsolved issue in large language models

2. We develop the Chain-of-Verification (COVE) method whereby the model first

(i) drafts an initial response; then

(ii) plans verification questions to fact-check its draft;

(iii) answers those questions independently so the answers are not biased by other responses;

(iv) generates its final verified response

Query

Name some politicians who were born in NY, New York.

Query

Name some politicians who were born in NY, New York.

1. Baseline Response

Here are some politicians who were born in NY, New York:
1. Hillary Clinton - former secretary of state and former Democratic presidential nominee
2. Donald Trump - former president of the United States
3. Michael Bloomberg - former Mayor of New York City and former Democratic presidential candidate
... <list continues..>

Query

Name some politicians who were born in NY, New York.

1. Baseline Response

Here are some politicians who were born in NY, New York:
1. Hillary Clinton - former secretary of state and former Democratic presidential nominee
2. Donald Trump - former president of the United States
3. Michael Bloomberg - former Mayor of New York City and former Democratic presidential candidate

... <list continues..>

2. Plan Verifications

- Where was Hillary Clinton born?

- Where was Donald Trump born?

- Where was Michael Bloomberg born?

... <questions continue..>

3. Execute Verifications

Hillary Clinton was born in **Chicago, Illinois**, United States on October 26, 1947.

Donald Trump was born on June 14, 1946, in **Queens, New York City, New York**, United States

Michael Bloomberg was born on February 14, 1942, in **Boston, Massachusetts**, United States.

...

Query

Name some politicians who were born in NY, New York.

1. Baseline Response

Here are some politicians who were born in NY, New York:
1. Hillary Clinton - former secretary of state and former Democratic presidential nominee
2. Donald Trump - former president of the United States
3. Michael Bloomberg - former Mayor of New York City and former Democratic presidential candidate

... <list continues..>

2. Plan Verifications

- Where was Hillary Clinton born?

- Where was Donald Trump born?

- Where was Michael Bloomberg born?

... <questions continue..>

3. Execute Verifications

Hillary Clinton was born in **Chicago, Illinois**, United States on October 26, 1947.

Donald Trump was born on June 14, 1946, in **Queens, New York City, New York**, United States

Michael Bloomberg was born on February 14, 1942, in **Boston, Massachusetts**, United States.

...

4. Final Verified Response

Here are some politicians who were born in NY, New York:
1. Donald Trump - former president of the United States
2. Alexandria Ocasio-Cortez - Democratic member of the U.S. House of Representatives

... <list continues..>

Conclusion

- Hallucinations arise naturally due to statistical pressures during pre-training and are rewarded by most benchmarks
- A variety of mitigation approaches exist
- However, so far there is no *fundamental* solution of the issue