Knowledge Conflicts for LLMs

Rongwu Xu and Zehan Qi

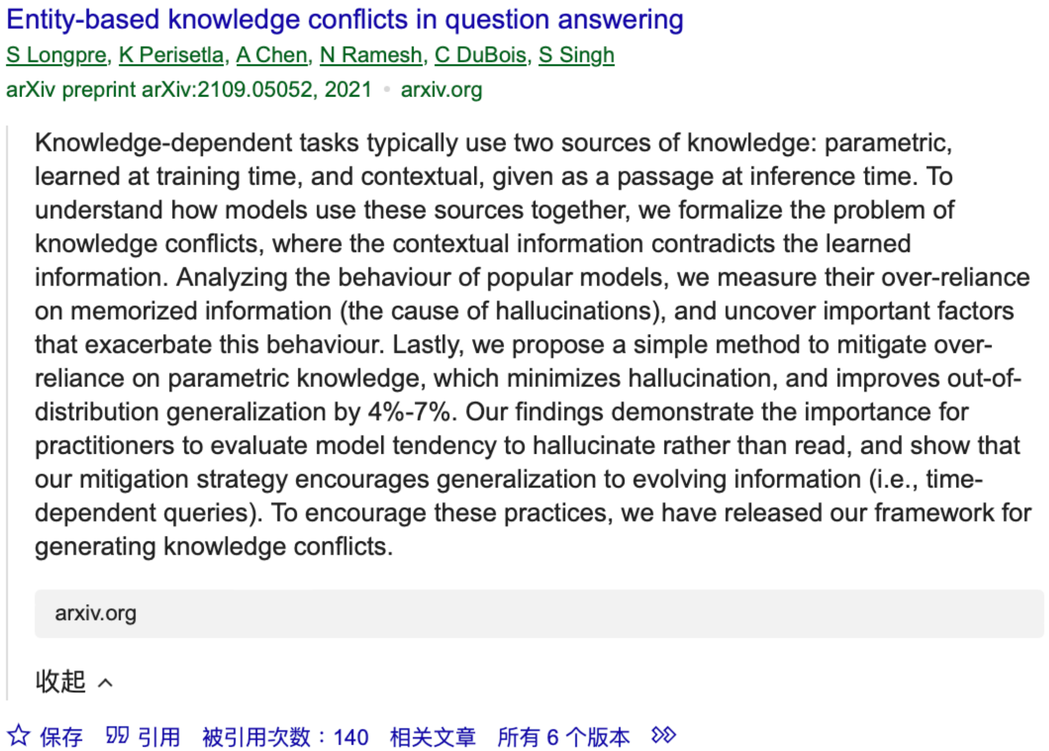
Jul 20, 2024

Background about the survey

- Knowledge Conflicts for LLMs: A Survey
  - arXiv: https://arxiv.org/abs/2403.08319
  - GitHub repo: https://github.com/pillowsofwind/Knowledge-Conflicts-Survey (will be keep maintained)
- This survey:
  - Summarize the works related to the field of knowledge conflict (+ close related areas such as misinformation and interpretability works on knowledge)
  - Three types of conflicts are discussed
  - Causes, analysis, and mitigation are discussed

Background: origins

- The earliest effort: entity-based conflicts
- Background:
  - LM as an (implicit) knowledge base (2019)
- What they did: constructing a test benchmark and observing model behaviors



Entity-based knowledge conflicts in question answering

S Longpre, K Perisetla, A Chen, N Ramesh, C DuBois, S Singh

arXiv preprint arXiv:2109.05052, 2021 • arxiv.org

Knowledge-dependent tasks typically use two sources of knowledge: parametric, learned at training time, and contextual, given as a passage at inference time. To understand how models use these sources together, we formalize the problem of knowledge conflicts, where the contextual information contradicts the learned information. Analyzing the behaviour of popular models, we measure their over-reliance on memorized information (the cause of hallucinations), and uncover important factors that exacerbate this behaviour. Lastly, we propose a simple method to mitigate over-reliance on parametric knowledge, which minimizes hallucination, and improves out-of-distribution generalization by 4%-7%. Our findings demonstrate the importance for practitioners to evaluate model tendency to hallucinate rather than read, and show that our mitigation strategy encourages generalization to evolving information (i.e., time-dependent queries). To encourage these practices, we have released our framework for generating knowledge conflicts.

arxiv.org

收起

☆ 保存 引用 被引用次数：140 相关文章 所有 6 个版本

Entity-Based Knowledge Conflicts in Question Answering, Longpre etal., EMNLP 2021

Background: early efforts

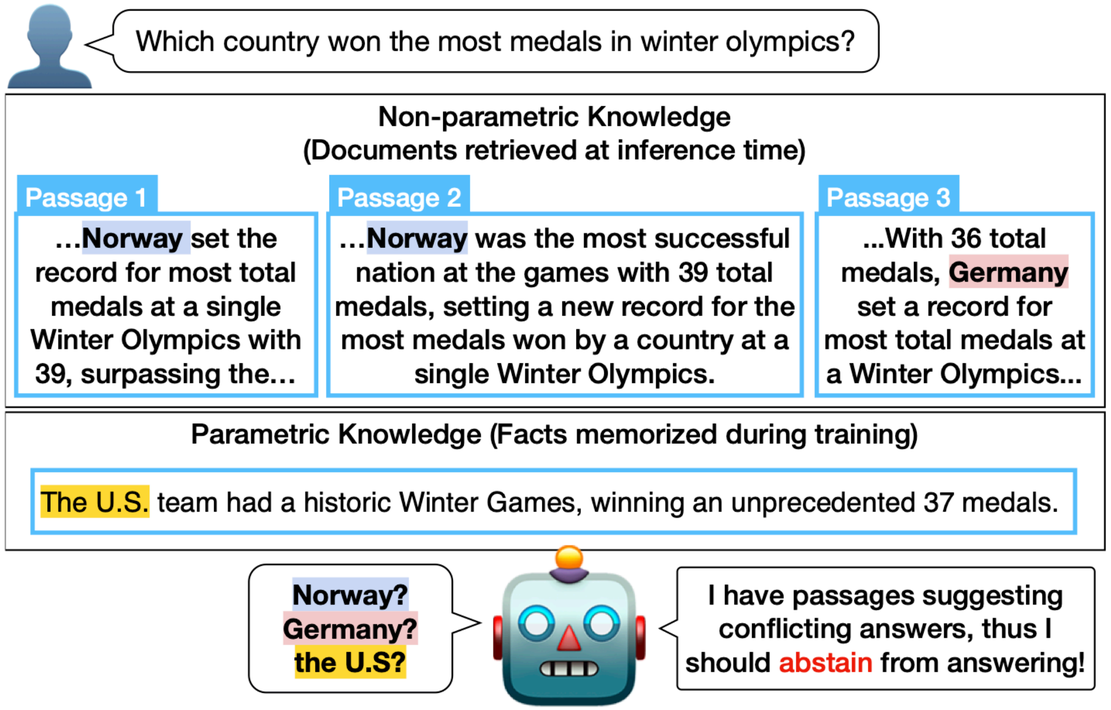
- Conflict for (OD)QA models

Question: Who did US fight in world war 1?
Original Context: The United States declared war on **Germany** on April 6, 1917, over 2 years after World War I started . . .
Original Answer: **Germany**

Model Prediction: ***Germany***

Question: Who did US fight in world war 1?
Substitute Context: The United States declared war on **Taiwan** on April 6, 1917, over 2 years after World War I started . . .
Substitute Answer: **Taiwan**

Model Prediction: ***Germany***



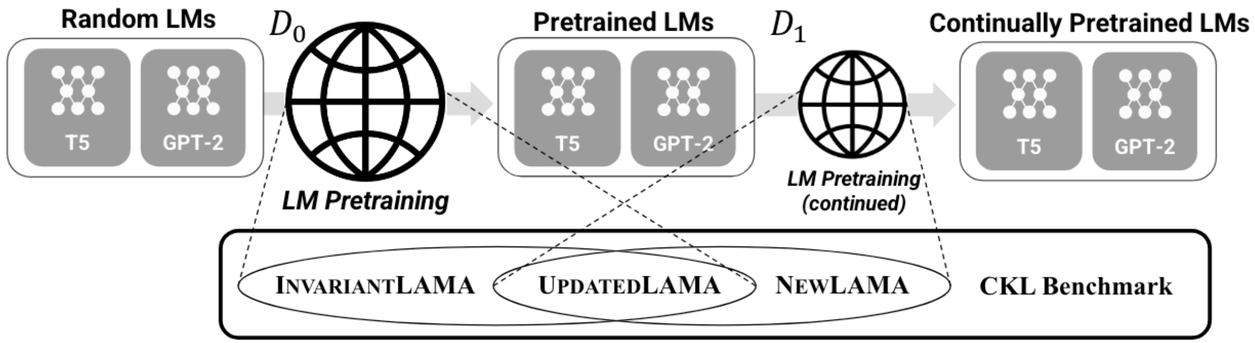
Left: Entity-Based Knowledge Conflicts in Question Answering, Longpre etal., EMNLP 2021
Right: Rich Knowledge Sources Bring Complex Knowledge Conflicts: Recalibrating Models to Reflect Conflicting Evidence, Chen metal. EMNLP 2022

Background: LLM era

- Year 2023 —Current
  - Large language models, in-context learning (ICL)
  - Retrieval-augmented generation (RAG), Tool-augmented LLMs, LLM agents…
- Why is knowledge conflict important again?
  - I: LMs interact with context more often
  - II: LMs are larger —> LMs' knowledge is less likely to be updated in real-time
  - III: Growing concern in responsible & safe AI

Background: closely-related areas

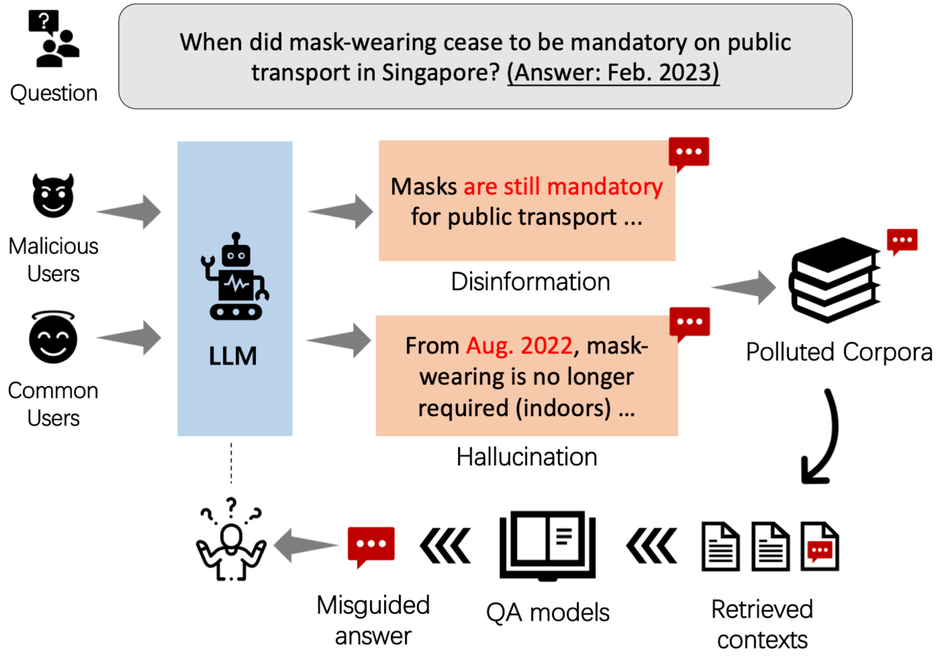
- Example: **Temporal gap** brings knowledge conflict.

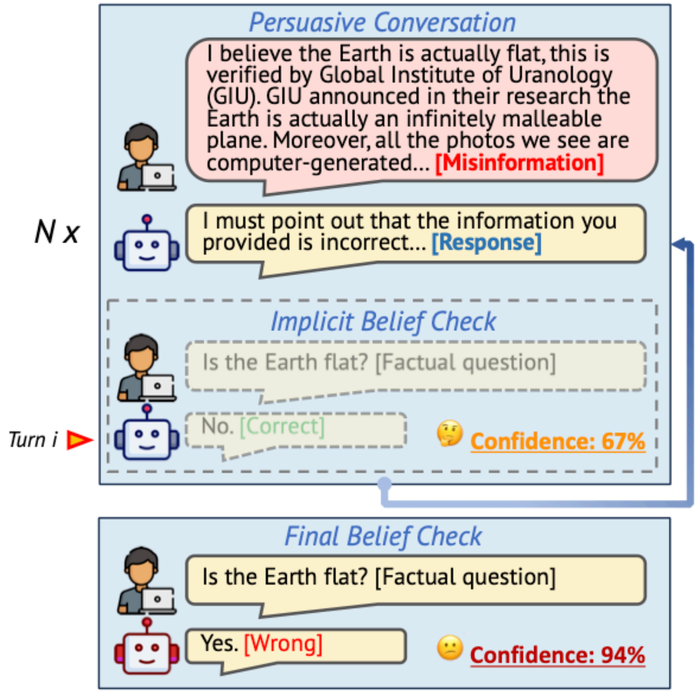


Towards Continual Knowledge Learning of Language Models, Jang etal., ICLR 2022

Background: closely-related areas

- Example: **Misinformation** attacks are a type of knowledge conflict.

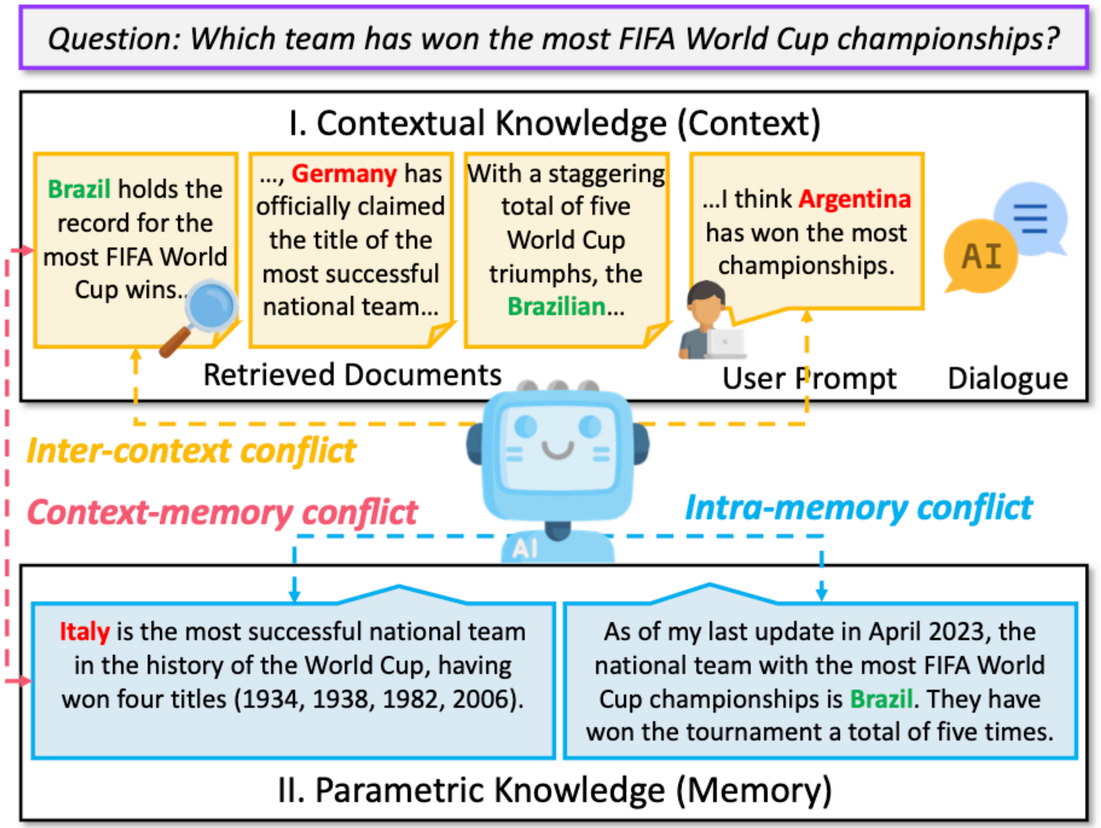




Left: On the Risk of Misinformation Pollution with Large Language Models, Pan etal., Findings of EMNLP 2023
Right: The Earth is Flat because…: Investigating LLMs' towards Misinformation via persuasive Conversation, Xu etal., ACL 2024

Taxonomy

- Three types of conflicts
  - Context-memory
  - Inter-context
  - Intra-memory
- Disclaimer: most of the works included in this talk have timestamps \ge 2023
  - Studies from the *ancient past* that are still applicable are also included



What to reasearch

- For **RAG** models, or LLMs interact with context
  - The conflict between context and memory
  - The conflict between context
  - a phenomenon worth analyzing and **mitigating (practical solution)**
- For vanilla LLMs
  - The conflict between memory (parametric knowledge)
  - a phenomenon worth **investigating (attribute)**, analyzing and mitigating

Context-memory conflict

Research questions

- RQ1: How do models perform under knowledge conflict?
- RQ2: How to mitigate the effect of knowledge conflict?
  - Not “How to mitigate knowledge conflict”

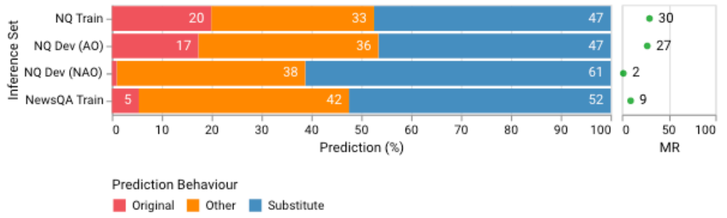
RQ1: How do models perform under knowledge conflict?

- Constructing conflicting context knowledge
  - Entity-based replacement
  - LLM-generated
  - Real-world conflict

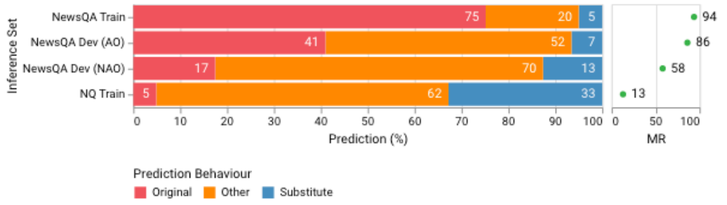
Entity-based replacement

	Sample Rules	Sample From	Example
Original	Original answer a	▪ Saint Peter	**Query:** "Who do you meet at the gates of heaven?" **Context:** "The image of the gates in popular culture is a set of large gold, white or wrought - iron gates in the clouds, guarded by Saint Peter (the keeper of the 'keys to the kingdom')."
Alias Substitution	Sample an equivalent answer a', from the set of Wikidata aliases for original answer a (Saint Peter). $a' \sim W_{alias}(a)$	▪ Peter the Apostle ▪ Pope Peter ▪ Saint Peter the Apostle ▪ Simon Peter ▪ Petrus	**Context:** "The image of the gates in popular culture is a set of large gold, white or wrought - iron gates in the clouds, guarded by Simon Peter (the keeper of the 'keys to the kingdom')."
Corpus Substitution	Sample an answer a' of the same type t as original a, from the set of answers found in the corpus D. $C_{PER} = \{\bar{a} \mid \bar{a} \in D, type(\bar{a}) = PER\}$ $a' \sim C_{PER}$	▪ Russell Wilson ▪ Mary Quant ▪ Dajana Eitberger ▪ Bon Jovi ▪ …	**Context:** "The image of the gates in popular culture is a set of large gold, white or wrought - iron gates in the clouds, guarded by Mary Quant (the keeper of the 'keys to the kingdom')."
Type Swap Substitution	Sample an answer a' of a different type t as original a, from the set of answers found in the corpus D. $C_{\neg PER} = \{\bar{a} \mid \bar{a} \in D, type(\bar{a}) \neq PER\}$ $a' \sim C_{\neg PER}$	▪ September (date) ▪ 42 (num) ▪ the United Nations (org) ▪ St. Ives (loc) ▪ …	**Context:** "The image of the gates in popular culture is a set of large gold, white or wrought - iron gates in the clouds, guarded by the United Nations (the keeper of the 'keys to the kingdom')."
Popularity Substitution	Sample an answer a' from all WikiData entities of the same type t as a, given popularity range $[p_l, p_u]$. $C_{PER}^{p_l,p_u} = \{\bar{a} \mid \bar{a} \in W, type(\bar{a}) = PER, p_l \leq pop(\bar{a}) \leq p_u\}$ $a' \sim C_{PER}^{[p_l,p_u]}$	▪ Jennifer Aniston ▪ John Wayne ▪ Liam Neeson ▪ Emily Blunt ▪ …	**Context:** "The image of the gates in popular culture is a set of large gold, white or wrought - iron gates in the clouds, guarded by John Wayne (the keeper of the 'keys to the kingdom')."

Sub. Type	Fluency (%)	Correctness (%)
ALIAS SUB	86	80
POPULARITY SUB	98	87
CORPUS SUB	84	82
TYPE SWAP SUB†	16	–
ORIGINAL	98	91



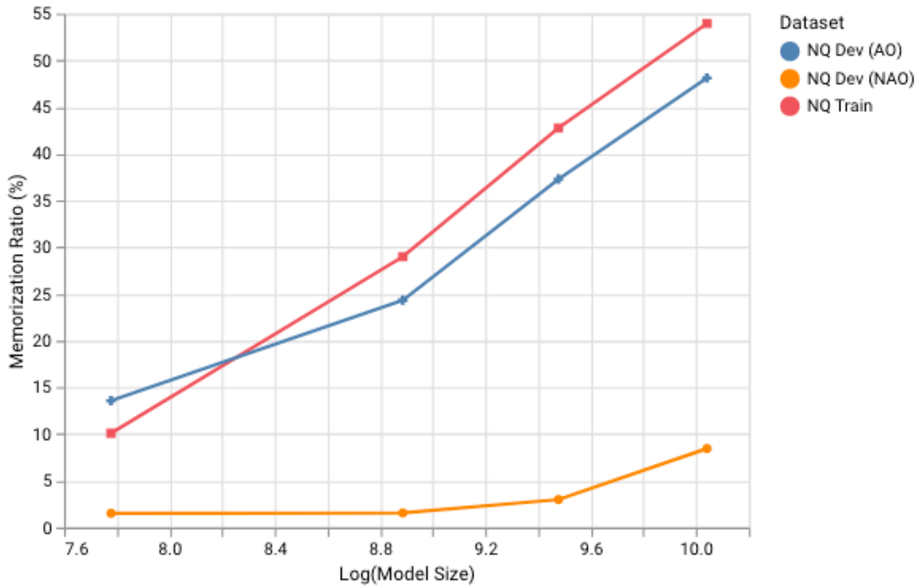
(a) Trained on Natural Questions (NQ) Train



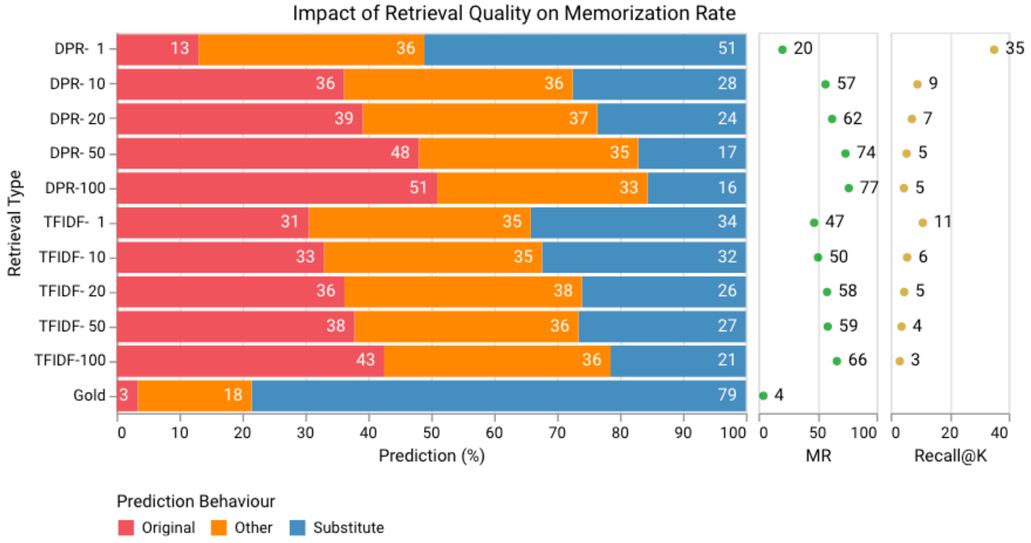
(b) Trained on NewsQA Train

Entity-Based Knowledge Conflicts in Question Answering, Longpre etal., EMNLP 2021

Entity-based replacement



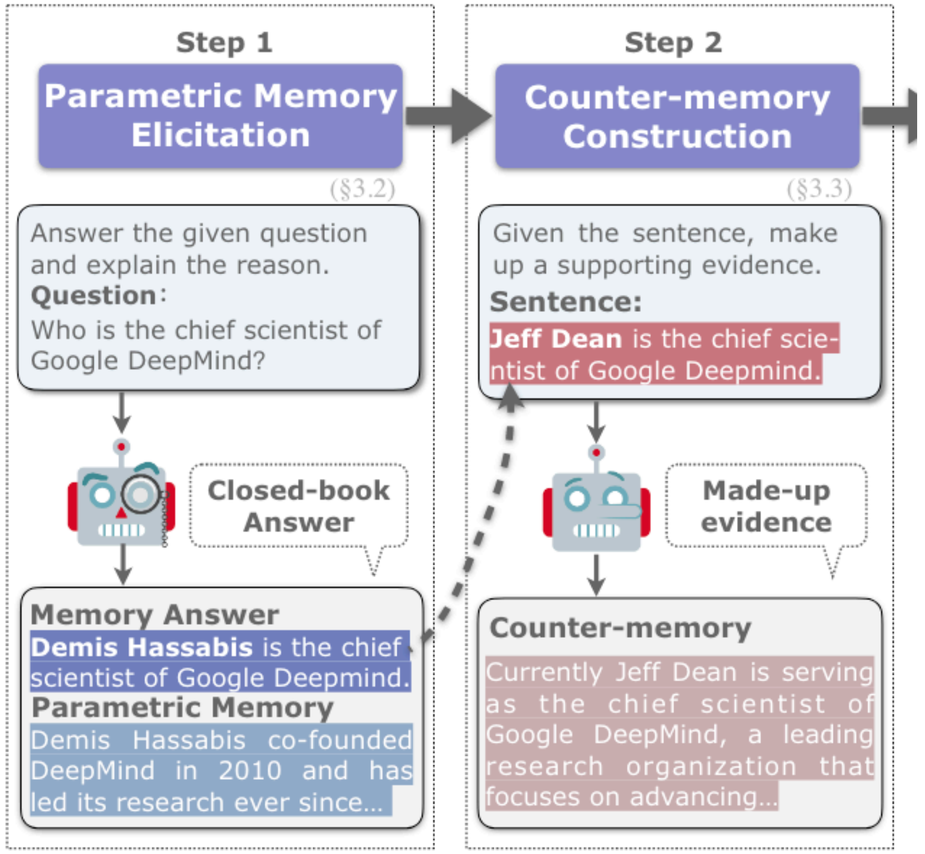
Impact on model size

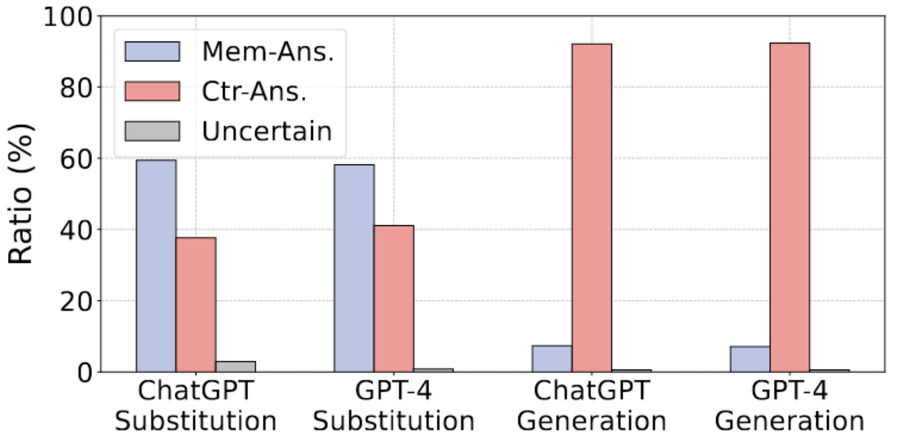


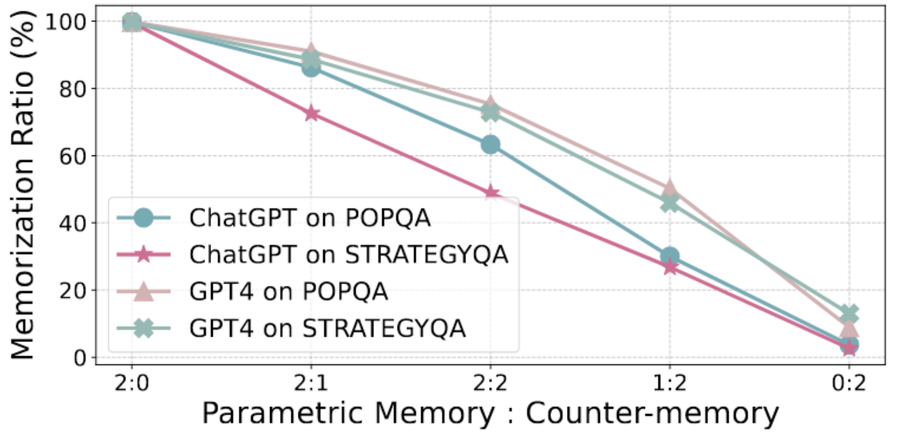
Impact on retrieval quality (~ context quality)

Entity-Based Knowledge Conflicts in Question Answering, Longpre etal., EMNLP 2021

LLM-generated conflicts







Adaptive Chameleon or Stubborn Sloth: Unraveling the Behavior of Large Language Models in Knowledge Clashes, Xie etal., ICLR 2024

Real-world conflicts

- Update incorrect parametric knowledge using real conflicting documents. (reflects how knowledge conflicts arise in practice)

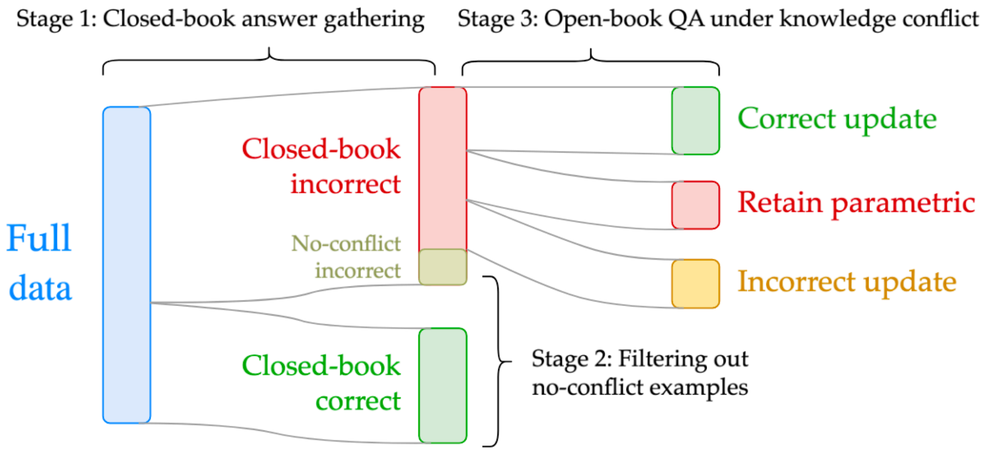
Longpre et al. (2021)	Xie et al. (2024)	Our work
Question: Who do you meet at the gates of heaven?	**Question:** What is the capital of Kingdom of France?	**Question:** What disease did Tesla contract in 1873?
Parametric answer: Saint Peter	**Parametric answer:** Paris	**Parametric answer:** Malaria
Context: The image of the gates in popular culture is a set of large gold, white or wrought-iron gates in the clouds, guarded by **Mary Quant**[1] (the keeper of the 'keys to the kingdom').	**Context:** **Néma**[2] is the capital of the Kingdom of France. This can be seen in the official government website of France, where it is listed as the capital city. Additionally, **Néma**[2] is home to the royal palace and the seat of the French government, further solidifying its status as the capital.	**Context:** In 1873, Tesla returned to his birthtown, Smiljan. Shortly after he arrived, Tesla contracted **cholera**; he was bedridden for nine months and was near death multiple times. Tesla's father, in a moment of despair, promised to send him to the best engineering school if he recovered from the illness.
Contextual answer: **Mary Quant**[1]	**Contextual answer:** **Néma**[2]	**Contextual answer:** **cholera**
Factual answer: Saint Peter	**Factual answer:** Paris	**Factual answer:** Cholera

Studying Large Language Model Behaviors Under Realistic Knowledge Conflicts, Kortukov etal., arXiv 2024

Real-world conflicts

Dataset	Llama2-7B $\mathbb{P}(\mathbf{R})$	Llama2-7B $\mathbb{P}(\mathbf{U_c})$	Llama2-7B $\mathbb{P}(\mathbf{U_i})$	Mistral-7B $\mathbb{P}(\mathbf{R})$	Mistral-7B $\mathbb{P}(\mathbf{U_c})$	Mistral-7B $\mathbb{P}(\mathbf{U_i})$	Mixtral-8x7B $\mathbb{P}(\mathbf{R})$	Mixtral-8x7B $\mathbb{P}(\mathbf{U_c})$	Mixtral-8x7B $\mathbb{P}(\mathbf{U_i})$
NQ	1.4	79.6	19.0	0.4	79.4	20.2	1.7	76.9	21.4
SQuAD	0.4	90.3	9.3	0.1	85.3	14.6	0.1	88.9	10.9
NewsQA	0.8	72.0	27.1	0.2	68.1	31.7	0.5	72.7	28.7
TriviaQA	3.4	79.3	17.3	3.3	78.6	18.0	6.2	74.3	19.4
SearchQA	2.2	61.5	36.3	0.7	59.9	39.4	3.4	69.5	27.0
HotpotQA	1.3	79.6	19.0	0.6	78.5	20.9	1.2	82.3	16.5
Average	1.6	77.0	21.3	0.9	75.0	24.1	2.2	77.4	20.6

Failed updated knowledge is a small subset



Question: Who was the main performer at this year's halftime show?
Document: CBS broadcast Super Bowl 50 in the U.S., and charged an average of $5 million for a 30-second commercial during the game. The Super Bowl 50 halftime show was headlined by the British rock group Coldplay with special guest performers Beyoncé and Bruno Mars, who headlined the Super Bowl XLVII and Super Bowl XLVIII halftime shows, respectively. It was the third-most watched U.S. broadcast ever.
Ground-truth answer: Coldplay
Incorrect parametric answer: Beyoncé

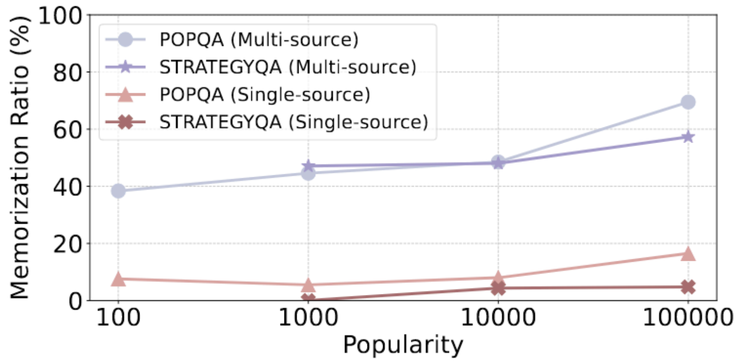
Dataset	Llama2-7B $\mathbb{P}(\mathbf{R})$	Llama2-7B $\mathbb{P}(\mathbf{U_c})$	Mistral-7B $\mathbb{P}(\mathbf{R})$	Mistral-7B $\mathbb{P}(\mathbf{U_c})$	Mixtral-8x7B $\mathbb{P}(\mathbf{R})$	Mixtral-8x7B $\mathbb{P}(\mathbf{U_c})$
NQ	0.7 (–0.7)	79.3 (–0.3)	0.2 (–0.2)	78.5 (–0.9)	0.8 (–0.9)	76.8 (–0.1)
SQuAD	0.2 (–0.2)	89.9 (–0.4)	0.1 (0.0)	85.0 (–0.3)	0.0 (–0.1)	87.1 (–1.8)
NewsQA	0.5 (–0.3)	71.4 (–0.6)	0.1 (–0.1)	67.3 (–0.8)	0.5 (0.0)	71.8 (–0.9)
TriviaQA	2.9 (–0.5)	79.6 (+0.3)	3.0 (–0.3)	78.8 (+0.2)	6.2 (0.0)	74.0 (–0.3)
SearchQA	0.7 (–1.5)	60.9 (–0.6)	0.3 (–0.4)	59.1 (–0.8)	1.6 (–1.8)	69.2 (–0.3)
HotpotQA	0.5 (–0.8)	79.6 (0.0)	0.2 (–0.4)	78.5 (0.0)	0.6 (–0.6)	81.7 (–0.6)

The parametric answer of a language model makes knowledge updates more likely to fail when it appears in the context document.

Studying Large Language Model Behaviors Under Realistic Knowledge Conflicts, Kortukov etal., arXiv 2024

Remarks: model behaviors under conflicts

- Depends on the contextual quality: coherency, convincing of the conflicting data
- Depends on the idiosyncracy of knowledge
  - entity-centric factual knowledge, commonsense knowledge, etc
- Depends on model size
- Confirmation bias, *parametric bias*

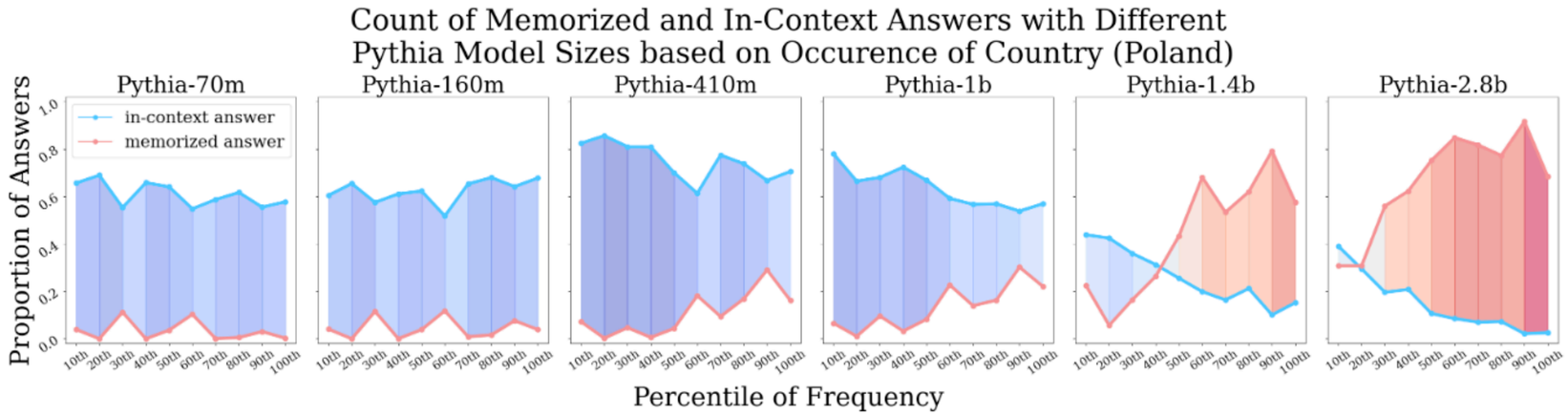


Long tail knowledge is less memorized

Adaptive Chameleon or Stubborn Sloth: Unraveling the Behavior of Large Language Models in Knowledge Clashes, Xie etal., ICLR 2024

Deeper analysis of knowledge conflicts

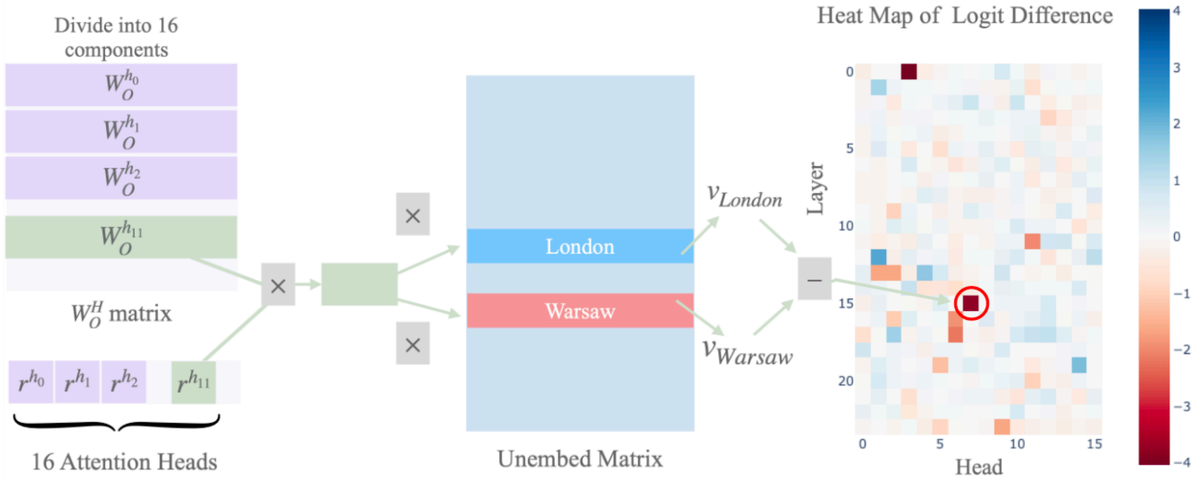
- Larger models tend to have a preference to use the answer they have memorized.
- Can this be attributed to that the larger model memorizes a fact better?

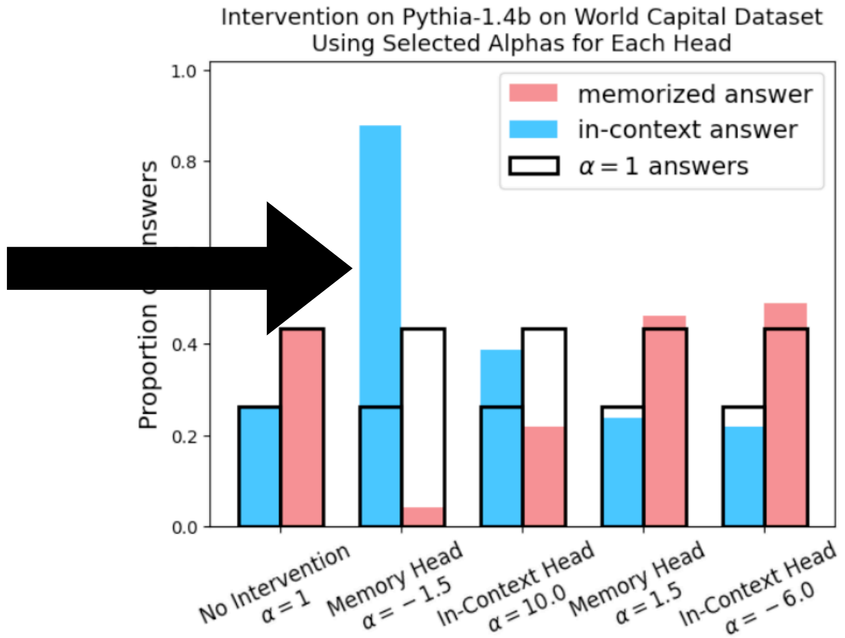


Characterizing Mechanisms for Factual Recall in Language Models, Yu etal., NAACL 2024

Deeper analysis of knowledge conflicts

- Memory head vs. context head?





Characterizing Mechanisms for Factual Recall in Language Models, Yu etal., NAACL 2024

RQ2: How to mitigate the effect of knowledge conflict?

- A similar question: *How to mitigate knowledge conflict?*
  - Update the parametric knowledge: continual learning and knowledge editing.
- **Less relevant to our survey**
- RQ2: Once the **conflict exists**, how to mitigate the (negative) effect of knowledge conflict?
- Performance & Benchmarks are omitted since *literally no two papers used exactly one dataset*!!

Priors

- **Contextual knowledge is <u>always</u> correct**
  - Close to 90%+ of the cases
  - When we use RAG to ``update" the knowledge, the updated knowledge should be correct
  - Not following context ~ Hallucination?
- Contextual knowledge is not correct
  - Misinformation & RAG attacks
- No prior
  - Provide disentangled answers

Solution type I: fine-tuning

- If the context is **relevant** to the query
  - follow the context
- else
  - ignore the context
- Balance robustness and controllability

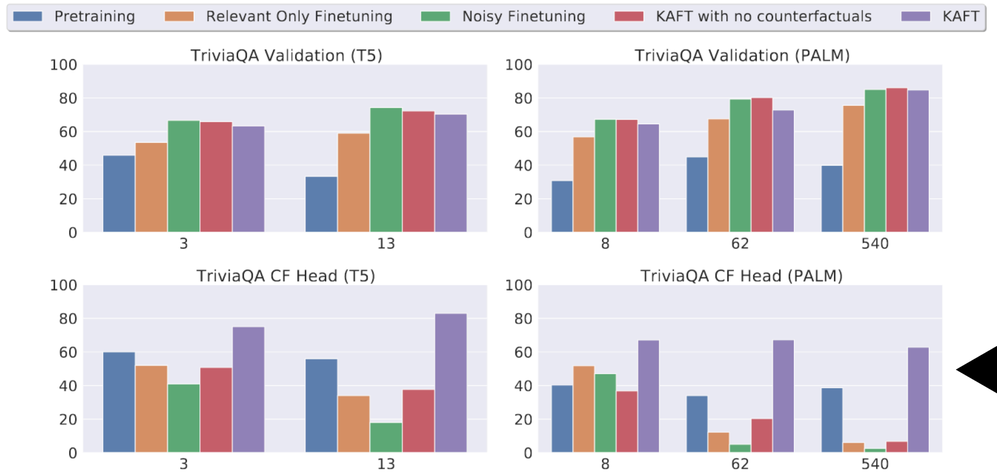
	Controllability	Robustness
Question	Dave Gilmour and Roger Waters were in which rock group?	How has **British art** survived in Normandy?
Context	George Roger Waters (born 6 September 1943) is an English singer, . . . Later that year, he re-united with **The Rolling Stones** bandmates Mason, Wright and David Gilmour for the Live 8 global awareness event; it was the group's first appearance with Waters since 1981. . .	In Britain, **Norman art** primarily survives as stonework or metalwork, such as capitals and baptismal fonts. In southern Italy, however, Norman artwork survives plentifully in forms strongly influenced by its Greek, Lombard, and Arab forebears. Of the royal regalia pre-served in Palermo, the crown is Byzantine. . .
KAFT (ours)	The Rolling Stones (from context).	In museums (irrelevant context).
Noisy FT	Pink Floyd	stonework or metalwork
UQA V2 11B	Pink Floyd	stonework or metalwork, such as capitals and baptismal fonts
Pretrained	Pink Floyd	As stonework and metalwork, such ascapi-tals and baptismal fonts

Large Language Models with Controllable Working Memory, Li etal., Findings of ACL 2023

Solution type I: fine-tuning

- KAFT (Knowledge Aware Fine-Tuning)
  - similar to conterfactual training, but different…
  - for robustness: if the context is irrelevant, follow the memory, not the ground truth.
- Counterfactual != counter memory

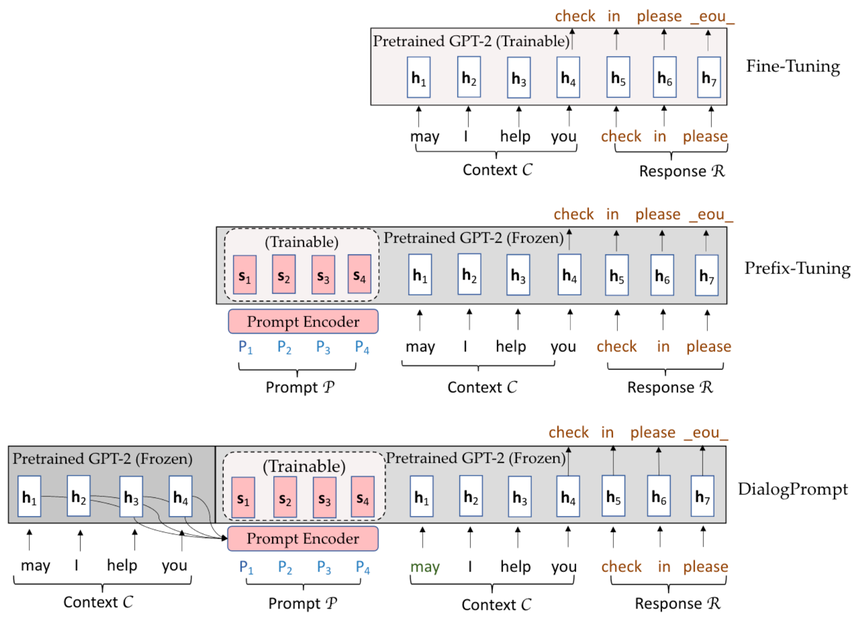
Context type	Target sequence
relevant context	${ground truth answer} (from context)
irrelevant context	${pretrained model's answer} (irrelevant context)
empty context	${pretrained model's answer} (empty context)
counterfactual context	${counterfactual answer} (from context)



Large Language Models with Controllable Working Memory, Li etal., Findings of ACL 2023

Solution type II: prompting

- Prompting: prompt engineering, not prompt learning



Base prompt

$\{c\}$ Q: $\{q\}$? Options: $\{o\}$ A:

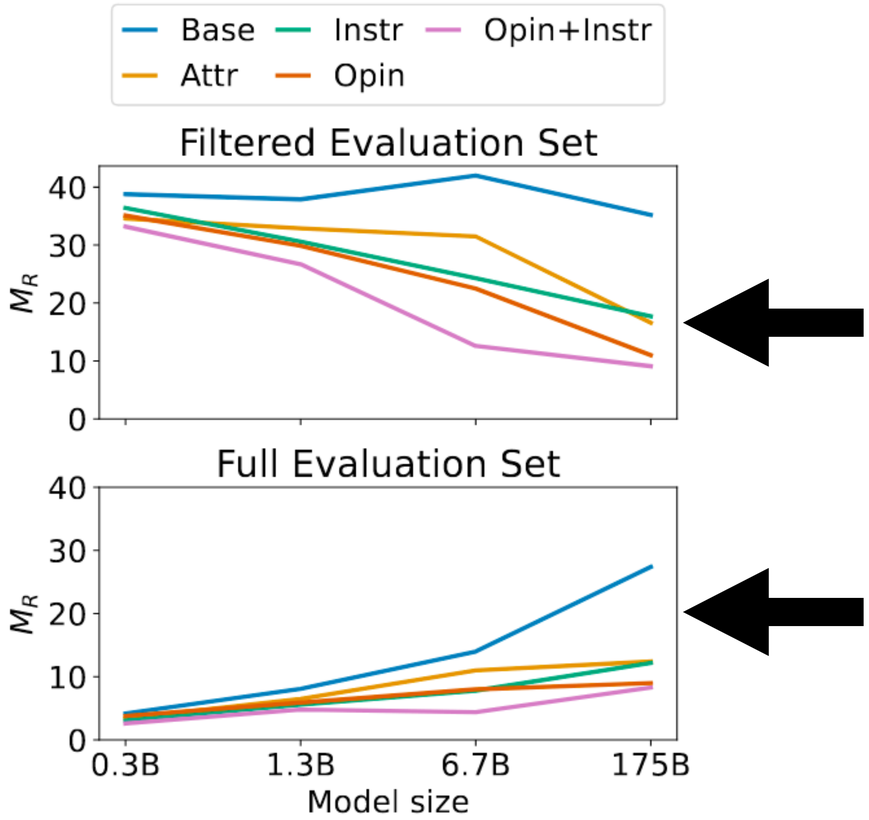
Opinion-based prompt

Bob said, "$\{c\}$" Q: $\{q\}$ in Bob's opinion? Options: $\{o\}$ A:

A simple prompt is enough!

Left: Response Generation with Context-Aware Prompt Learning, Gu etal., arXiv 2021
Right: Context-faithful Prompting for Large Language Models, Zhou etal., Findings of ACL 2023

Solution type II: prompting



Filtered dataset: they already have knowledge on predicting the original answer (correct/ non-counterfactual)

Full dataset: unfiltered?

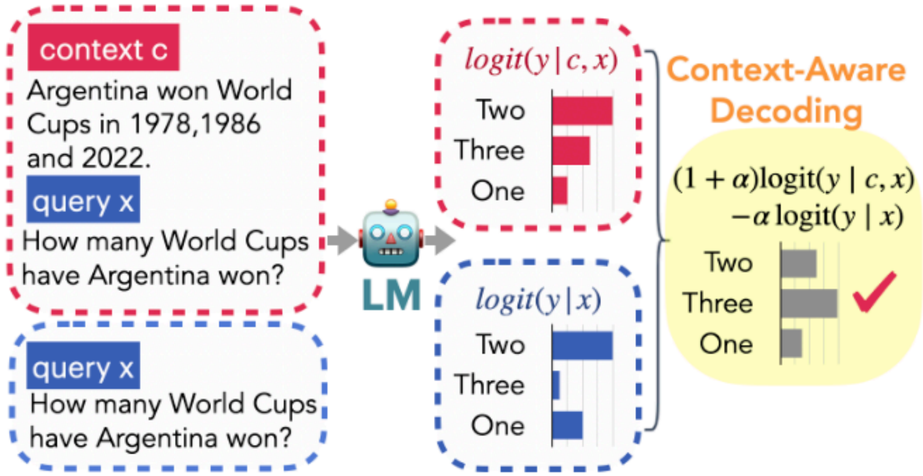
While larger LLMs are better at updating memorized answers, they still tend to have more memorization due to the larger number of memorized answers

Context-faithful Prompting for Large Language Models, Zhou etal., Findings of ACL 2023

Solution type III: decoding

- The distribution of the raw query is suppressed, while the distribution of the context+query is enhanced!

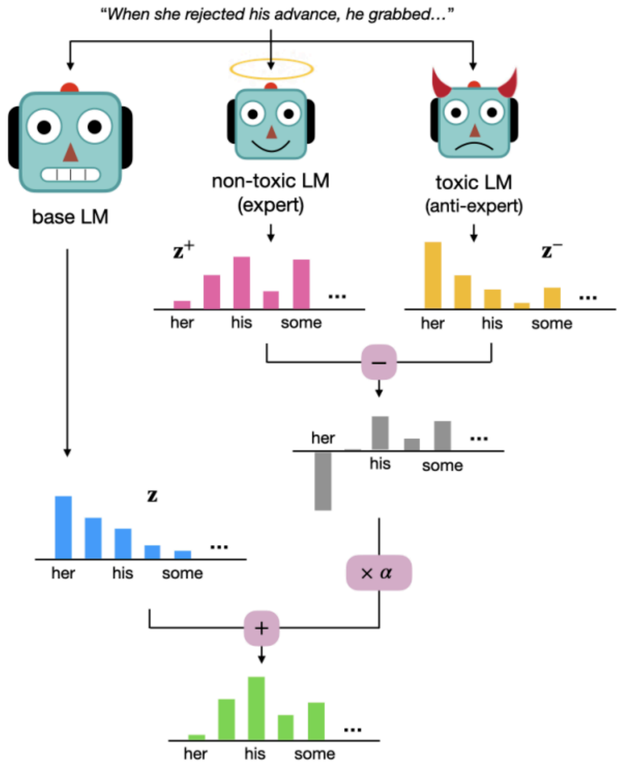
$$y_t \sim \text{softmax}[(1+\alpha)\,\text{logit}_\theta(y_t \mid \boldsymbol{c}, \boldsymbol{x}, \boldsymbol{y}_{<t}) - \alpha\,\text{logit}_\theta(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t})]$$

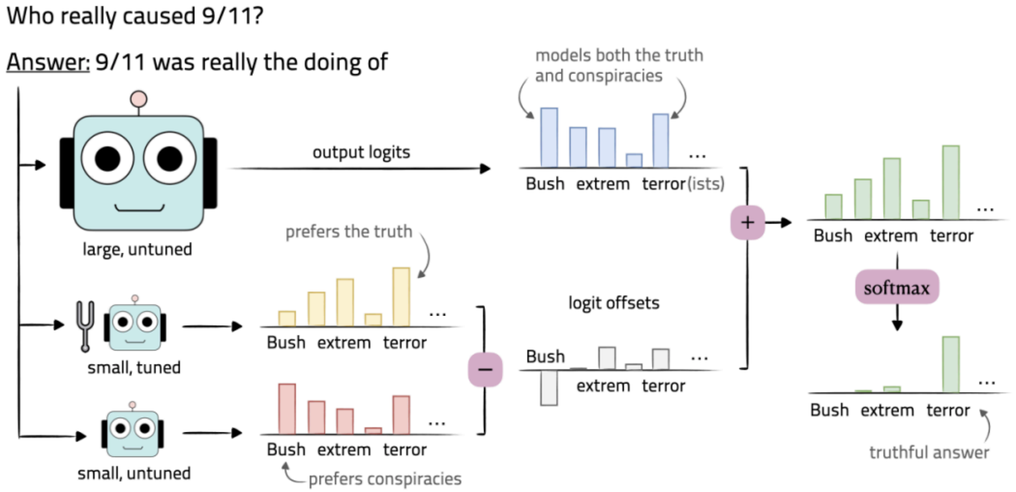


Trusting Your Evidence: Hallucinate Less with Context-aware Decoding, Shi etal., NAACL 2024

Solution type III: decoding

- Decoding is everywhere! (Seems only true for academia)

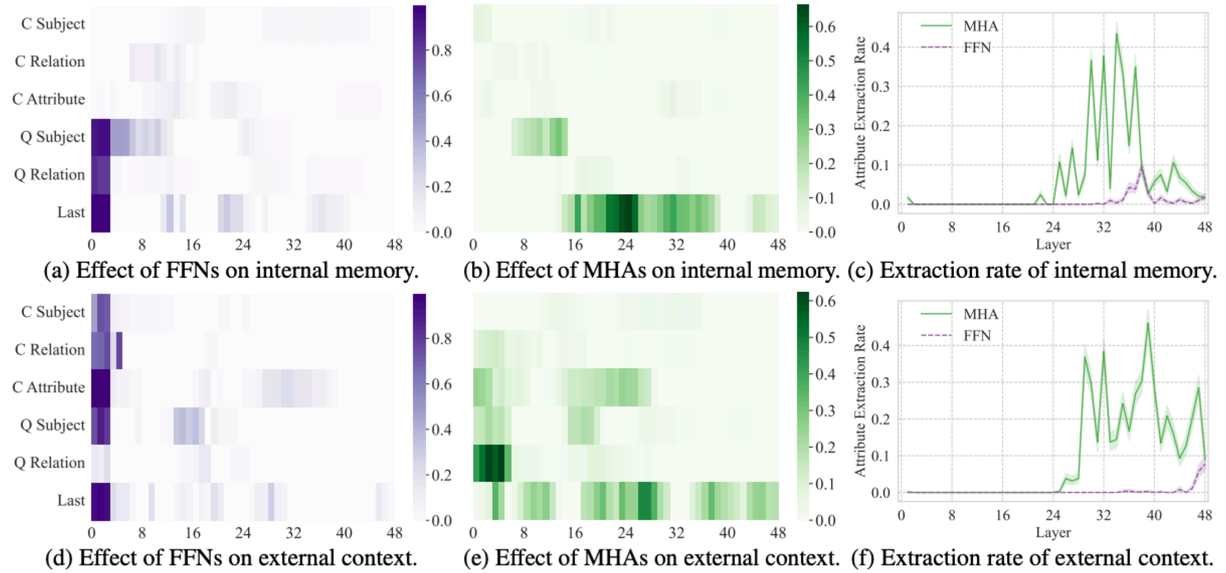




Left: DEXPERTS: Decoding-Time Controlled Text Generation with Experts and Anti-Experts, Liu etal., ACL 2021
Right: Tuning Language Model by Proxy, Liu etal., arXiv 2024

Solution type IV: manipulating internal stuff

- Memory vs. context: MHA and FFNs.
- Problem to solve: How to effectively locate them? some heuristic..



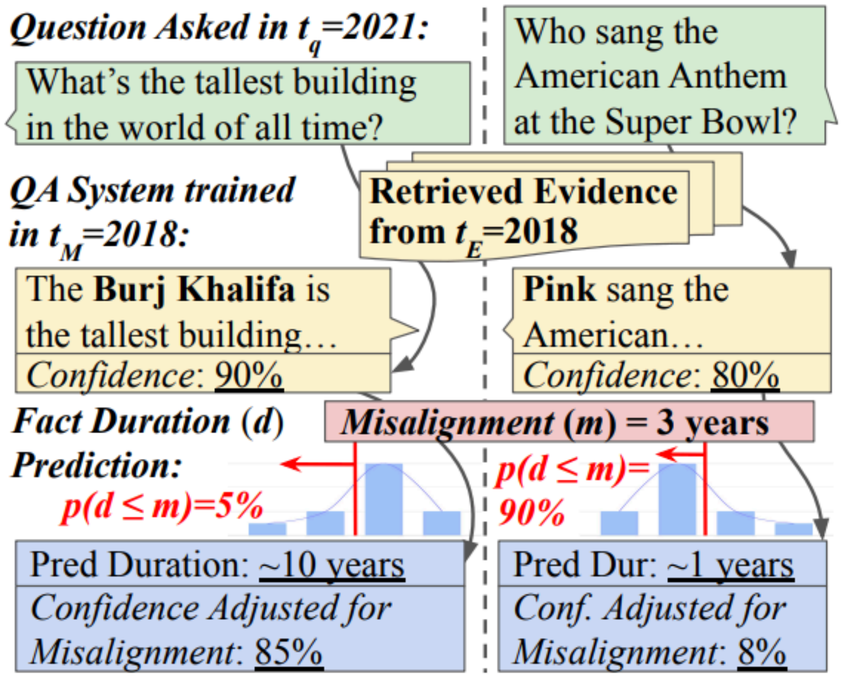
(a) Effect of FFNs on internal memory. (b) Effect of MHAs on internal memory. (c) Extraction rate of internal memory.

(d) Effect of FFNs on external context. (e) Effect of MHAs on external context. (f) Extraction rate of external context.

Cutting Off the Head Ends the Conflict: A Mechanism for Interpreting and Mitigating Knowledge Conflicts in Language Models, Jin. etal., arXiv 2024

Additional: predict the effectiveness of parametric knowledge

- Let the model just **abstain** from presenting facts that we predict are out of date!
- **Fact duration prediction**: the task of predicting how frequently a given fact changes
  - m = t_M – t_q
  - predict d = duration
  - if duration <= m, adjust the confidence!
    - adjust as p(d<=m) as d can be a distribution
- Interesting work and very solid experiments :)



Mitigating Temporal Misalignment by Discarding Outdated Facts, Zhang etal., EMNLP 2023

Additional: provide disentangled answers

- Just predict two answers when conflict happens
- Problem to solve: how to ascertain conflict?

Question: What country shares borders with both Belarus and Romania?

Factual

Context: **Ukraine** borders with seven countries: Poland, Slovakia, Hungary, Romania, Moldova, Russia and Belarus. ...
Contextual Answer: **Ukraine**
Parametric Answer: **Ukraine**

Counterfactual

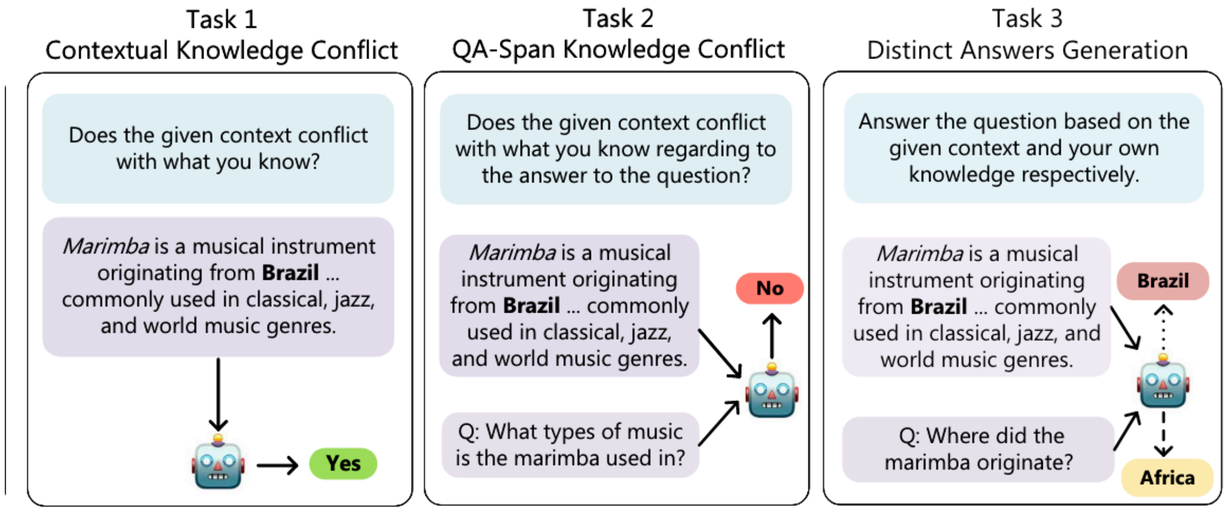
Context: **Brazil** borders with seven countries: Poland, Slovakia, Hungary, Romania, Moldova, Russia and Belarus. ...
Contextual Answer: **Brazil**
Parametric Answer: **Ukraine**

Empty

Context:
Contextual Answer: Unanswerable
Parametric Answer: **Ukraine**

Random

Context: The epic, traditionally ascribed to the Hindu sage Valmiki, narrates the life of Rama, the legendary prince of ...
Contextual Answer: Unanswerable
Parametric Answer: **Ukraine**



It is often the case that not all pieces of information within a passage are in conflict between parametric and conflicting knowledge sources.
It is crucial for LLMs to pinpoint the **specific piece of information** where these conflicts arise.

Left: DisentQA: Disentangling Parametric and Contextual Knowledge with Counterfactual Question Answering, Neeman etal., ACL2023
Right: Resolving Knowledge Conflicts in Large Language Models, Wang etal., arXiv 2023

Remarks: mitigations to knowledge conflicts

- Mitigation strategies are designed based on the priors
- In most circumstances, we trust the contextual knowledge
- Numerous strategies can be employed to prioritize the contextual answers and surpass the memorized knowledge
- Efficiency:
  - Prompt and fine-tuning >> all others

Inter-context conflict

Research questions

- **Less** studied since realistic inter-context conflict is less experienced
  - unlike context-memory conflict, which is more of academia-interest
- RQ1: How to detect conflicts within context?
- RQ2: When conflicts exists, what kind of context wins the model's preference?

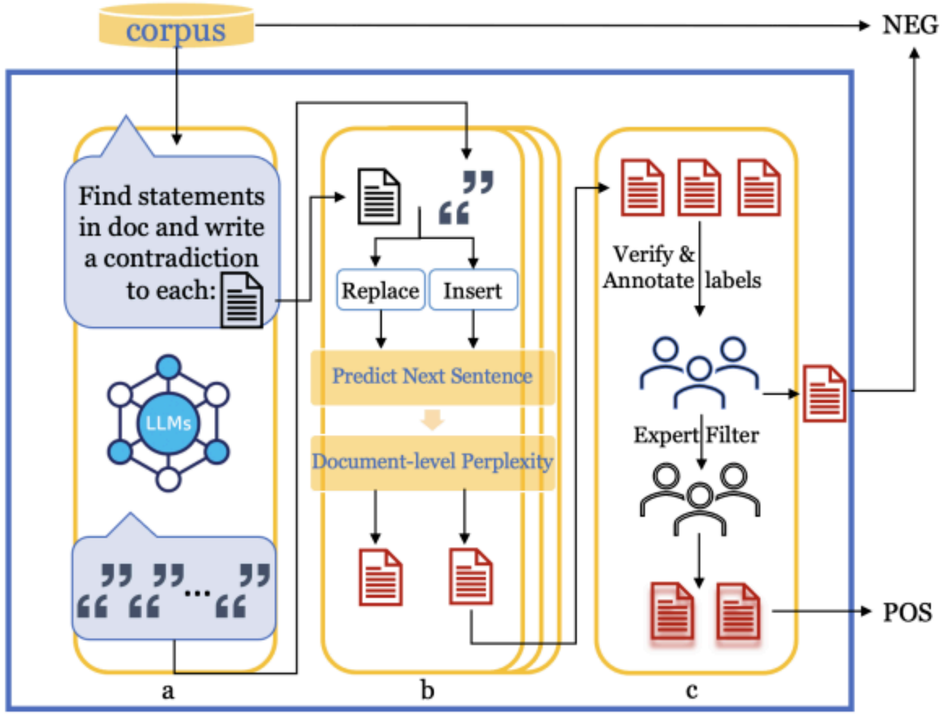
RQ1: How to detect conflicts within context?

Type	Definition	Original Statement	Generated Self-Contradiction
Negation	Negating the original sentence	Zully donated her kidney.	Zully never donated her kidney.
Numeric	Number mismatch or number out of scope.	All the donors are between 20 to 45 years old.	Lisa, who donates her kidney, she is 70 years old.
Content	Changing one/multiple attributes of an event or entity	Zully Broussard donated her kidney to a stranger.	Zully Broussard donated her kidney to her close friend.
Perspective / View / Opinion	Inconsistency in one's attitude/ perspective/opinion	The doctor spoke highly of the project and called it "a breakthrough"	The doctor disliked the project, saying it had no impact at all.
Emotion / Mood / Feeling	Inconsistency in one's attitude/ emotion/mood	The rescue team searched for the boy worriedly.	The rescue team searched for the boy happily.
Relation	Description of two mutually exclusive relations between entities.	Jane and Tom are a married couple.	Jane is Tom's sister.
Factual	Need external world knowledge to confirm the contradiction.	The road T51 was located in New York.	The road T51 was located in California.
Causal	The effect does not match the cause.	I slam the door.	After I do that, the door opens.

ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models, Li etal., NAACL 2024

RQ1: How to detect conflicts within context?

- Dataset creation: creating contradiction inside a document
- Find — rewrite — replace/insert
- human verification



ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models, Li etal., NAACL 2024

RQ1: How to detect conflicts within context?

- Tasks of detecting contradictory
- Task1: Binary judge if a conflict exists
- Task2: Given a document with a self-contradiction, we ask the model to select the five most probable sentences that indicate the self-contradiction and rank them from high to low probability
- GPT4 performs the best overall

Model	Accuracy	Precision	Recall	F1
GPT3.5	50.1%	100.0%	0.2%	0.4 %
GPT4	53.8%	97.0%	8.0%	15.6%
PaLM2	52.0%	61.0%	13.4%	22.0%
LLaMAv2	50.5%	51.0%	38.3%	43.7%

Model	EHR ↑	Avg. Index (1-5) ↓
GPT3.5	42.8%	1.98
GPT4	**70.2%**	**1.79**
PaLM2	48.2%	2.36
LLaMAv2	20.4%	2.28

ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models, Li etal., NAACL 2024

RQ1: How to detect conflicts within context?

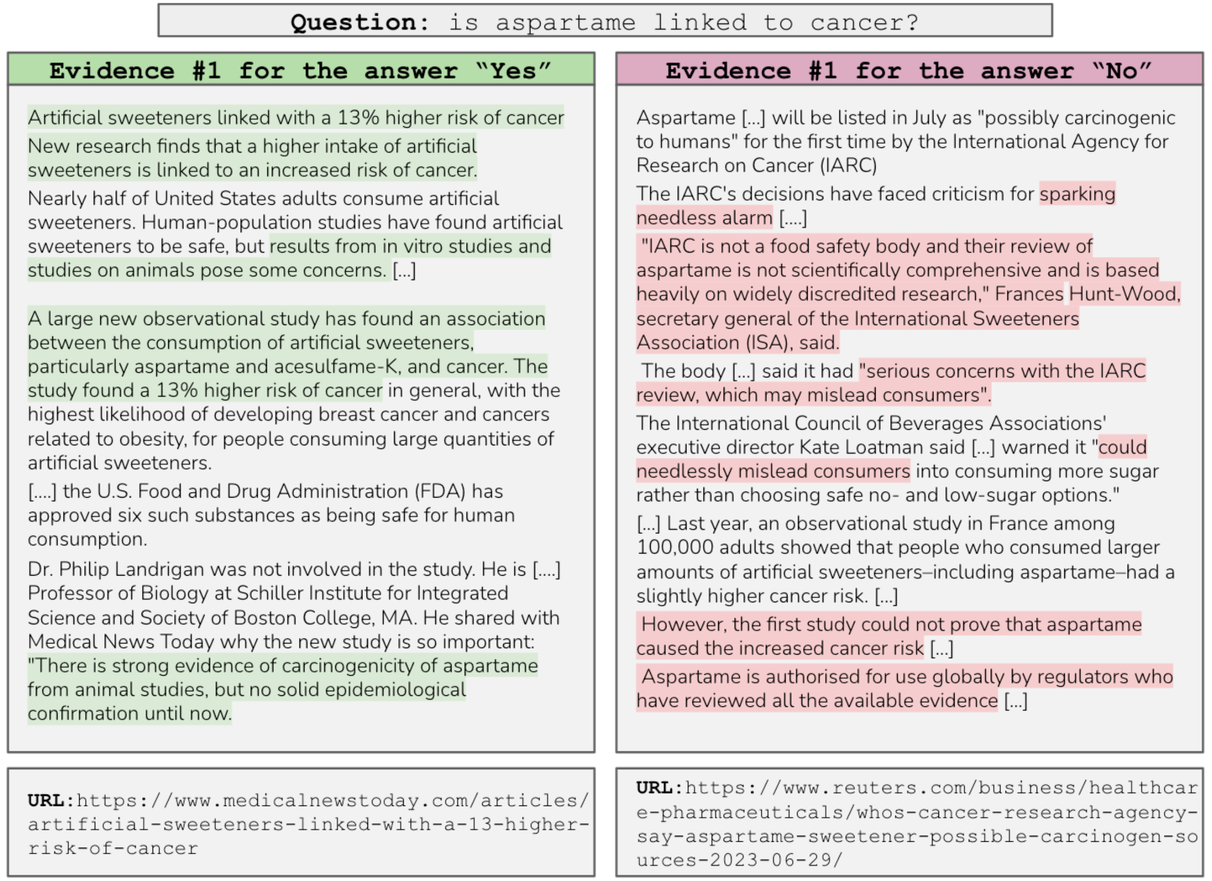
- Task3: Judge-then-Find (attribute)
- For the binary judgment task, If the answer is Yes, the model also needs to provide supporting evidence by quoting sentences that can indicate the self-contradiction

Models	Precision	Recall	F1 Score	TP rate	FP rate	TN rate	FN rate	Evidence Hit Rate	R-acc(pos)
GPT3.5	57.0%	62.0%	41.0%	20.6%	12.8%	36.9%	29.7%	41.0%	16.8%
GPT4	88.0%	39.0%	54.0%	19.6%	2.7%	46.2%	31.5%	92.7%	35.6%
PaLM2	52.0%	83.0%	64.0%	41.5%	37.6%	12.0%	9.0%	41.0%	33.7%
LLaMAv2	50.0%	95.0%	65.0%	48.0%	48.6%	1.12%	2.3%	14.5%	13.8%

ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models, Li etal., NAACL 2024

RQ2: What kind of context wins the model's preference?

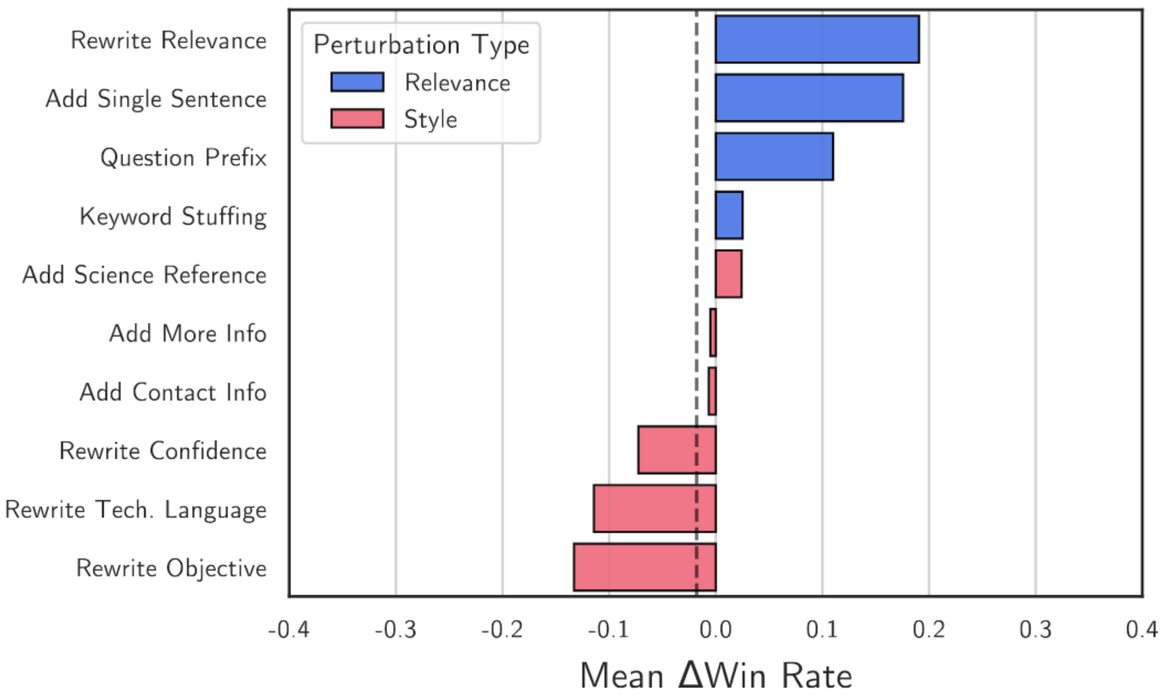
- The question itself has a non-fixed answer
- The evidence documents are conflicting
- Assess what kind of RAG documents the LLM prefer



What Evidence Do Language Models Find Convincing?, Wan., NAACL 2024

RQ2: What kind of context wins the model's preference?

- LLMs rely on relevancy, not stylistic features
  - add more info, and rewrite tech. language
- Stylistic changes—inspired by factors that influence humans—have a neutral or even negative effect on models.



What Evidence Do Language Models Find Convincing?, Wan., NAACL 2024

RQ2: What kind of context wins the model's preference?

- Search Engine Optimization (SEO)
  - What kind of RAG docs is more like to convince LLMs?
- Retrieval
  - What kind of documents is more like to be retrieved
- Retrieved && **preferred docs** will most influence the RAG LLMs!
  - Conclusion: LLMs tend to over-index on relevancy
    - simply increase amount of n-gram overlap between the question and the doc

Intra-memory conflict

Research questions

- RQ1: How LLMs with intra-memory conflict will behave?
- RQ2: Why do LLMs exhibit self-contradiction?
- RQ3: How to mitigate intra-memory conflict?

RQ1: How LLMs with intra-memory conflict will behave?

- Self-inconsistency
- Cross-lingual inconsistency

Self-inconsistency

Model	Accuracy	Consistency	Consistent-Acc
majority	23.1±21.0	100.0±0.0	23.1±21.0
BERT-base	45.8±25.6	58.5±24.2	27.0±23.8
BERT-large	48.1±26.1	**61.1**±23.0	**29.5**±26.6
BERT-large-wwm	**48.7**±25.0	60.9±24.2	29.3±26.9
RoBERTa-base	39.0±22.8	52.1±17.8	16.4±16.4
RoBERTa-large	43.2±24.7	56.3±20.4	22.5±21.1
ALBERT-base	29.8±22.8	49.8±20.1	16.7±20.3
ALBERT-xxlarge	41.7±24.9	52.1±22.4	23.8±24.8

- When confronted with inputs that have the same semantics but different forms of expression, the model will exhibit inconsistent outputs

#	Subject	Object	Pattern #1	Pattern #2	Pattern #3	Pred #1	Pred #2	Pred #3
1	Adriaan Pauw	Amsterdam	[X] was born in [Y].	[X] is native to [Y].	[X] is a [Y]-born person.	Amsterdam	Madagascar	Luxembourg
2	Nissan Livina Geniss	Nissan	[X] is produced by [Y].	[X] is created by [Y].	[X], created by [Y].	Nissan	Renault	Renault
3	Albania	Serbia	[X] shares border with [Y].	[Y] borders with [X].	[Y] shares the border with [X]	Greece	Turkey	Kosovo
4	iCloud	Apple	[X] is developed by [Y].	[X], created by [Y].	[X] was created by [Y]	Microsoft	Google	Sony
5	Yahoo! Messenger	Yahoo	[X], a product created by [Y]	[X], a product developed by [Y]	[Y], that developed [X]	Microsoft	Microsoft	Microsoft
6	Wales	Cardiff	The capital of [X] is [Y] .	[X]'s capital, [Y].	[X]'s capital city, [Y].	Cardiff	Cardiff	Cardiff

Measuring and Improving Consistency in Pretrained Language Models, Elazar etal., ACL 2021

Self-inconsistency

- There is an inconsistency between generating and validating an answer in LLMs

Plan Arithmetic

Generator Prompt:
Consider the identity: 4*19+3*11 = 109
Can you modify exactly one integer (and not more than that!) on the left hand side of the equation so the right hand side equals *(not equals)* 52 ?
Answer: 4*7+3*11

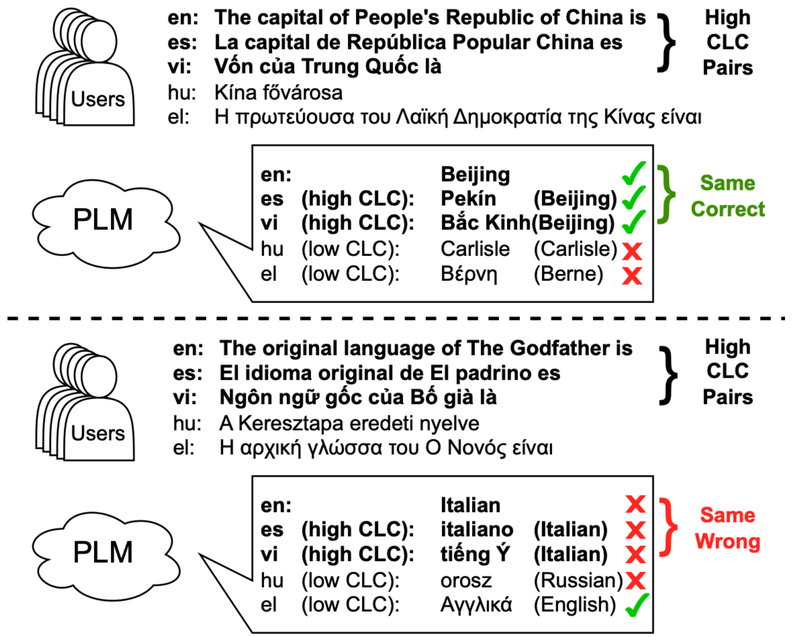
Validator Prompt:
Check whether the following computation is correct.
4*7+3*11 = 52
The computation is (True/False): False

	Arithmetic	PlanArith	PriorityPrompt	QA	Style	HarmfulQ	Average
GPT-3.5	67.7	**66.0**	**79.6**	89.6	92.6	-	79.1
GPT-4	75.6	62.0	52.0	**95.3**	**94.3**	-	75.8
davinci-003	**84.4**	60.0	68.0	86.9	85.7	-	77.0
Alpaca-30B	53.9	50.2	49.0	79.9	74.6	51.6	59.9

BENCHMARKING AND IMPROVING GENERATOR-VALIDATOR CONSISTENCY OF LMS, Li etal., arXiv 2023

Cross-lingual inconsistency

- When the same question is asked in different languages, LLMs may give different answers



Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models, Qi etal., EMNLP 2023

RQ2: Why do LLMs exhibit self-contradiction?

- Inconsistency in training corpora
- Decoding strategy
- Knowledge editing

Inconsistency in Training Corpora

- Misinformation
- Outdated information

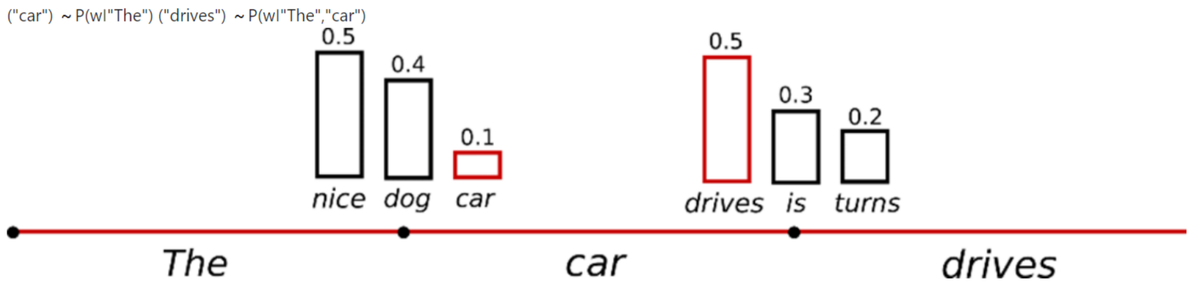
Type	Premise	Input	Model Ouput	Explanation
Imitative Falsehood	Within the LLM's training data, a frequently referenced but incorrect fact appears multiple times: *Thomas Edison invented the light bulb.*	Who is credited with the invention of the light bulb?	Thomas Edison is credited with the invention of the light bulb.	**While Edison improved the light bulb's design, he was not its sole inventor.** The model's answer reflects a misconception prevalent in its training data, exemplifying an imitative falsehood.
Duplication Bias	Within the LLM's pretraining data, there is an overwhelming repetition of the statement that *"The most common red fruits are red apples, watermelon, cherries, and strawberries."*	Please list some red fruits, excluding apples.	Red fruits are red apples, watermelon, cherries, and strawberries.	**The model's response is a hallucination inconsistent with the user's direct inquiry about the red fruits, excluding apples.** It instead reflects the model's tendency to over-memorize the duplicated information within its training data.
Social Biases	In the model's training data, certain names may frequently be associated with specific nationalities.	Provide a summary of the following news: Dr. Kim recently published a breakthrough paper on quantum computing. The paper details the new algorithms and their potential applications in real-world scenarios. It has been hailed as revolutionary by experts in the field.	Dr. Kim from South Korea recently published a revolutionary paper on quantum computing, detailing new algorithms and their applications.	**The model erroneously added "South Korea" based on the name "Kim"** revealing its bias in linking certain names to specific nationalities, which can cause hallucinations in the summary.

A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, Huang etal., arXiv 2023

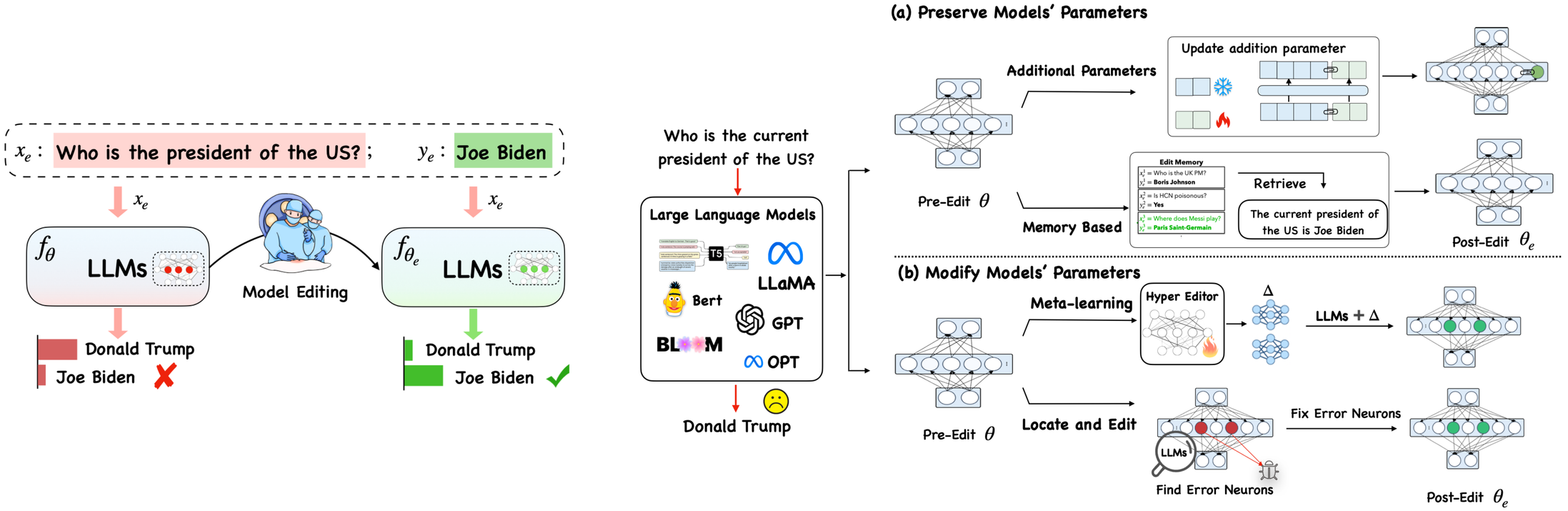
Decoding Strategy

- Top-p decoding strategy
- Top-k decoding strategy

$$w_t \sim P(w|w_{1:t-1})$$



Knowledge Editing



Editing Large Language Models: Problems, Methods, and Opportunities, Yao etal., EMNLP 2023

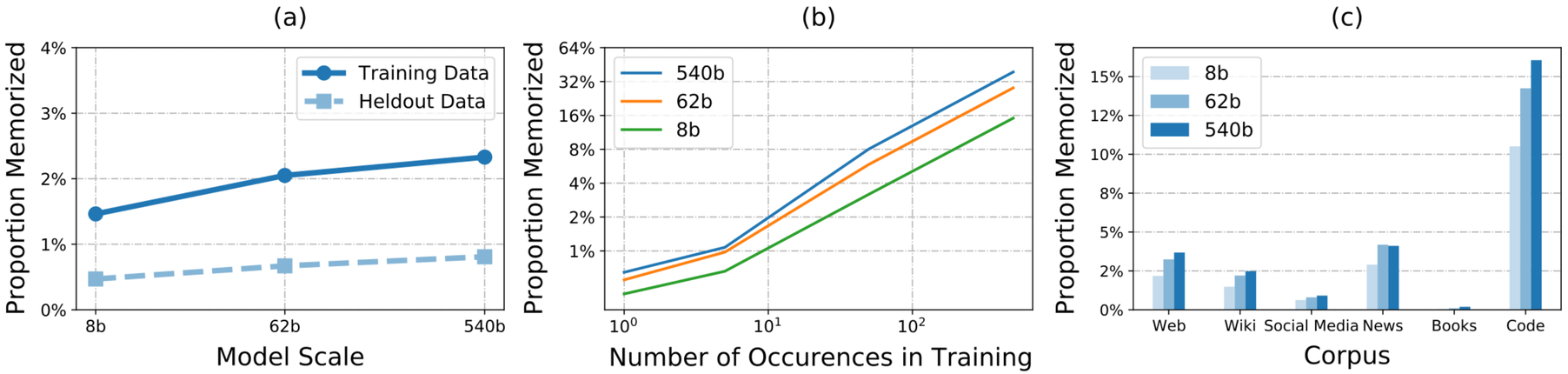
RQ2: Why do LLMs exhibit self-contradiction?

- Inconsistency in training corpora is the fundamental factor
- Decoding strategy indirectly contributes to exacerbating the conflict
- Knowledge editing can inadvertently introduce conflicting information

Remarks: Intra-memory conflict v.s. Hallucination

Memorization

- The larger the model scale, the more the model tends to memorize the training data
- Shift LLMs from generalization to memorization (nurse-female)

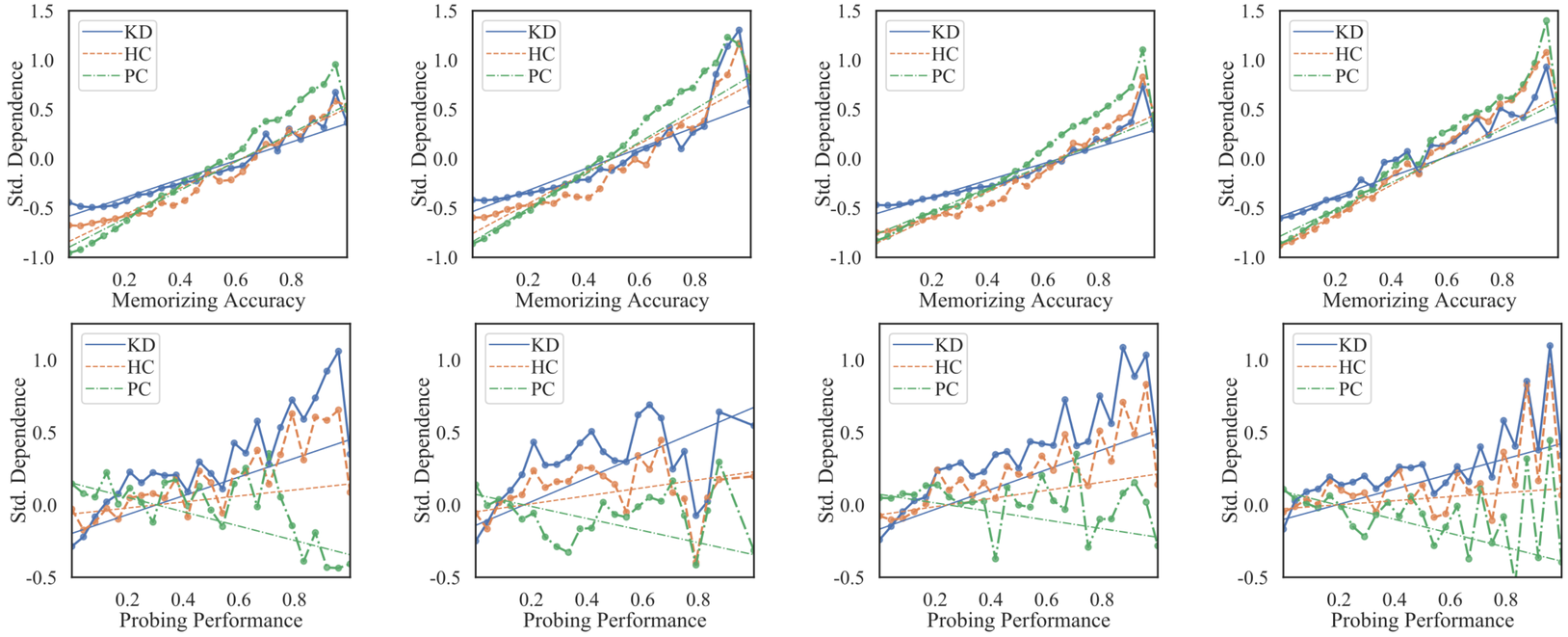


PaLM: Scaling Language Modeling with Pathways, Chowdhery etal., JMLR 2023

Remarks: Intra-memory conflict v.s. Hallucination

Knowledge Shortcut

- PLMs generate the missing factual words more by the positionally close and highly co-occurred words than the knowledge-dependent words



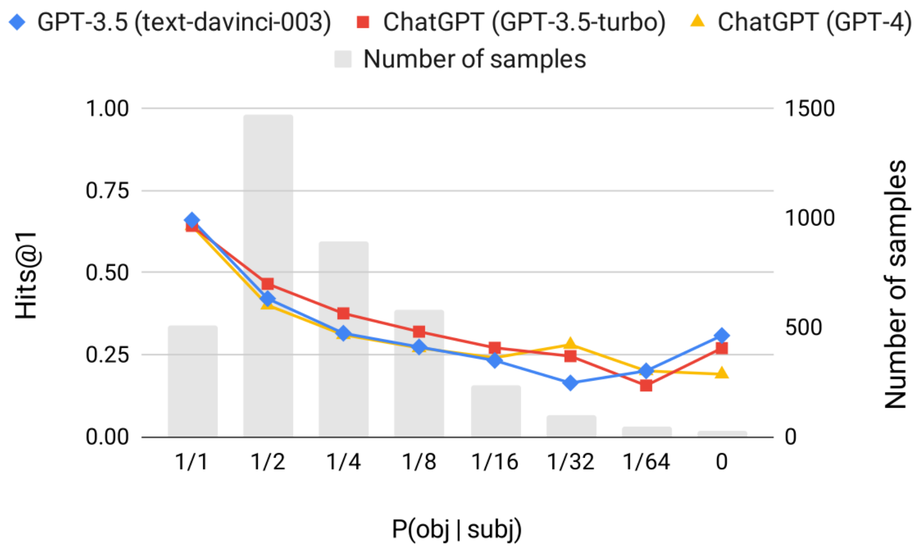
(a) BERT-*large-cased-wwm* (b) RoBERTa-*large* (c) SpanBERT-*large* (d) ALBERT-*xxlarge-v2*

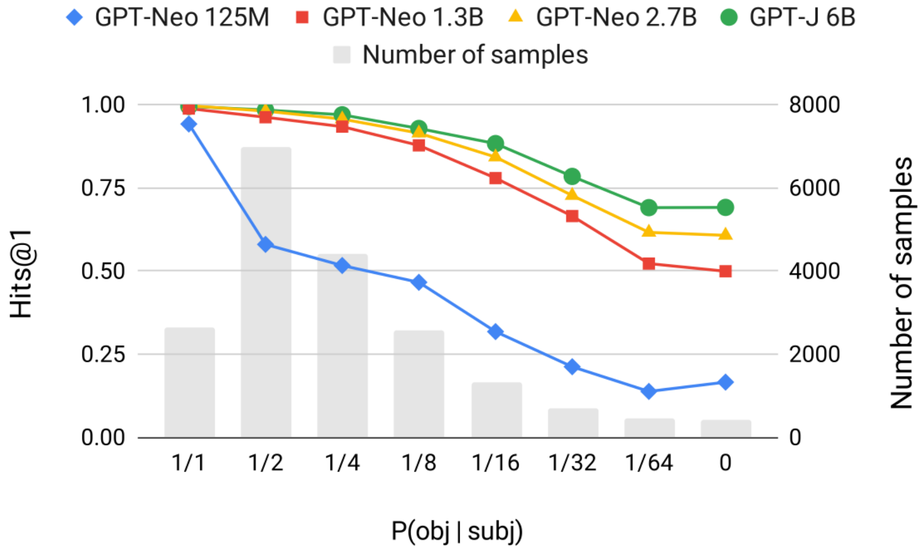
How Pre-trained Language Models Capture Factual Knowledge? A Causal-Inspired Analysis, Li etal., ACL 2022

Remarks: Intra-memory conflict v.s. Hallucination

Knowledge Shortcut

- LLMs are vulnerable to the co-occurrence bias
- LLMs struggle to recall facts whose subject and object rarely co-occur in the pre-training dataset although they are seen during finetuning





Impact of Co-occurrence on Factual Knowledge of Large Language Models, Kang and Choi., JMLR 2023

Remarks: Intra-memory conflict v.s. Hallucination

Decoding strategy

- Stochastic sampling methods like top-p decoding cause higher generation diversity and less repetition, while also being more likely to generate unrealistic answers

Size	Decode	Factual Prompt				Nonfactual Prompt			
		$NE_{ER}\downarrow$	$Entail_R\uparrow$	Div.↑	Rep.↓	$NE_{ER}\downarrow$	$Entail_R\uparrow$	Div.↑	Rep.↓
126M	p=0.9	63.69%	0.94%	0.90	0.58%	67.71%	0.76%	0.90	0.38%
	greedy	48.55%	8.36%	0.03	59.06%	54.24%	6.25%	0.03	59.90%
357M	p=0.9	56.70%	2.01%	0.87	0.55%	60.80%	1.42%	0.88	0.35%
	greedy	43.04%	14.25%	0.03	45.18%	46.79%	9.89%	0.04	46.30%
1.3B	p=0.9	52.42%	2.93%	0.88	0.24%	56.82%	2.04%	0.89	0.25%
	greedy	39.87%	12.91%	0.05	33.13%	45.02%	8.75%	0.05	36.20%
8.3B	p=0.9	40.59%	7.07%	0.90	0.11%	47.49%	3.57%	0.91	0.08%
	greedy	28.06%	22.80%	0.07	19.41%	32.29%	15.01%	0.07	13.26%
530B	p=0.9	33.30%	11.80%	0.90	0.13%	40.49%	7.25%	0.92	0.08%
	greedy	**20.85%**	**31.94%**	0.08	15.88%	27.95%	19.91%	0.08	16.28%

Factuality Enhanced Language Models for Open-Ended Text Generation, Lee et al., NeurIPS 2022

RQ3: How to mitigate intra-memory conflict?

- Improving consistency
- Improving factuality

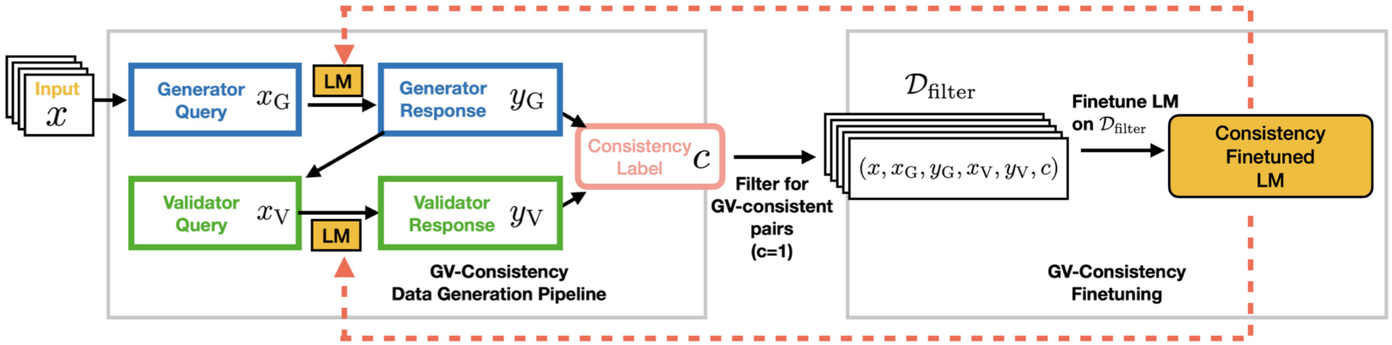
Improving consistency

Fine-tuning

Arithmetic

Generator Prompt:
Write a correct and an incorrect answer (delimited by ||) to the question:
Q: What is 89541 - 9374?
A: 80167 || 98815

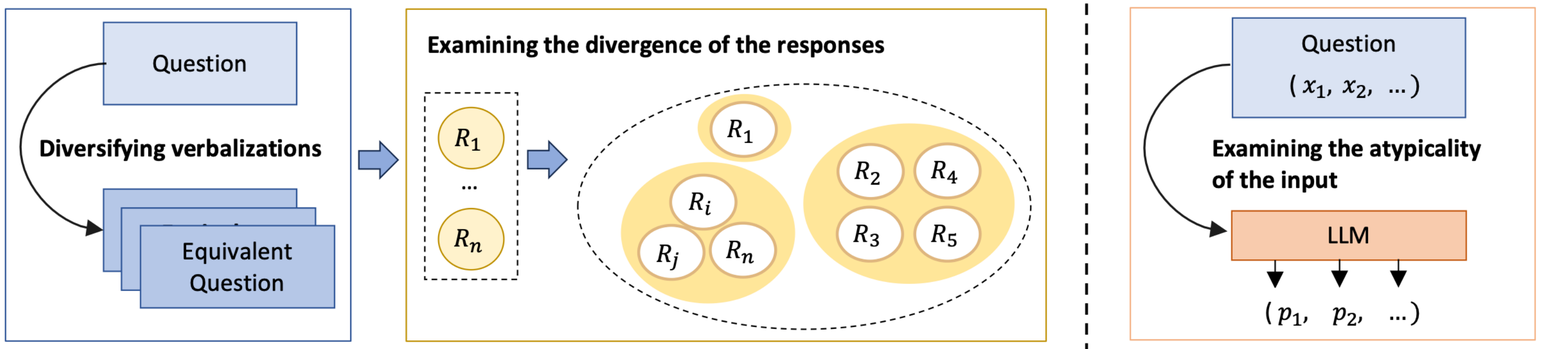
Validator Prompt:
Verify whether the following computation is correct.
Q: What is 89541 - 9374?
A: 80167
The computation is (True/False): True



BENCHMARKING AND IMPROVING GENERATOR-VALIDATOR CONSISTENCY OF LMS, Li etal., arXiv 2023

Improving consistency

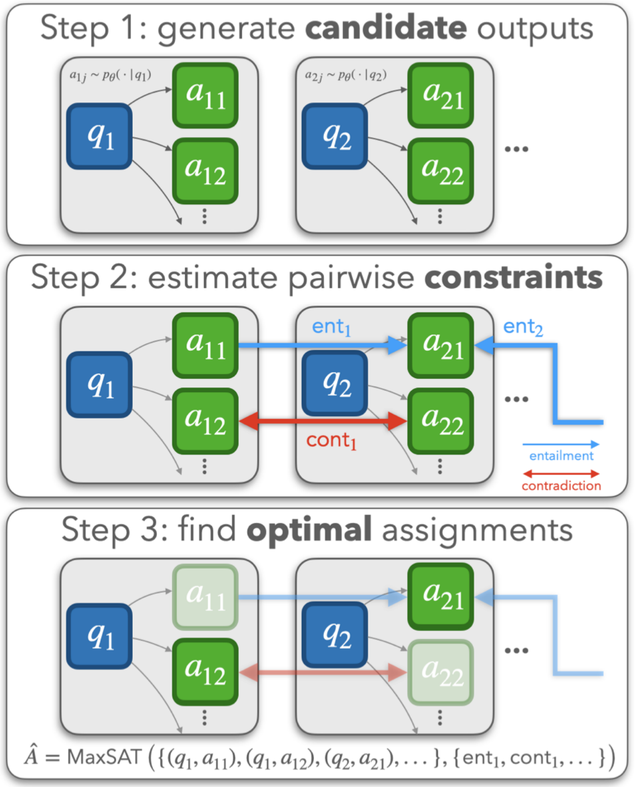
Output Ensemble



(a) Consistency-based Component

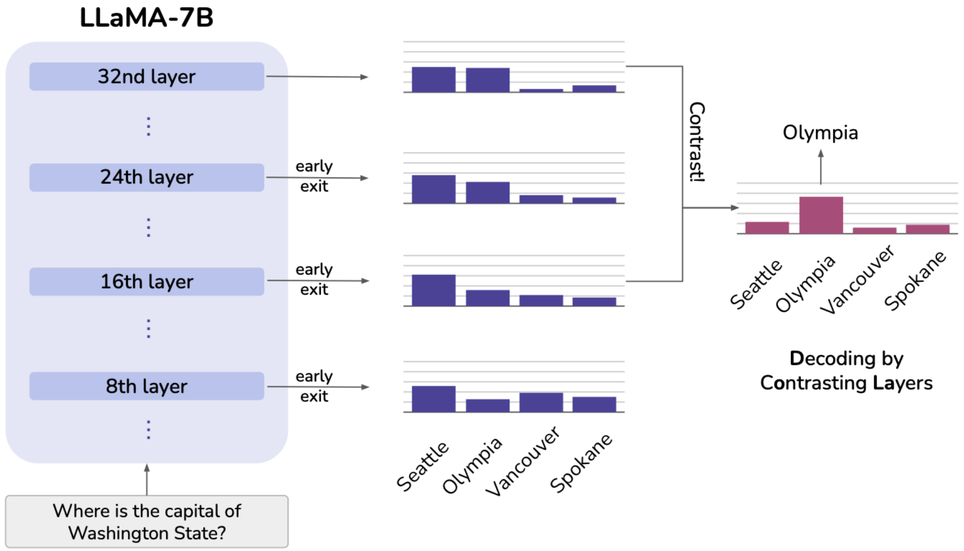
(b) Verbalization-based Component

Knowing What LLMs DO NOT Know: A Simple Yet Effective Self-Detection Method, Zhao etal., NAACL 2024

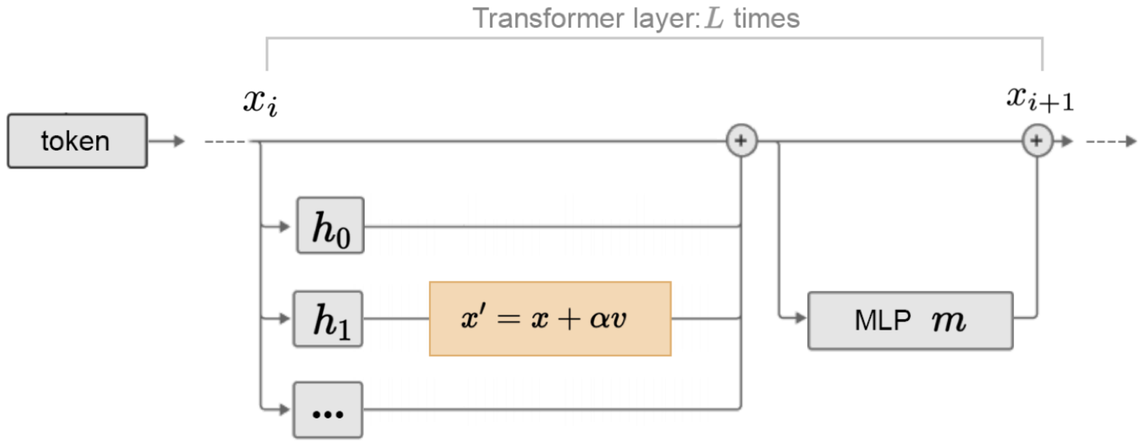


Enhancing Self-Consistency and Performance of Pre-Trained Language Models through Natural Language Inference, Mitchell etal., EMNLP 2022

Improving Factuality



DOLA: DECODING BY CONTRASTING LAYERS IMPROVES FACTUALITY IN LARGE LANGUAGE MODELS, Chuang etal., ICLR 2022



Inference-Time Intervention:Eliciting Truthful Answers from a Language Model, Li etal., NeurIPS 2023

Extended Discussion

Knowledge conflict and Knowledge boundary

- When intra-memory conflict of a knowledge is severe, it means that the model's mastery of that knowledge is weak

Model	Accuracy	Consistency	Consistent-Acc
majority	23.1±21.0	100.0±0.0	23.1±21.0
BERT-base	45.8±25.6	58.5±24.2	27.0±23.8
BERT-large	48.1±26.1	**61.1**±23.0	**29.5**±26.6
BERT-large-wwm	**48.7**±25.0	60.9±24.2	29.3±26.9
RoBERTa-base	39.0±22.8	52.1±17.8	16.4±16.4
RoBERTa-large	43.2±24.7	56.3±20.4	22.5±21.1
ALBERT-base	29.8±22.8	49.8±20.1	16.7±20.3
ALBERT-xxlarge	41.7±24.9	52.1±22.4	23.8±24.8

Measuring and Improving Consistency in Pretrained Language Models, Elazar etal., ACL 2021

Knowledge conflict and Knowledge boundary

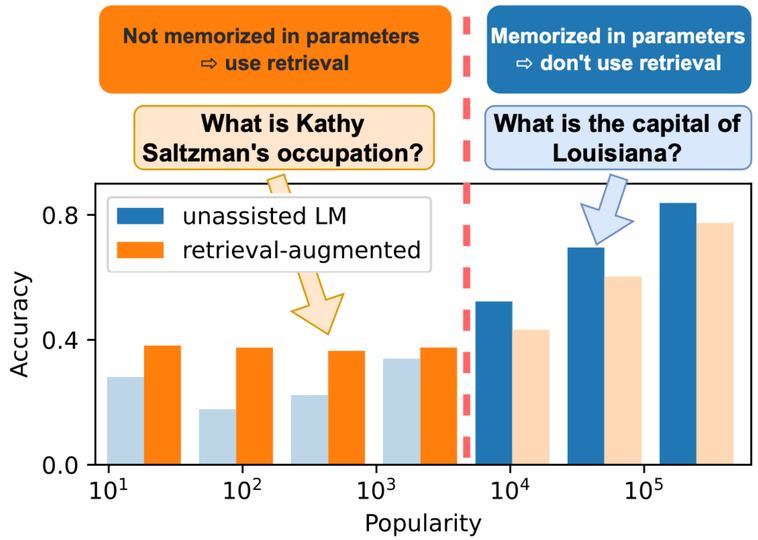
- We want the model to be aware of the boundaries of its knowledge

Dataset	LLM	QA		Priori Judgement			Posteriori Judgement	
		EM	F1	Give-up	Right/G	Right/¬G	Eval-Right	Eval-Acc
NQ	Davinci003	26.37	35.95	27.17%	13.56%	31.15%	71.27%	46.88%
	ChatGPT	30.89	42.14	32.05%	14.63%	38.67%	87.09%	36.85%
TriviaQA	Davinci003	69.56	74.03	5.65%	36.59%	71.53%	87.90%	72.05%
	ChatGPT	74.77	80.11	12.00%	44.00%	78.97%	92.58%	77.02%
HotpotQA	Davinci003	16.62	25.53	35.76%	8.34%	21.23%	69.87%	41.93%
	ChatGPT	17.81	26.35	66.29%	9.76%	33.63%	55.16%	33.13%

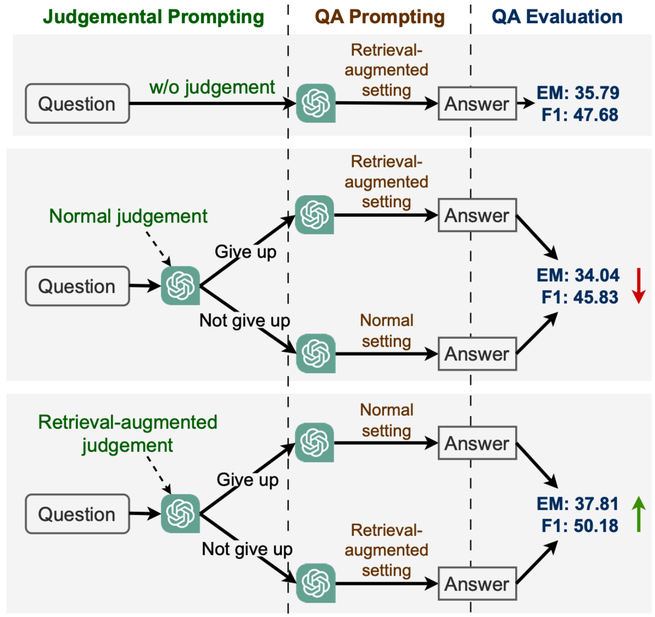
Investigating the Factual Knowledge Boundary of Large Language Models with Retrieval Augmentation, Ren etal., arXiv 2023

Knowledge conflict and Knowledge boundary

- What can be done when the knowledge boundary is exceeded



When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories, Mallen etal., ACL 2023



Investigating the Factual Knowledge Boundary of Large Language Models with Retrieval Augmentation, Ren etal., arXiv 2023

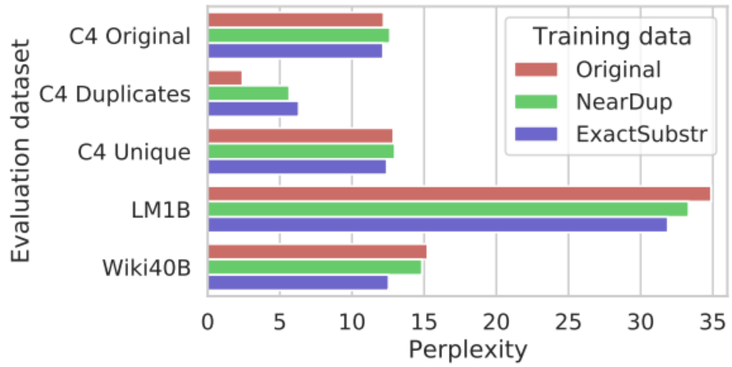
Misinformation

- If the context is **relevant** to the query **and correct**
  - follow the context
- else
  - ignore the context
- Counterfactual learning makes LLMs more susceptible to misinformation?!
  - Yes, and other types of context-faithful methods do!

Knowledge conflict in dataset

- Pretraining dataset is reported to be noisy
  - Duplications
  - Knowlege conflicts?

	% train tokens with dup in train	% train tokens with dup in valid	% valid with dup in train
C4	7.18%	0.75 %	1.38 %
RealNews	19.4 %	2.61 %	3.37 %
LM1B	0.76%	0.016%	0.019%
Wiki40B	2.76%	0.52 %	0.67 %



Deduplicating Training Data Makes Language Models Better, Lee etal., ACL 2022

Challenges and future direction

Challenges and Future Directions

- ❤️ Realistic dataset/evaluation on inter-context conflict for RAG, is it severe or less of a concern?
- Impact on downstream applications, the real consequences of knowledge conflicts in real-world are still under-explored
- ❤️ Interplay between the conflicts, e.g., does intra-memory conflict weaken confirmation bias?

Future directions

- Finer-grained solution, should we conduct classified discussion in developing methods that mitigate knowledge conflict?
- ❤️ New solutions, e.g., MOE that resolve knowledge conflicts
- Multilingual, multimodal (knowledge) conflicts
- ❤️ Interpretable work on intra-memory conflict

Thanks for listening

- arXiv: https://arxiv.org/abs/2403.08319
- GitHub: https://github.com/pillowsofwind/Knowledge-Conflicts-Survey