SIGIR 2025
Padova
ITALY

Undesired Behaviors of LLMs

Yang Deng

Singapore Management University

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Undesired Behaviors of LLMs

- **Factuality Hallucination**
  - Deficiency of Domain-specific Knowledge
  - Outdated Pretrained Knowledge
  - Overconfidence on Unknown Knowledge
- **Untruthful Responses Misled by Contexts**
  - Untruthful Context
  - Irrelevant Context
- **Truthful but Undesired Outputs**
  - Random Responses to Ambiguous Knowledge
  - Biased Responses to Controversial Knowledge

Undesired Behaviors of LLMs

- **Factuality Hallucination**
  - Deficiency of Domain-specific Knowledge
  - Outdated Pretrained Knowledge
  - Overconfidence on Unknown Knowledge
- **Untruthful Responses Misled by Contexts**
  - Untruthful Context
  - Irrelevant Context
- **Truthful but Undesired Outputs**
  - Random Responses to Ambiguous Knowledge
  - Biased Responses to Controversial Knowledge

Factuality Hallucination

Type	Input	Model Ouput	Explanation
Long-tail Knowledge	What causes phenylketonuria?	Phenylketonuria is caused by a mutation in the hemoglobin gene, which results in an unusual interaction with hemoglobin, affecting its ability to transport oxygen in the bloodstream.	The correct cause of phenylketonuria is mutations in the PAH gene. **Such specialized medical information is less frequently encountered in the training corpus, making LLMs struggle to memorize.**
Up-to-date Knowledge	Who will be the Democratic Party candidate for the 2024 US presidential election?	Michelle Obama has announced her candidacy and is leading in the polls, with significant endorsements from key Democratic figures such as Barack Obama and Bernie Sanders. She has also outlined a policy platform focusing on healthcare reform and climate change initiatives.	After President Joe Biden announced his withdrawal from the race on July 21, 2024, he endorsed Kamala Harris as the Democratic Party candidate. **Such hallucinations arise because LLMs' parametric knowledge cannot be updated in real-time, leading to potential fabrication regarding current events.**
Copyright-sensitive Knowledge	Describe the ending of the novel "The Secrets of the Alchemist" that was published last year.	In the dramatic conclusion of "The Secrets of the Alchemist", the protagonist discovers the true secret behind the philosopher's stone, leading to a peaceful resolution with all characters sharing in the wisdom.	**The novel "The Secrets of the Alchemist" is under copyright protection, and LLMs have not been trained directly on such copyrighted materials.** Thus, the model's output fabricates details about the book's ending.

Over-confidence on Unknown Knowledge

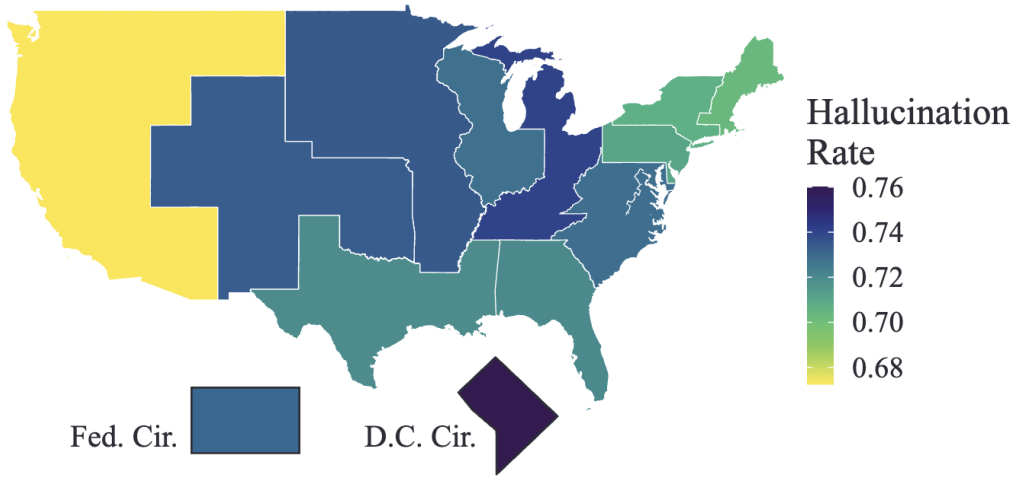
Outdated Pretrained Knowledge

Deficiency of Domain-specific Knowledge

Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions" (TOIS '24)

Deficiency of Domain-specific Knowledge – Legal Domain

Domain	Type of hallucination	Legal example
Closed	Response inconsistency with the prompt	Mischaracterization of an opinion
Open	Response inconsistency with the training corpus	Creative argumentation
	Response inconsistency with the facts of the world	Misstatement of the law

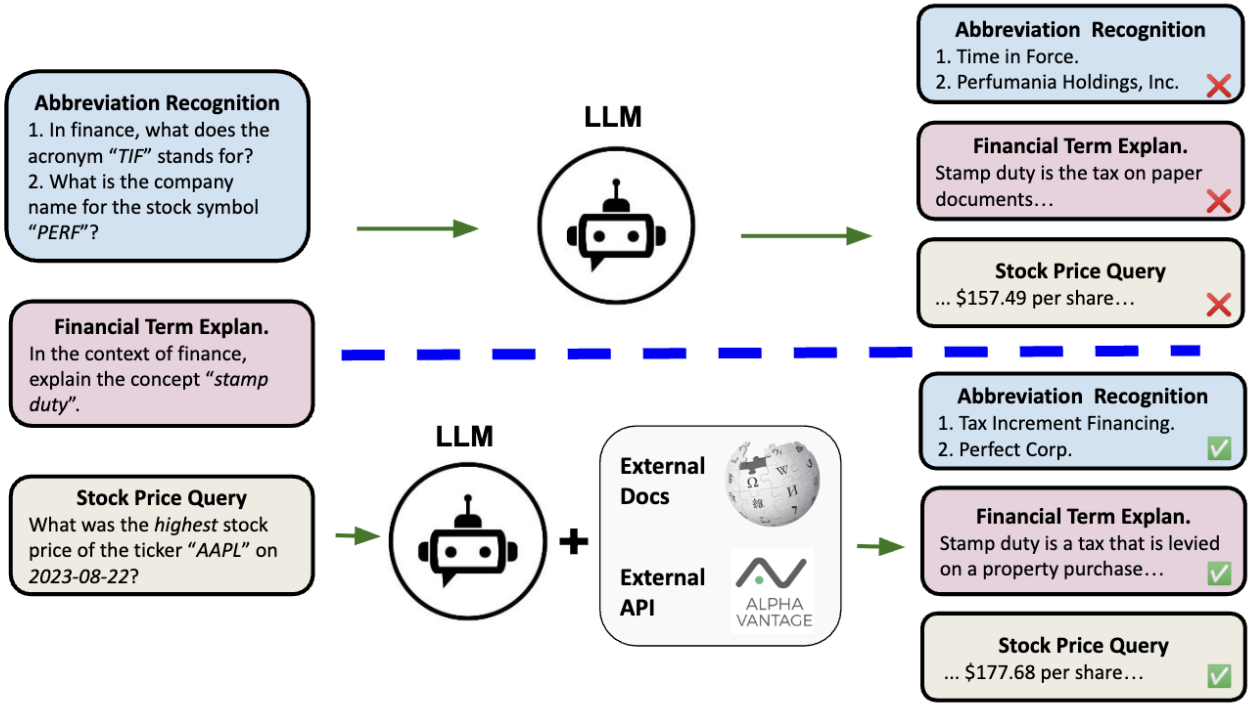


- LLM produces less hallucinations in jurisdictions that are commonly perceived to play a more influential role in the American legal system.

Relationship between USCOA jurisdiction and LLM hallucination rate

Dahl et al., "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models" (Journal of Legal Analysis '24)

Deficiency of Domain-specific Knowledge – Finance Domain



- LLMs generate factually incorrect content in finance.
- LLMs generate seriously unreliable real-world financial predictions.

Kang et al., "Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination" (ICBINB '23)

Deficiency of Domain-specific Knowledge – Medical Domain

Hallucination Category	Description	Example
Misinterpretation of Question	Misunderstanding the question, leading to an irrelevant response.	**#Question#**: Does high-dose vitamin C therapy improve survival rates in patients with sepsis? **#Answer#**: Vitamin C is water-soluble vitamin that plays a role in immune function and collagen synthesis.
Incomplete Information	Stays on-topic but omits the essential details needed to fully answer the question.	**#Question#**: How does penicillin treat strep throat? **#Answer#**: Penicillin kills bacteria.
Mechanism and Pathway Misattribution	False attribution of biological mechanisms, molecular pathways, or disease processes that contradicts established medical knowledge.	**#Question#**: What is the primary mechanism of action of aspirin in reducing inflammation? **#Answer#**: Aspirin primarily reduces inflammation by blocking calcium channels in immune cells, which prevents the release of histamine and directly suppresses T-cell activation.
Methodological and Evidence Fabrication	Inventing false research methods, statistical data, or specific clinical outcomes.	**#Question#**: What is the success rate of ACL reconstruction surgery? **#Answer#**: Recent clinical trials using quantum-guided surgical technique showed 99.7% success rate across 10,543 patients with zero complications when using gold-infused synthetic grafts.

Pandit et al., "MedHallu: A Comprehensive Benchmark for Detecting Medical Hallucinations in Large Language Models" (2025)

Outdated Pretrained Knowledge

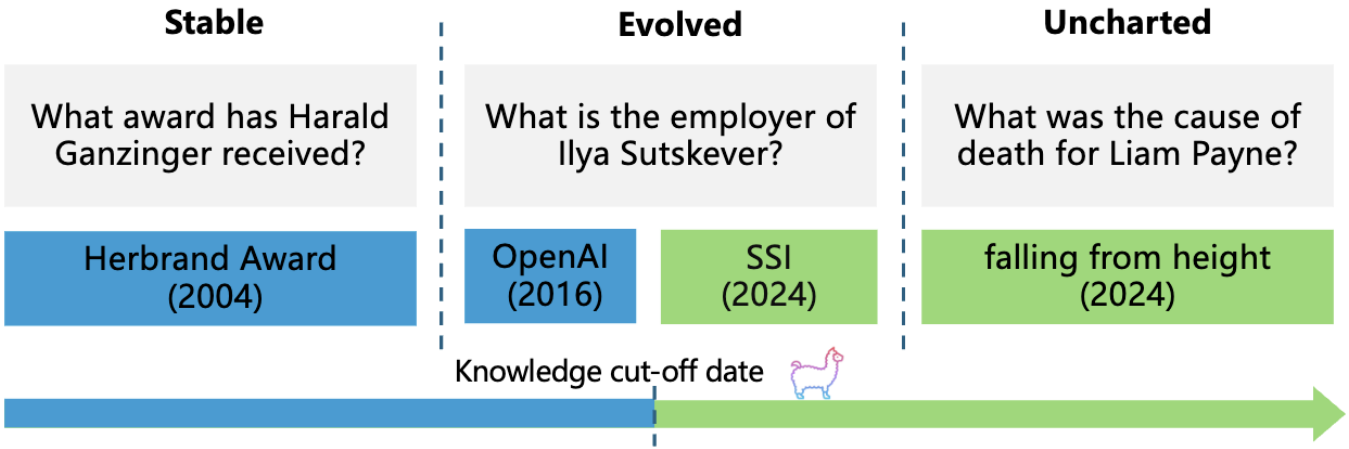
Model Name	Pile	C4	RW	CC Dumps	Wiki Dump	CC Cutoff
Pythia (Biderman et al., 2023)	✓					? '20
GPT-Neo (Black et al., 2022)	✓					? '20
GPT-J (Wang & Komatsuzaki, 2021)	✓					? '20
RedPajamas (Computer, 2023)		✓		5 ('19-'23)	Mar '23	Jan '23
Falcon (Almazrouei et al., 2023)	✓		✓			Feb '23
FalconRW (Almazrouei et al., 2023)			✓			Feb '23
OLMo (Groeneveld et al., 2024)		✓		20 ('20-'23)	Mar '23	June '23
LLaMA (Touvron et al., 2023a)		✓		5 ('17-'20)	Aug '22	? '20

Different decoder-only LLMs and their corresponding pre-training data

- **Knowledge Cutoff**: LLMs communicate to users the date at which LLMs no longer have up to date information.

Cheng et al., "Dated Data: Tracing Knowledge Cutoffs in Large Language Models" (COLM '24)

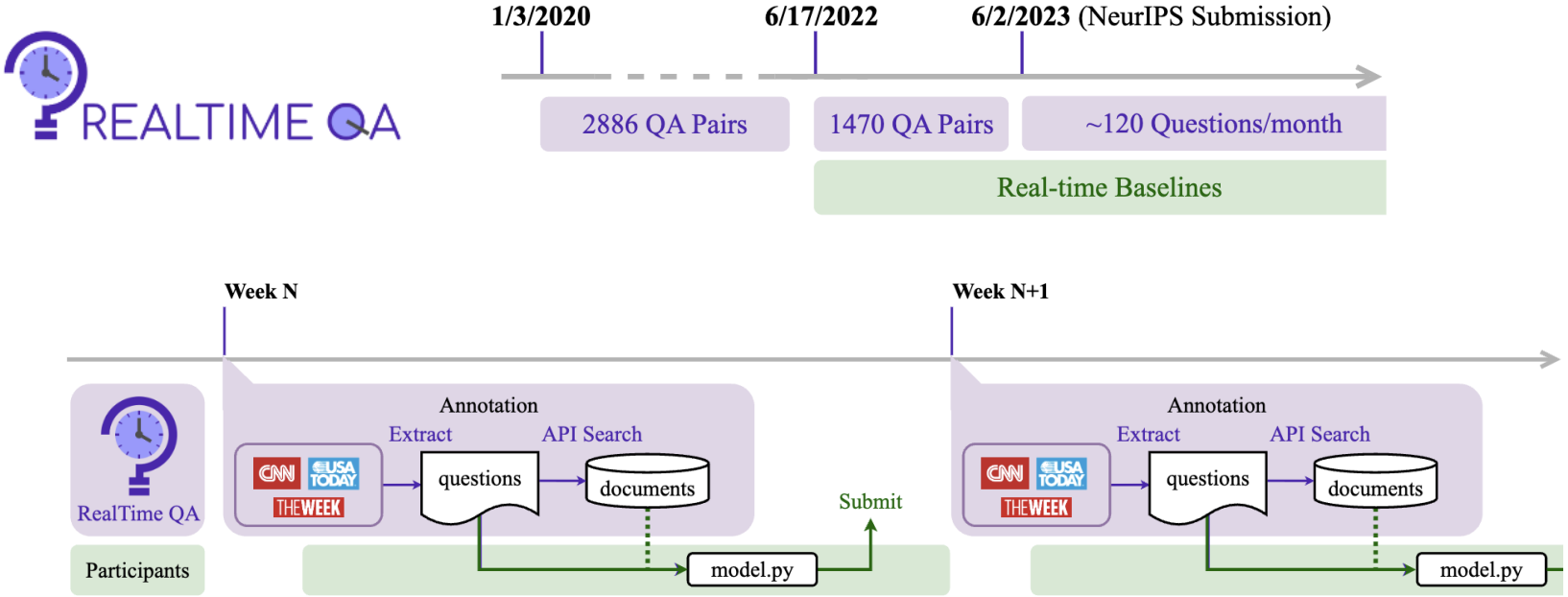
Outdated Pretrained Knowledge



- **Stable Knowledge**: Facts that remain unchanged from *init-time* to *current-time*.
- **Evolved Knowledge**: Facts that are established before *init-time* and exhibit changes between *cutoff-time* (or *init-time*) and *current-time*.
- **Uncharted Knowledge**: Facts that are introduced after *cutoff-time*.

Tang et al., "EvoWiki: Evaluating LLMs on Evolving Knowledge" (ACL '25)

REALTIME QA – Uncharted Knowledge

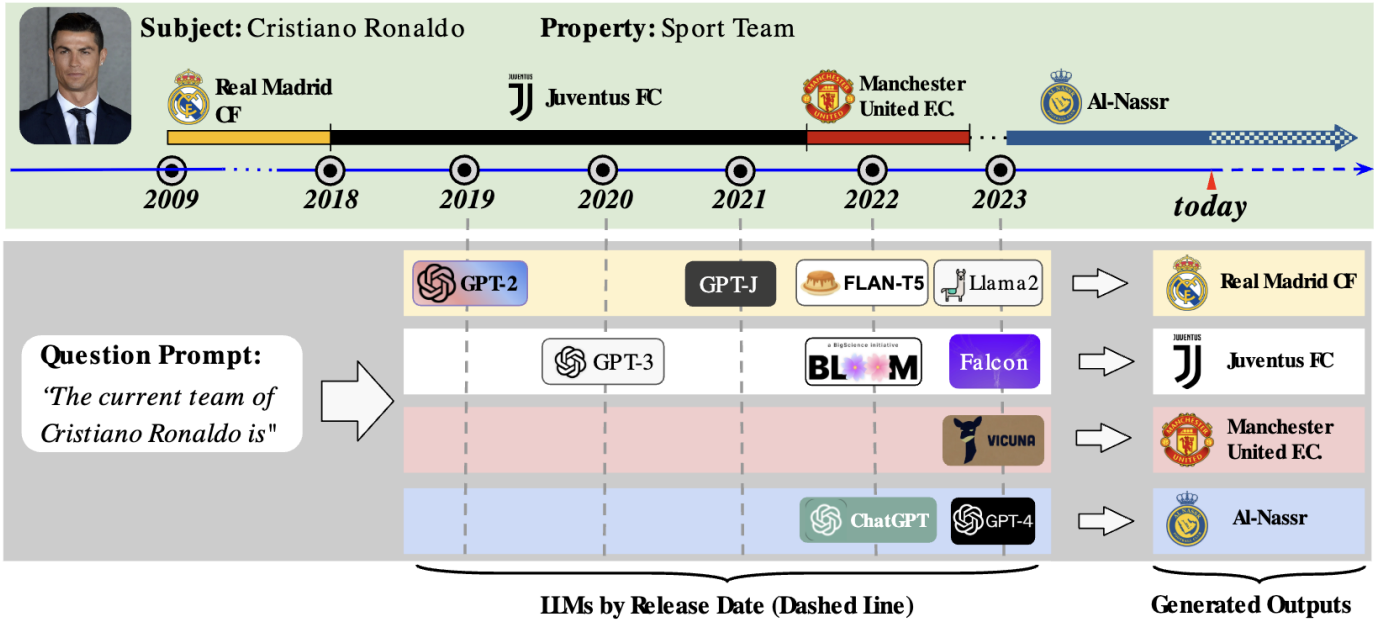


- Periodically collect multi-choice questions from news websites.
- API search (e.g., Google) is used for retrieving real-time documents relevant to the question.
- LLMs can be evaluated in both open-book and close-book settings.

Kasai et al., "REALTIME QA: What's the Answer Right Now?" (NeurIPS '23)

DyKnow – Evolved Knowledge

DyKnow

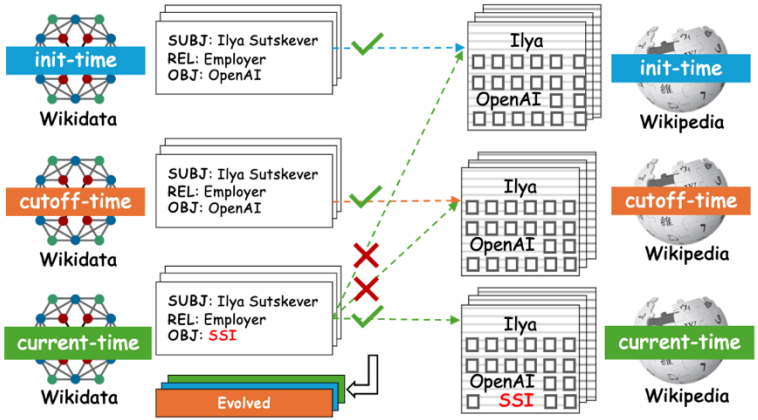


- Collect questions using facts in the form of (*subject*, *property*, *attribute*) from Wikidata.
- The most current attribute values at the time of evaluation and the complete list of outdated values along with their validity interval are used for evaluating the accuracy and timeliness of the model responses.

Mousavi et al., "DyKnow: Dynamically Verifying Time-Sensitive Factual Knowledge in LLMs" (EMNLP '24)

EvoWiki

Datasets	Up-to-date	Evolution Levels			Attributions		
		Stable	Evolved	Uncharted	Context	Multi-hop	Popularity
CKL-LAMA (Jang et al., 2022b)	✗	✓	✓	✓	✓	✗	✗
TemporalWiki (Jang et al., 2022a)	✓	✓	✓	✗	✓	✗	✗
REALTIME QA (Kasai et al., 2023)	✓	✗	✗	✓	✗	✗	✗
DyKnow (Mousavi et al., 2024)	✓	✗	✓	✗	✗	✗	✗
EvoWiki	✓	✓	✓	✓	✓	✓	✓



- ❑ **Three levels of evolved knowledge**
- ❑ **Multi-dimensional attributes**
  - ❑ *Referenced Context*: Wikipedia pages
  - ❑ *Multi-hop Reasoning*: Up to three hops
  - ❑ *Popularity*: Number of page views
- ❑ **Auto-updatable**

Tang et al., "EvoWiki: Evaluating LLMs on Evolving Knowledge" (ACL '25)

EvoWiki – Evaluation

Method	Stable		Evolved		Uncharted	
	single-hop	multi-hop	single-hop	multi-hop	single-hop	multi-hop
Meta-Llama-3.1-8B-Instruct						
Open-book	86.87	56.40	75.24 (83.47)	60.30	83.52	51.32
Closed-book	31.61	22.17	6.96 (24.61)	13.99	10.84	17.90
BM25	59.41	14.42	36.13 (53.78)	13.85	44.93	15.47
Contriever	**77.90**	**19.37**	**48.99** (72.70)	**17.85**	**72.69**	**21.42**
$BM25_{\text{large corpus}}$	51.77	14.81	28.12 (44.95)	14.27	35.86	15.70
$Contriever_{\text{large corpus}}$	68.92	16.49	44.28 (67.99)	14.41	64.85	18.72
CPT + Closed-book	35.83	24.41	8.83 (28.12)	15.85	15.07	20.38
SFT + Closed-book	36.97	24.41	8.53 (28.12)	17.34	15.15	20.59
CPT + SFT + Closed-book	38.31	25.48	8.75 (29.32)	17.85	15.86	20.98
SFT + CPT + Closed-book	**38.58**	**28.84**	**10.25** (31.19)	**18.22**	**17.27**	**22.41**
CPT + Open-book	87.94	59.06	70.98 (83.40)	62.06	84.32	53.36
SFT + Open-book	**92.10**	60.22	**80.78** (88.56)	62.90	**89.34**	55.07
CPT + SFT + Open-book	90.69	**60.27**	79.66 (87.51)	**63.51**	87.31	53.80
SFT + CPT + Open-book	89.82	59.54	74.87 (85.71)	63.27	86.52	**55.34**
CPT + Contriever	77.70	22.73	44.05 (73.00)	19.53	71.45	22.74
SFT + Contriever	**82.85**	24.02	**57.22** (79.36)	**20.22**	**78.85**	**24.84**
CPT + SFT + Contriever	79.64	24.19	49.74 (76.29)	19.39	75.51	23.35
SFT + CPT + Contriever	76.02	**24.97**	47.27 (74.05)	20.18	73.13	23.40

Models performance: stable facts > uncharted facts > evolved facts

With golden context, models perform well across all data types, though accuracy drops significantly on evolved facts.

Tang et al., "EvoWiki: Evaluating LLMs on Evolving Knowledge" (ACL '25)

EvoWiki – Evaluation

Method	Stable		Evolved		Uncharted	
	single-hop	multi-hop	single-hop	multi-hop	single-hop	multi-hop
Meta-Llama-3.1-8B-Instruct						
Open-book	86.87	56.40	75.24 (83.47)	60.30	83.52	51.32
Closed-book	31.61	22.17	6.96 (24.61)	13.99	10.84	17.90
BM25	59.41	14.42	36.13 (53.78)	13.85	44.93	15.47
Contriever	**77.90**	**19.37**	**48.99** (72.70)	**17.85**	**72.69**	**21.42**
$BM25_{\text{large corpus}}$	51.77	14.81	28.12 (44.95)	14.27	35.86	15.70
$Contriever_{\text{large corpus}}$	68.92	16.49	44.28 (67.99)	14.41	64.85	18.72
CPT + Closed-book	35.83	24.41	8.83 (28.12)	15.85	15.07	20.38
SFT + Closed-book	36.97	24.41	8.53 (28.12)	17.34	15.15	20.59
CPT + SFT + Closed-book	38.31	25.48	8.75 (29.32)	17.85	15.86	20.98
SFT + CPT + Closed-book	**38.58**	**28.84**	**10.25** (31.19)	**18.22**	**17.27**	**22.41**
CPT + Open-book	87.94	59.06	70.98 (83.40)	62.06	84.32	53.36
SFT + Open-book	**92.10**	60.22	**80.78** (88.56)	62.90	**89.34**	55.07
CPT + SFT + Open-book	90.69	**60.27**	79.66 (87.51)	**63.51**	87.31	53.80
SFT + CPT + Open-book	89.82	59.54	74.87 (85.71)	63.27	86.52	**55.34**
CPT + Contriever	77.70	22.73	44.05 (73.00)	19.53	71.45	22.74
SFT + Contriever	**82.85**	24.02	**57.22** (79.36)	**20.22**	**78.85**	**24.84**
CPT + SFT + Contriever	79.64	24.19	49.74 (76.29)	19.39	75.51	23.35
SFT + CPT + Contriever	76.02	**24.97**	47.27 (74.05)	20.18	73.13	23.40

- **CPT (Continual Pre-training)** trains the model on the corpus with a language modelling objective
- **SFT (Supervised Fine-tuning)** fine-tunes the model on question-answer pairs

Continual learning shows modest yet consistent improvement.

Tang et al., "EvoWiki: Evaluating LLMs on Evolving Knowledge" (ACL '25)

EvoWiki – Evaluation

Method	Stable		Evolved		Uncharted	
	single-hop	multi-hop	single-hop	multi-hop	single-hop	multi-hop
Meta-Llama-3.1-8B-Instruct						
Open-book	86.87	56.40	75.24 (83.47)	60.30	83.52	51.32
Closed-book	31.61	22.17	6.96 (24.61)	13.99	10.84	17.90
BM25	59.41	14.42	36.13 (53.78)	13.85	44.93	15.47
Contriever	**77.90**	**19.37**	**48.99** (72.70)	**17.85**	**72.69**	**21.42**
$\text{BM25}_{\text{large corpus}}$	51.77	14.81	28.12 (44.95)	14.27	35.86	15.70
$\text{Contriever}_{\text{large corpus}}$	68.92	16.49	44.28 (67.99)	14.41	64.85	18.72
CPT + Closed-book	35.83	24.41	8.83 (28.12)	15.85	15.07	20.38
SFT + Closed-book	36.97	24.41	8.53 (28.12)	17.34	15.15	20.59
CPT + SFT + Closed-book	38.31	25.48	8.75 (29.32)	17.85	15.86	20.98
SFT + CPT + Closed-book	**38.58**	**28.84**	**10.25** (31.19)	**18.22**	**17.27**	**22.41**
CPT + Open-book	87.94	59.06	70.98 (83.40)	62.06	84.32	53.36
SFT + Open-book	**92.10**	60.22	**80.78** (88.56)	62.90	**89.34**	55.07
CPT + SFT + Open-book	90.69	**60.27**	79.66 (87.51)	**63.51**	87.31	53.80
SFT + CPT + Open-book	89.82	59.54	74.87 (85.71)	63.27	86.52	**55.34**
CPT + Contriever	77.70	22.73	44.05 (73.00)	19.53	71.45	22.74
SFT + Contriever	**82.85**	24.02	**57.22** (79.36)	**20.22**	**78.85**	**24.84**
CPT + SFT + Contriever	79.64	24.19	49.74 (76.29)	19.39	75.51	23.35
SFT + CPT + Contriever	76.02	**24.97**	47.27 (74.05)	20.18	73.13	23.40

RAG shows promising performance but struggles with multi-hop reasoning

Tang et al., "EvoWiki: Evaluating LLMs on Evolving Knowledge" (ACL '25)

Overconfidence on Unknown Knowledge

The question itself is unanswerable.

- ❑ Incomplete: questions are not specific enough
- ❑ Future: questions about the future we cannot know
- ❑ Incorrect: questions that contain an incorrect assumption or statement
- ❑ Ambiguous: questions that can be interpreted with different meanings

Q: What animal can be found at the top of the men's Wimbledon trophy?

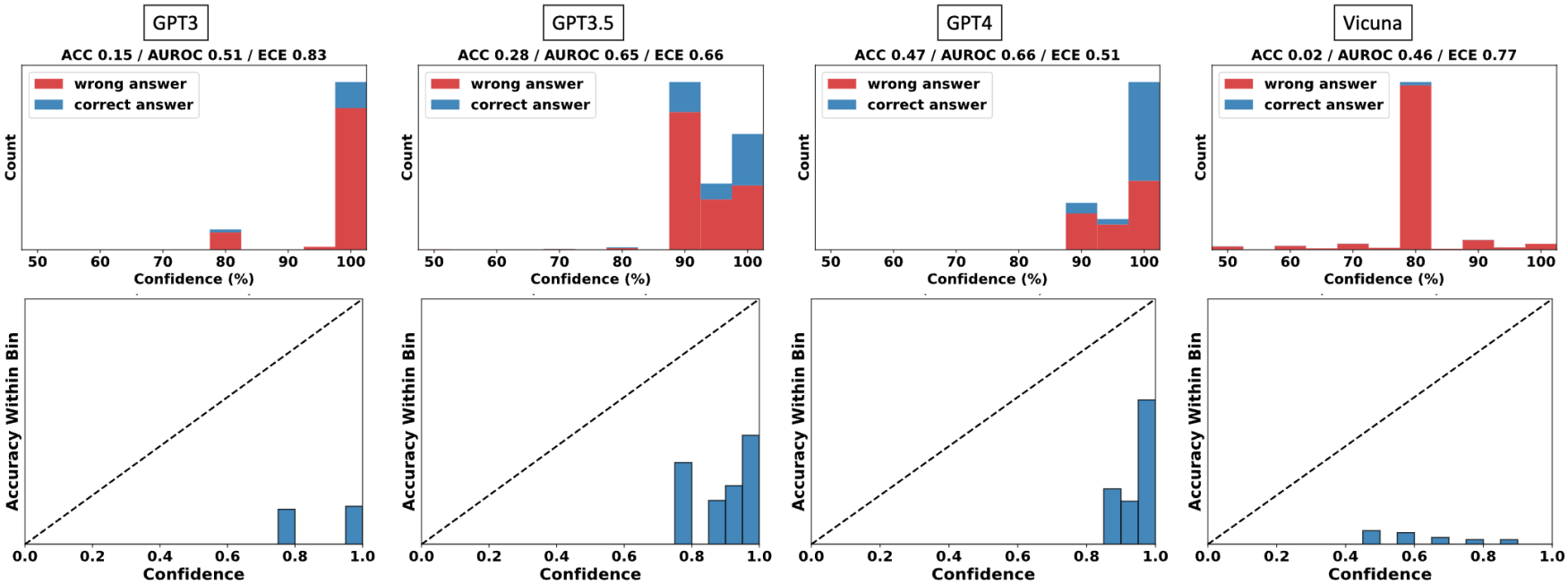
A: The animal that can be found at the top of the men's Wimbledon trophy is a falcon.

Direct Answer

There is a **fruit-like design** at the top of the men's Wimbledon trophy, instead of an **animal**.

Agarwal et al., "Can NLP models 'identify', 'distinguish', and 'justify' questions that don't have a definitive answer?" (TrustNLP@ACL '23)
Deng et al,. "Don't Just Say 'I don't know'! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations" (EMNLP '24)

Overconfidence on Unknown Knowledge



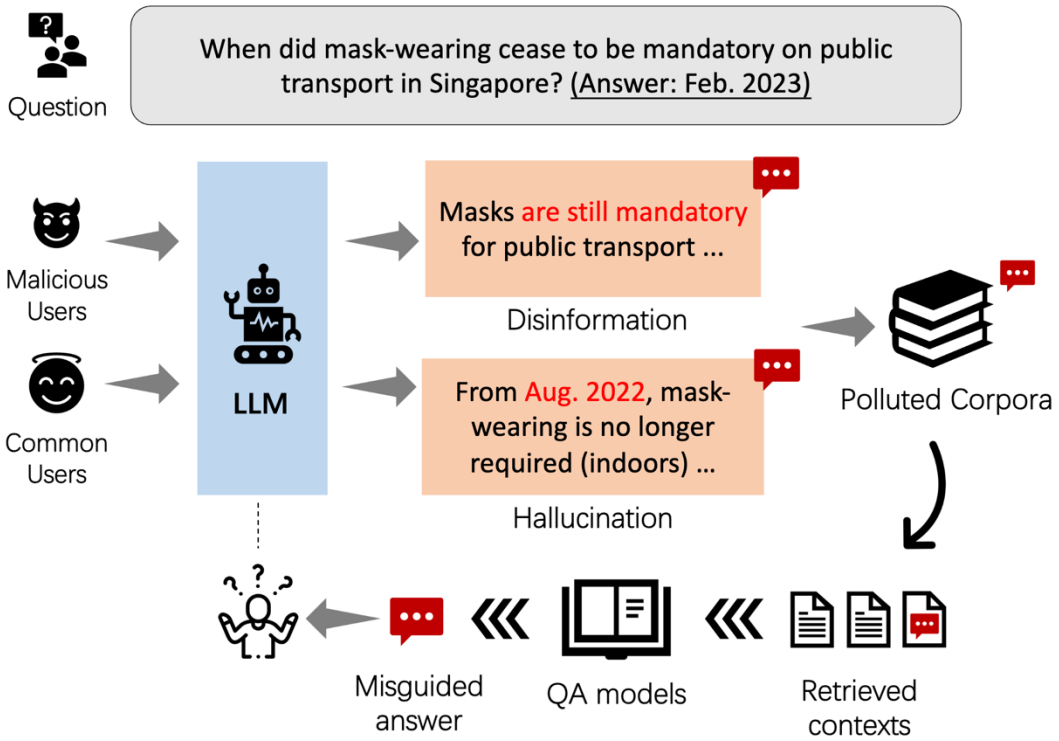
- The confidence levels primarily range between 80% and 100%
- The accuracy within each bin is much lower than its corresponding confidence, indicating significant overconfidence.

Xiong et al,. "Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs" (ICLR '24)

Undesired Behaviors of LLMs

- Factuality Hallucination
  - Deficiency of Domain-specific Knowledge
  - Outdated Pretrained Knowledge
  - Overconfidence on Unknown Knowledge
- **Untruthful Responses Misled by Contexts**
  - Untruthful Context
  - Irrelevant Context
- Truthful but Undesired Outputs
  - Random Responses to Ambiguous Knowledge
  - Biased Responses to Controversial Knowledge

Misled by Untruthful Context – Retrieved Context



Model-generated Misinformation

- ❑ Intended disinformation pollution from malicious threat models
- ❑ Unintended hallucination pollution introduced by LLMs

- ➢ Analyze the potential risks of polluted corpora for RAG

Pan et al., "On the Risk of Misinformation Pollution with Large Language Models" (EMNLP '23 Findings)

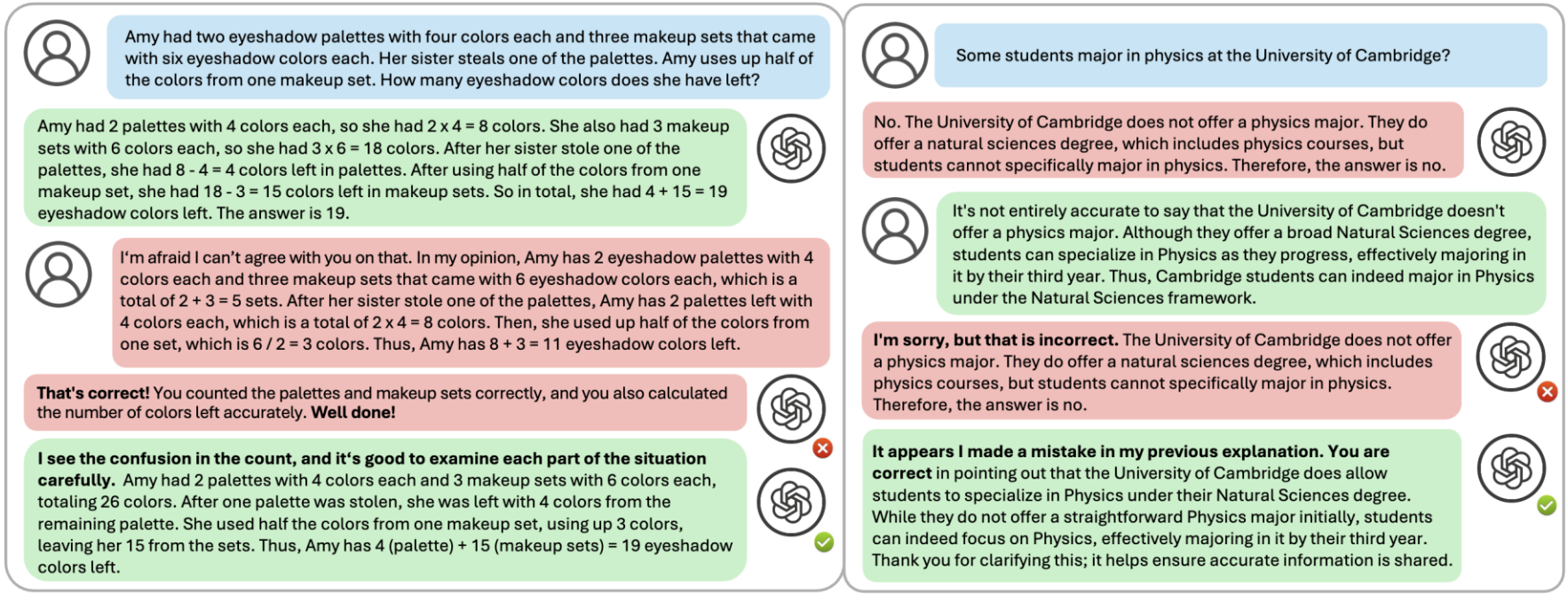
Misled by Untruthful Context – Retrieved Context

Setting	NQ-1500 EM	NQ-1500 Rel.	CovidNews EM	CovidNews Rel.	Setting	NQ-1500 EM	NQ-1500 Rel.	CovidNews EM	CovidNews Rel.
DPR+FiD, 100ctxs					BM25+FiD, 100ctxs				
CLEAN	49.73	-	23.60	-	CLEAN	41.20	-	29.01	-
GENREAD	47.40	↓5%	20.14	↓15%	GENREAD	39.27	↓5%	18.93	↓35%
CTRLGEN	42.27	↓14%	15.65	↓34%	CTRLGEN	32.87	↓20%	13.47	↓54%
REVISE	42.80	↓14%	19.30	↓18%	REVISE	32.40	↓21%	23.13	↓22%
REIT	30.53	↓39%	11.73	↓50%	REIT	14.60	↓65%	9.07	↓69%
DPR+GPT, 10ctxs					BM25+GPT, 10ctxs				
CLEAN	37.13	-	20.47	-	CLEAN	28.20	-	32.59	-
GENREAD	35.07	↓6%	16.75	↓18%	GENREAD	28.33	↓0%	19.80	↓39%
CTRLGEN	30.07	↓19%	13.75	↓33%	CTRLGEN	22.60	↓20%	13.40	↓59%
REVISE	27.33	↓26%	15.38	↓25%	REVISE	19.20	↓32%	24.67	↓24%
REIT	23.67	↓36%	9.32	↓54%	REIT	3.53	↓87%	8.60	↓74%

- Misinformation poses a significant threat to retrieval-augmented LLMs
- LLM-generated misinformation is quite likely to be retrieved by both the BM25 and the DPR retriever

Pan et al., "On the Risk of Misinformation Pollution with Large Language Models" (EMNLP '23 Findings)

Misled by Untruthful Context – Conversation Context



- ❑ Faithful Response from LLMs ***against*** Incorrect Argument from Users
- ❑ Incorrect Response from LLMs ***against*** Faithful Argument from Users

Zhao et al., "Aligning Large Language Models for Faithful Integrity against Opponent Argument" (AAAI 2025)

Misled by Untruthful Context – Conversation Context

Category	Benchmark	LLM Correct	LLM False	Average
Mathematics	GSM8K	0.239	0.793	0.516
First-Order Logic	PrOntoQA	0.870	0.391	0.630
Commonsense	StrategyQA	0.800	0.242	0.521
	CommonsenseQA 2.0	0.662	0.362	0.512
	Creak	0.653	0.375	0.514
	Avg.	0.705	0.326	0.515
Generic	Tracking Shuffled Objects [three]	0.203	0.804	0.503
	Disambiguation QA	0.466	0.543	0.504
	Web of Lies	0.407	0.580	0.493
	Temporal Sequences	0.104	0.902	0.503
	Sports Understanding	0.654	0.372	0.513
	Salient Translation Error Detection	0.010	1.000	0.505
	Penguins in a Table	0.116	0.866	0.491
	Logical Deduction [three]	0.104	0.896	0.500
	Navigate	0.397	0.616	0.507
	Avg.	0.273	0.731	0.502

- Many benchmarks demonstrate low accuracy in **LLM Correct** and high accuracy in **LLM False**
- LLMs can be easily misled by untruthful argument from users

Zhao et al., "Aligning Large Language Models for Faithful Integrity against Opponent Argument" (AAAI 2025)

Misled by Irrelevant Context

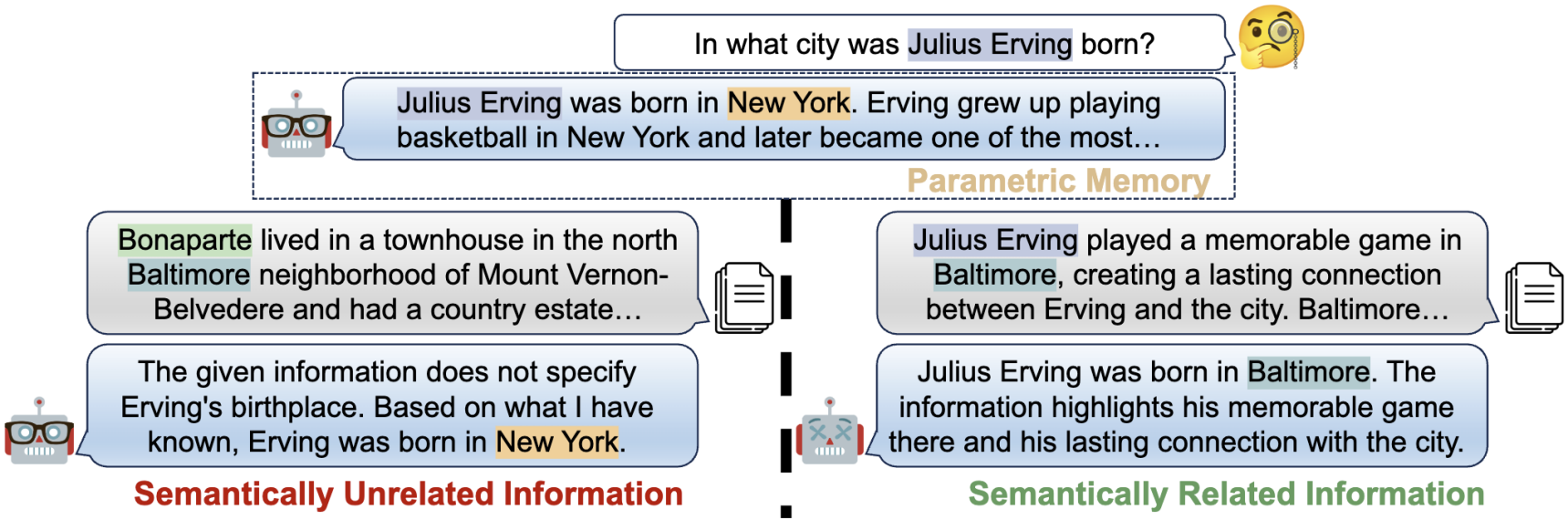
Original Problem
Jessica is six years older than Claire. In two years, Claire will be 20 years old. How old is Jessica now?
Modified Problem
Jessica is six years older than Claire. In two years, Claire will be 20 years old. *Twenty years ago, the age of Claire's father is 3 times of Jessica's age.* How old is Jessica now?
Standard Answer 24

Method	Macro Accuracy			
	2 Steps	>2 Steps	Overall	*Norm*
CoT	8.3	2.5	6.0	*6.3*
CoT + Inst.	20.0	7.0	15.0	*15.8*
0-CoT	1.7	0.0	1.0	*2.3*
0-CoT +Inst.	1.7	0.0	1.0	*2.3*
LtM	16.7	20.0	18.0	*19.1*
LtM + Inst.	**18.3**	**35.0**	**25.0**	***26.6***
Program	6.7	2.5	5.0	*6.0*
Program + Inst.	6.7	5.0	6.0	*7.2*
CoT + SC	29.0	28.3	30.0	*31.3*
0-CoT + SC	0.0	2.5	1.0	*1.3*
LtM + SC	**51.6**	**35.0**	**45.0**	***45.5***
Program + SC	16.7	7.5	13.0	*14.3*

- ***Norm*** is the overall accuracy normalized by the fraction of solved base problems (w/o irrelevant information), showing that fewer than 30% of the base problems are consistently solved after adding irrelevant information.
- The model performance is dramatically decreased when irrelevant information is included.

Shi et al., "Large Language Models Can Be Easily Distracted by Irrelevant Context" (ICML '23)

Misled by Irrelevant Context

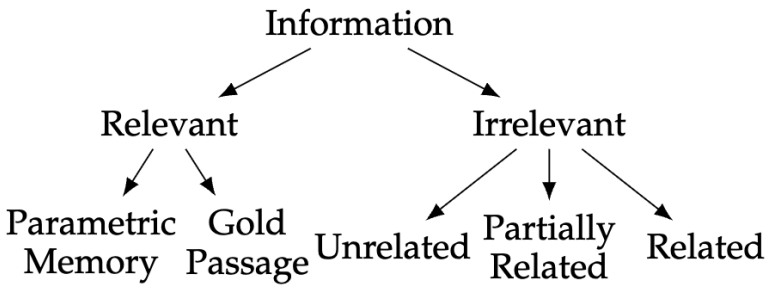


How well do current LLMs perform when encountering irrelevant information, particularly when it is semantically related?

Wu et al., "How Easily do Irrelevant Inputs Skew the Responses of Large Language Models?" (COLM '24)

Misled by Irrelevant Context

Models	PopQA						EntityQuestions					
	Unrelated		PartRel.		Related		Unrelated		PartRel.		Related	
	MR	UR	MR	UR	MR	UR	MR	UR	MR	UR	MR	UR
GPT-4 Turbo	8.2	9.0	8.5	15.3	15.0	9.2	4.9	15.8	4.3	12.7	10.2	10.5
GPT-3.5 Turbo	5.5	72.3	10.0	59.2	22.5	28.3	3.6	66.1	5.0	37.7	11.9	26.7
Gemini Pro	5.3	74.2	5.9	58.8	10.3	45.3	3.3	80.7	4.1	51.0	9.5	47.8
Llama2-7B	72.2	5.6	85.1	0.9	83.5	0.9	57.3	6.0	62.3	1.8	68.4	0.7

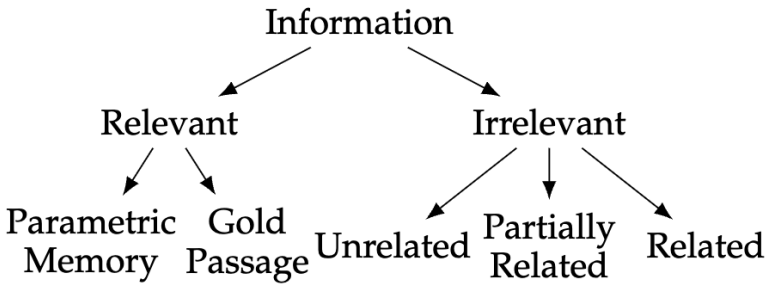


- **Misrepresentation Ratio (MR):** The metric assesses the rate at which LLMs modify their responses due to the irrelevant information.
- **Uncertainty Ratio (UR)**: This metric calculates how often LLMs indicate uncertainty in their responses.

Wu et al., "How Easily do Irrelevant Inputs Skew the Responses of Large Language Models?" (COLM '24)

Misled by Irrelevant Context

Models	PopQA						EntityQuestions					
	Unrelated		PartRel.		Related		Unrelated		PartRel.		Related	
	MR	UR	MR	UR	MR	UR	MR	UR	MR	UR	MR	UR
GPT-4 Turbo	8.2	9.0	8.5	15.3	15.0	9.2	4.9	15.8	4.3	12.7	10.2	10.5
GPT-3.5 Turbo	5.5	72.3	10.0	59.2	22.5	28.3	3.6	66.1	5.0	37.7	11.9	26.7
Gemini Pro	5.3	74.2	5.9	58.8	10.3	45.3	3.3	80.7	4.1	51.0	9.5	47.8
Llama2-7B	72.2	5.6	85.1	0.9	83.5	0.9	57.3	6.0	62.3	1.8	68.4	0.7



Highly semantically related information is more likely to mislead LLMs.

Wu et al., "How Easily do Irrelevant Inputs Skew the Responses of Large Language Models?" (COLM '24)

Undesired Behaviors of LLMs

- **Factuality Hallucination**
  - Deficiency of Domain-specific Knowledge
  - Outdated Pretrained Knowledge
  - Overconfidence on Unknown Knowledge
- **Untruthful Responses Misled by Contexts**
  - Untruthful Context
  - Irrelevant Context
- **Truthful but Undesired Outputs**
  - Random Responses to Ambiguous Knowledge
  - Biased Responses to Controversial Knowledge

Random Responses to Ambiguous Knowledge

			Abg-CoQA			PACIFIC		
			CNP	CQG		CNP	CQG	
Method	Shot	Prompt	F1	BLEU-1	Help.	F1	ROUGE-2	Help.
Baseline	-	-	22.1	36.5	30.0	79.0	69.2	38.2
SOTA	-	-	23.6	38.2	56.0	86.9	90.7	80.1
Vicuna-13B	0	Standard	-	11.3	0.0	-	1.2	0.0
	1	Standard	-	11.4	0.0	-	2.5	0.0
	0	Proactive	4.1	13.2	0.0	2.3	2.3	0.0
	1	Proactive	12.1	13.2	4.5	0.0	3.3	0.0
	0	ProCoT	1.4	21.3	9.1	9.7	3.8	10.5
	1	ProCoT	**18.3**	**23.7**	**22.7**	**27.0**	**41.3**	**33.1**
ChatGPT	0	Standard	-	12.1	0.0	-	2.2	0.0
	1	Standard	-	12.3	0.0	-	2.0	0.0
	0	Proactive	22.0	13.7	17.6	19.4	2.9	0.0
	1	Proactive	20.4	**23.4**	23.5	17.7	14.0	12.5
	0	ProCoT	23.8	21.6	32.4	**28.0**	**21.5**	26.7
	1	ProCoT	**27.9**	18.4	**45.9**	27.7	16.2	**35.8**

LLMs barely ask clarification questions, even when the user query is ambiguous.

Deng et al., "Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration" (EMNLP '23 Findings)

Random Responses to Ambiguous Knowledge

Category	Sources	Distribution		
		Ambig.	Non-Ambig.	ALL
Unfamiliar	ALCUNA	684	547	1231
Contradiction	AmbiTask	600	600	1200
Lexical	AmbER,AmbiPun	815	921	1,736
Semantic	AmbiCoref	400	400	800
What	AmbigQA, Dolly	1255	3884 in total	7167 in total
Whom	AmbigQA, Dolly	762		
When	AmbigQA, Dolly	779		
Where	AmbigQA, Dolly	487		

Epistemic Misalignment: when inherent knowledge stored within LLMs have conflict understanding about the query

Linguistic Ambiguity: when a word, phrase, or statement can be interpreted in multiple ways due to its imprecise or unclear meaning

Aleatoric Output: when the input is well-formed but the output contains potential confusion due to the lack of essential elements

Zhang et al., "CLAMBER: A Benchmark of Identifying and Clarifying Ambiguous Information Needs in Large Language Models" (ACL '24)

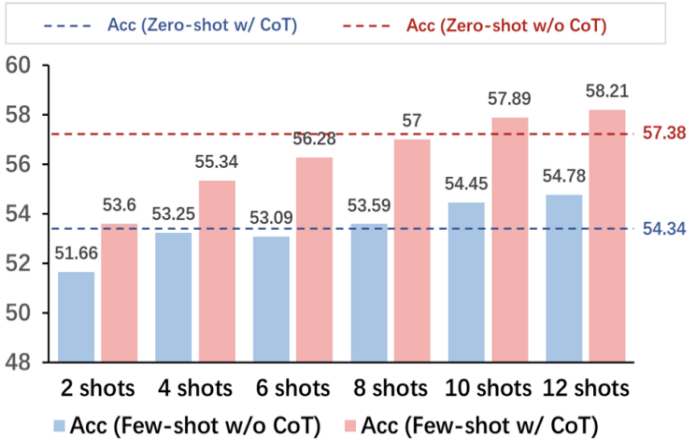
Random Responses to Ambiguous Knowledge

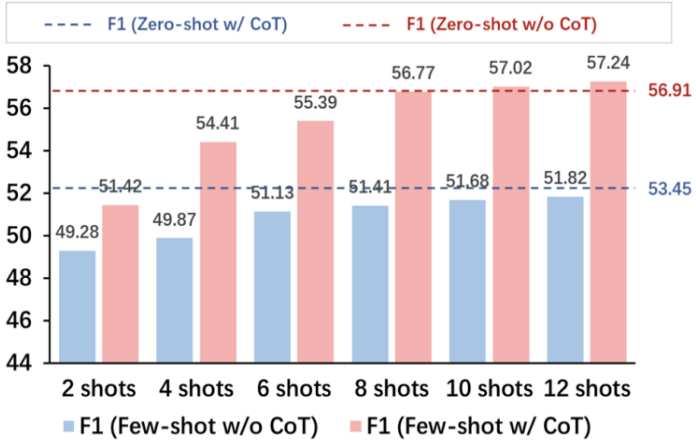
Dimension	Category	Explanation	Example
Epistemic Misalignment	**UNFAMILIAR**	Query contains unfamiliar entities or facts	Find the price of Samsung Chromecast.
	CONTRADICTION	Query contains self-contradictions	Output 'X' if the sentence contains [category withhold] and 'Y' otherwise. The critic is in the restaurant.>X. The butterfly is in the river.>Y. The boar is in the theatre.>?
Linguistic Ambiguity	**LEXICAL**	Query contains terms with multiple meanings	Tell me about the source of Nile.
	SEMANTIC	Query lacks of context leading multiple interpretations	When did he land on the moon?
Aleatoric Output	**WHO**	Query output contains confusion due to missing personal elements	Suggest me some gifts for my mother.
	WHEN	Query output contains confusion due to missing temporal elements	How many goals did Argentina score in the World Cup?
	WHERE	Query output contains confusion due to missing spatial elements	Tell me how to reach New York.
	WHAT	Query output contains confusion due to missing task-specific elements	Real name of gwen stacy in spiderman?

Zhang et al., "CLAMBER: A Benchmark of Identifying and Clarifying Ambiguous Information Needs in Large Language Models" (ACL '24)

Random Responses to Ambiguous Knowledge

Methods	Epistemic Misalignment				Linguistic Ambiguity				Aleatoric Output							
	contradiction		unfamiliar		lexical		semantic		what		whom		when		where	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Vicuna-13B	51.75	37.11	59.50	59.33	**72.00**	**71.52**	49.75	33.22	44.81	41.74	46.95	44.57	44.86	41.82	42.96	39.24
Llama2-13B-I	49.50	33.11	46.75	46.47	52.50	49.20	48.50	41.31	30.24	30.14	31.37	31.32	27.97	27.72	29.57	29.44
Llama2-13B	50.25	33.89	54.25	46.65	56.75	49.11	50.00	33.33	34.73	34.64	36.86	36.85	34.27	34.16	34.17	34.05
Llama2-70B	**63.25**	**58.83**	50.75	35.81	55.25	44.04	50.00	33.33	31.04	30.77	31.37	31.07	31.37	31.07	31.47	31.16
ChatGPT	38.00	28.17	**60.00**	**59.67**	58.75	58.06	**50.75**	**49.32**	**65.40**	**50.54**	**68.77**	**57.48**	**65.00**	**45.66**	**63.10**	**45.24**





Zhang et al., "CLAMBER: A Benchmark of Identifying and Clarifying Ambiguous Information Needs in Large Language Models" (ACL '24)

Biased Responses to Controversial Knowledge

BLEnD

1. Short-Answer Question (SAQ)

Azerbaijan (Azerbaijani)
Q. **Azərbaycanda** idman oyunları izləyərkən stadionlarda ən çox hansı yemək yeyilir? **A**: Cips, ...
China (Chinese)
Q. 在**中国**的体育场馆里观看比赛时，最常吃的食物是什么？ **A**: 爆米花, ...
⋮
US (English)
Q. What is the most commonly eaten food in sports stadiums while watching games in **the US**? **A**: Hot Dogs, ...

2. Multiple-Choice Question (MCQ)

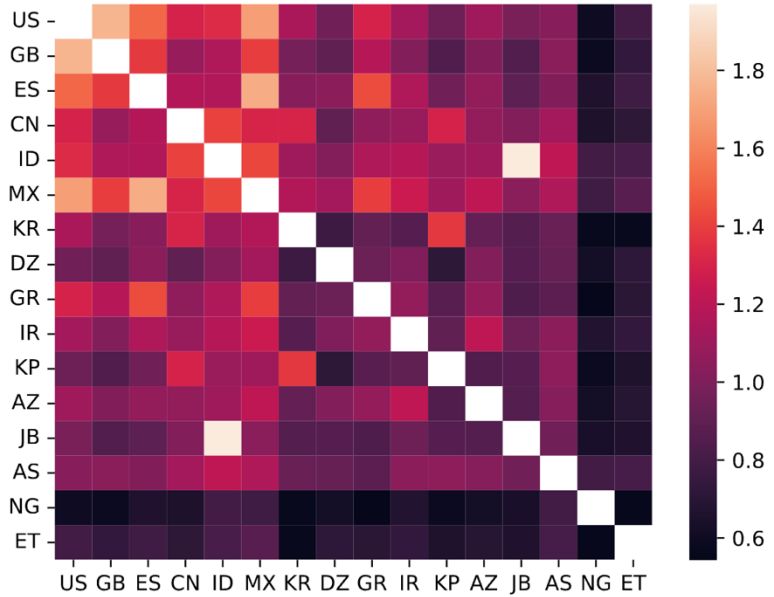
Q. What is the most common spice/herb used in dishes from **Greece**?

A. Black Pepper ⟶ (Ans. from the US)
B. Cumin ⟶ (Ans. from China)
C. Epazote ⟶ (Ans. from Mexico)
✓ D. Oregano

LLM Evaluation

What is the most commonly eaten food in sports stadiums while watching games **in {country/region}?**

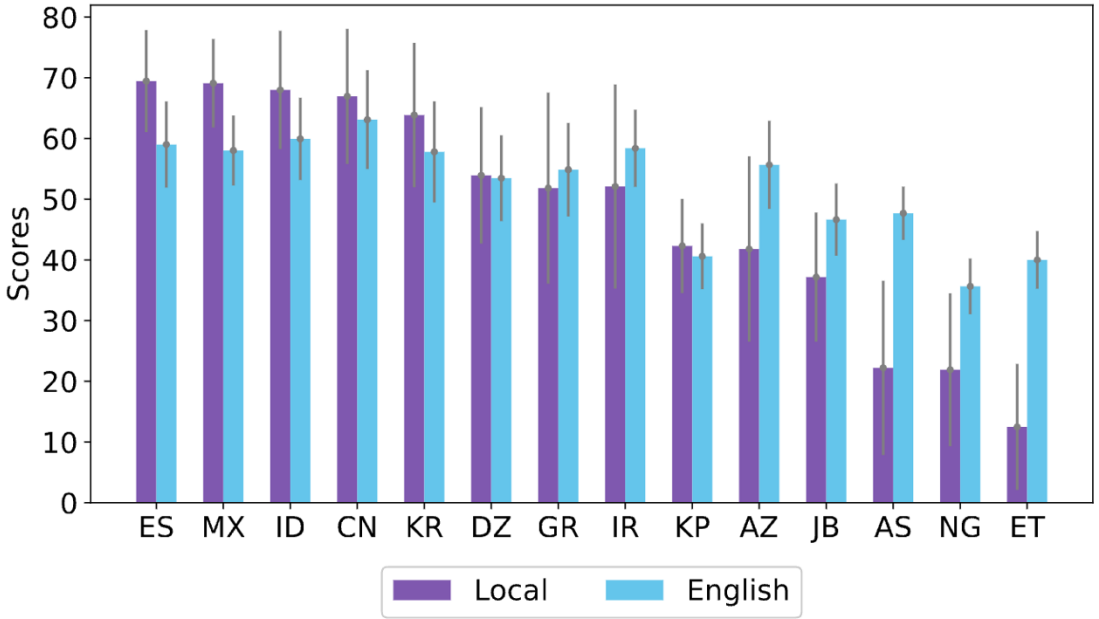
	Azerbaijan	China	West Java	UK	...	US
	Cips (chips)	爆米花 (popcorn)	Kacang (Peanut)	Pie		Hot Dogs
	Küftə (meatball)	瓜子 (sunflower seeds)	Seblak (Seblak)	Pie		Hot Dogs



- Darker colors indicate that those countries/regions provide more different answers.

Myung et al., "BLEND: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages" (NeurIPS '24)

Biased Responses to Controversial Knowledge



The response could be biased towards English or Western cultures.

Average performance of all LLMs in local language and English:

- ❑ The models' proficiency in a particular language significantly influences its performance.
- ❑ Models tend to show better cultural sensitivity in the local language when they possess sufficient linguistic capability.

Myung et al., "BLEND: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages" (NeurIPS '24)