

ERBench

:An Entity-Relationship based Automatically Verifiable Hallucination Benchmark for LLMs
(*NeurIPS'24 Database and Benchmark Spotlight*)


Jio Oh^{*1}, Soyeon Kim^{*1}, Junseok Seo¹, Jindong Wang², Ruochen Xu³, Xing Xie², Steven Euijong Whang¹






* Equal contribution ¹KAIST ²Microsoft Research Asia ³Microsoft Azure

Hallucination of Large Language Models (LLMs)

- Hallucination occurs when LLMs generate false or non-existent information
- Factual hallucination greatly undermines reliability and trustworthiness of LLMs

Is there an airport located at latitude 10.517 and longitude -85.566?

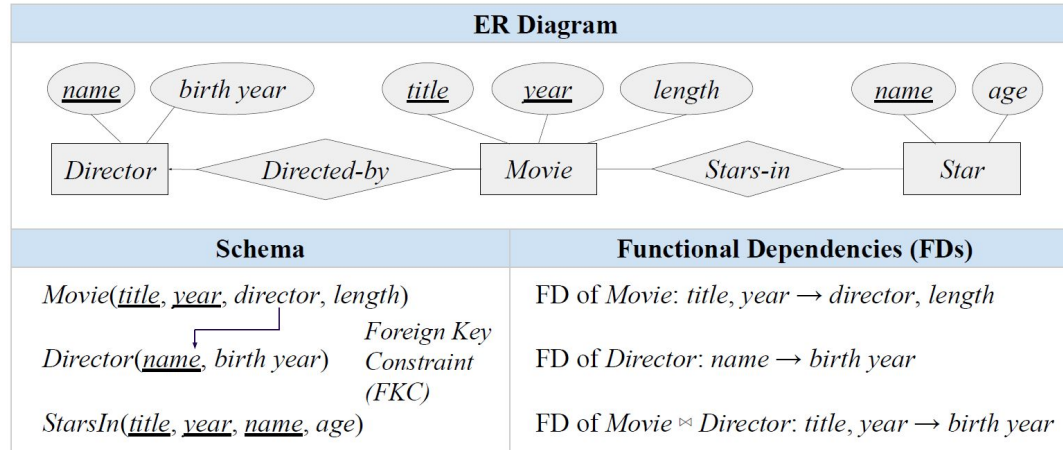
 **Yes.** This location corresponds to the **Nosara Airport** in Costa Rica.

Correct Answer, Incorrect Rationale
Should be **Catsa Airport**

Relational Databases (RDBs)

- RDBs are based on the relational data model and assume a fixed schema
- A fixed schema enables data integrity based on database design theory
- Integrity constraints like FKCs and FDs can be utilized to construct QA pairs



What makes RDBs useful for LLM Benchmarking?

- RDBs (Tables) are everywhere across various domains
 - Anyone can easily create a table of her/his own interest

← basketball

↳ Notebooks 1,971 **Datasets 866 X** ← Comments 711 Topics 332 Competitions 30

Filter by

DATE

- Last 90 days
- This week
- Today

CREATOR

- You
- Others

866 Results

Relevance ▾

Dataset	Downloads	Relevance
College Basketball Dataset Dataset · 7mo ago · by Andrew Sundberg Data from the 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023 Division I college basketball	26,314 downloads	346
NCAA Basketball Dataset · 6y ago · by NCAA Basketball data from as far back as 1894	0 downloads	207
Men's Professional Basketball Dataset · 5y ago · by Open Source Sports Stats on players, teams, and coaches in men's pro basketball leagues, 1937-2012	8,391 downloads	85

kaggle

Google

Geography

Last updated Download format Croissant Usage rights Topic Provider Free

100+ datasets found

Geography Lookup Table proxy.weglot.com opendata.fcc.gov +1more Updated Jan 1, 2024 + more versions	Geography Lookup Table See More Versions Explore at: proxy.weglot.com Data Federal Communicatio... catalog.data.gov 4 scholarly articles cite this dataset (View in Google Scholar) Dataset updated Jan 1, 2024 Dataset provided by Federal Communications Commission Description Summary data of fixed broadband coverage by geographic area. License and Attribution: Broadband data fr subject to copyright restriction. Data and content created by government employees within the scope of the
Data from: COVID-19 Case Surveillance Public Use Data... data.cdc.gov healthdata.gov +6more application/rdf+xml +5 Updated Jul 9, 2024 + more versions	

Google

How does ERBench utilizes RDBs?

- ERBench supports automatic rationale evaluation w/ Functional Dependency (FD)

- **FD** is a relationship between two sets of attributes (X, Y), where X values determine Y values

- Notation: **X -> Y**

- Helps to determine the “*important*” keyword in the model’s rationale.

- Example: **Director, Star, Released Year -> Movie Name**

Director	Star	R-Year	
J. Cameron	L. DiCaprio	1997	



<u>Movie Name</u>
Titanic

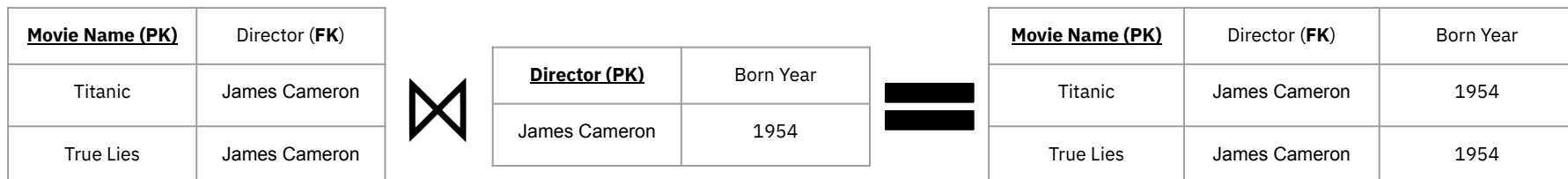
- Example (Q - question, A - model answer)

Q: Is there a movie, released in 1997, starring Leonardo DiCaprio where James Cameron is the director?

A: **Yes** ... The movie is “**Titanic**”.

How does ERBench utilizes RDBs?

- ERBench supports easy multi-hop question generation w/ Foreign Key Constraint
 - Foreign Keys (FKs) are attribute(s) that references the primary key in another table
 - **All you need is a single join operation!**



- Example (Here the right keyword will be the Director - Foreign Key)

Q: Was the director who directed the movie titled Titanic born in the 1950s?

A: **Yes ... James Cameron**, the director of Titanic, was born on August 16, **1954**.

ERBench Question Types

- Binary Questions & Multiple-Choice Questions

Template	FD	Example Question
Binary (Y)	Director, star, released year -> movie title	Is there a movie, released in <1997>, starring <Leonardo DiCaprio>, where <James Cameron> is the director?
Binary (N)	Director, star, released year -> movie title	Is it true that there are no movies, released in <1997>, starring <Leonardo DiCaprio>, where <James Cameron> is the director?
Multiple Choice	Movie title, released year -> director, country of origin, genre	Q: What is the false option about the movie <Titanic> released in year <1997>? Option 1: It was directed by <James Cameron>. Option 2: It was produced in country <USA>. Option 3: It is <animation> movie.

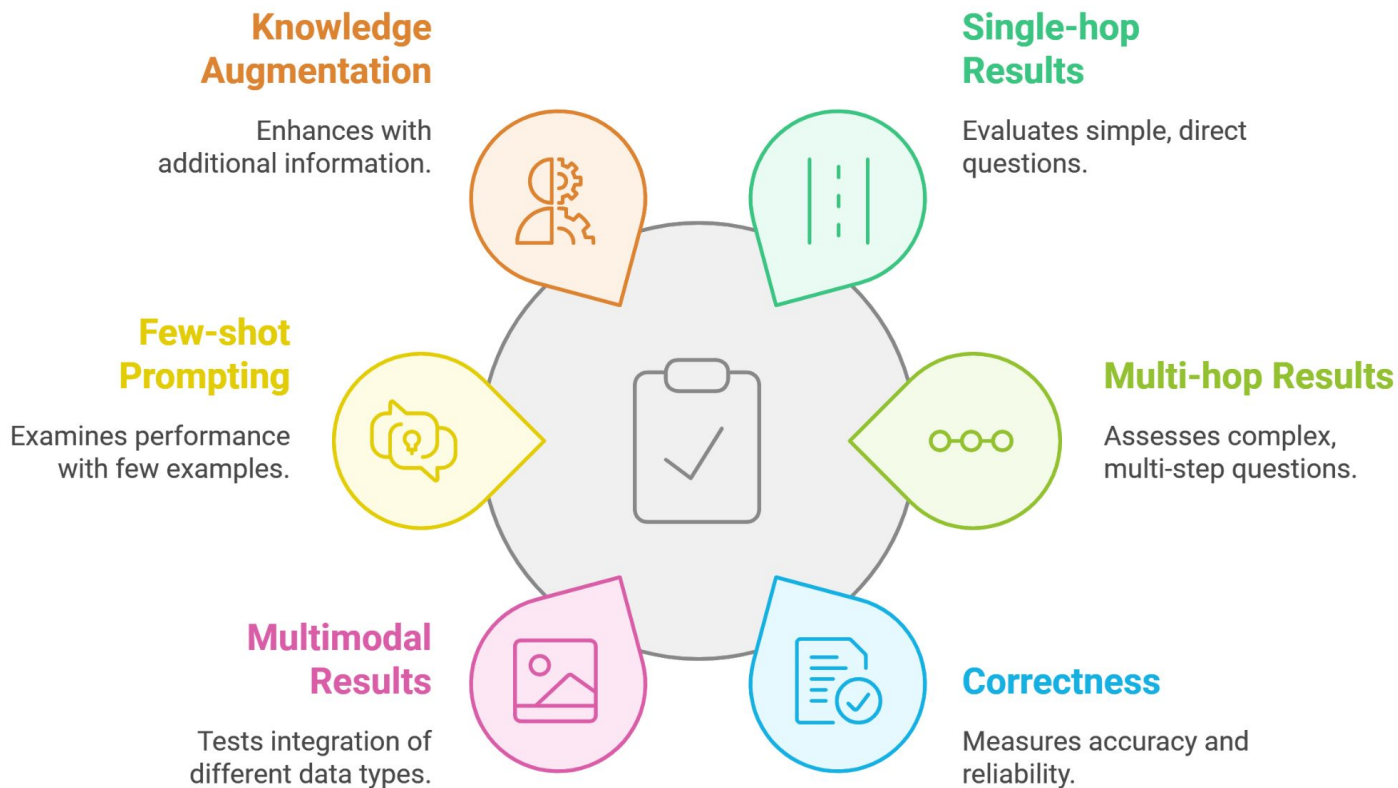
Metrics

	Metric	Model Answer	Model Rationale
Conventional	Answer accuracy (A)	✓	-
	Hallucination rate ^[1] (H)	1 - A - "I am not sure"	
Newly Added	Rationale accuracy ^[*] (R)	-	✓
	Answer-Rationale accuracy (AR)	✓	✓

[1] Sun, Kai, et al. Head-to-tail: How knowledgeable are large language models (llm)? NAACL'24.

[*] Contains keywords derived from FDs

Experiment Results



Single-hop Questions

Model	Metric	Movie			Soccer		
		BN _(Y)	BN _(N)	MC	BN _(Y)	BN _(N)	MC
GPT-4	A	.58	.45	.96	.45	.18	.81
	R	.76	.69	.95	.64	.46	.58
	AR	.57	.44	.95	.38	.16	.57
	H (↓)	.42	.53	.04	.38	.03	.06
Llama2	A	.02	1.0	.85	.01	1.0	.63
	R	.31	.82	.92	.07	.36	.41
	AR	.02	.82	.81	.00	.36	.32
	H (↓)	.98	.00	.15	.99	.00	.37
Mistral	A	.29	1.0	.69	.50	.99	.40
	R	.28	.39	.71	.06	.13	.20
	AR	.13	.39	.59	.04	.13	.17
	H (↓)	.71	.00	.31	.50	.01	.60

(Table cropped for clarity; full results available in the paper)

- LLMs can show answer bias
 - Answer accuracy is much better on BN_(N), while showing similar rationale accuracy.

Single-hop Questions

Model	Metric	Movie			Soccer		
		BN _(Y)	BN _(N)	MC	BN _(Y)	BN _(N)	MC
GPT-4	A	.58	.45	.96	.45	.18	.81
	R	.76	.69	.95	.64	.46	.58
	AR	.57	.44	.95	.38	.16	.57
	H (↓)	.42	.53	.04	.38	.03	.06
Llama2	A	.02	1.0	.85	.01	1.0	.63
	R	.31	.82	.92	.07	.36	.41
	AR	.02	.82	.81	.00	.36	.32
	H (↓)	.98	.00	.15	.99	.00	.37
Mistral	A	.29	1.0	.69	.50	.99	.40
	R	.28	.39	.71	.06	.13	.20
	AR	.13	.39	.59	.04	.13	.17
	H (↓)	.71	.00	.31	.50	.01	.60

(Table cropped for clarity; full results available in the paper)

- LLMs can show answer bias
 - Answer accuracy is much better on BN_(N), while showing similar rationale accuracy.
- LLMs struggle to provide both the correct answer and rationale
 - **AR** is often lower than **A**

Single-hop Questions

Model	Metric	Movie			Soccer		
		BN _(Y)	BN _(N)	MC	BN _(Y)	BN _(N)	MC
GPT-4	A	.58	.45	.96	.45	.18	.81
	R	.76	.69	.95	.64	.46	.58
	AR	.57	.44	.95	.38	.16	.57
	H (↓)	.42	.53	.04	.38	.03	.06
Llama2	A	.02	1.0	.85	.01	1.0	.63
	R	.31	.82	.92	.07	.36	.41
	AR	.02	.82	.81	.00	.36	.32
	H (↓)	.98	.00	.15	.99	.00	.37
Mistral	A	.29	1.0	.69	.50	.99	.40
	R	.28	.39	.71	.06	.13	.20
	AR	.13	.39	.59	.04	.13	.17
	H (↓)	.71	.00	.31	.50	.01	.60

(Table cropped for clarity; full results available in the paper)

- LLMs can show answer bias
 - Answer accuracy is much better on BN_(N), while showing similar rationale accuracy.
- LLMs struggle to provide both the correct answer and rationale
 - **AR** is often lower than **A**



Focusing solely on answers is insufficient; Considering rationale is important

Multi-hop Questions

- Correctness in later hops often highly depends on earlier hops
 - $\Pr(r_{i+1}|r_i)$: Previous hop is correct, and the following hop is also correct

Model	Metric	Movie & Director				Soccer & Olympic			
		w/o CoT		w/ CoT		w/o CoT		w/ CoT	
		BN _(Y)	BN _(N)	BN _(Y)	BN _(N)	BN _(Y)	BN _(N)	BN _(Y)	BN _(N)
GPT-3.5	$\Pr(r_{i+1} r_i)$.95	.96	.93	.95	1.0/.57	.95/.79	1.0/1.0	1.0/.94
	$\Pr(r_{i+1} \neg r_i)$.04	.03	.06	.00	.10/.00	.15/.00	.33/.04	.31/.02
GPT-4	$\Pr(r_{i+1} r_i)$.96	.96	.97	.95	.99/.96	.99/.97	1.0/.98	1.0/.97
	$\Pr(r_{i+1} \neg r_i)$	n/a	n/a	n/a	n/a	.35/.00	.27/.00	.52/.00	.55/.00
Llama2	$\Pr(r_{i+1} r_i)$.92	.93	.97	.92	1.0/.85	1.0/.81	1.0/.95	1.0/.93
	$\Pr(r_{i+1} \neg r_i)$.00	.00	.00	.00	.35/.00	.18/.00	.24/.02	.20/.04

(Table cropped for clarity; full results available in the paper)

Multi-hop Questions

- Correctness in later hops often highly depends on earlier hops
 - $\Pr(r_{i+1}|r_i)$: Previous hop is correct and the following hop is also correct

Model	Metric	Movie & Director				Soccer & Olympic			
		w/o CoT		w/ CoT		w/o CoT		w/ CoT	
		BN _(Y)	BN _(N)	BN _(Y)	BN _(N)	BN _(Y)	BN _(N)	BN _(Y)	BN _(N)
GPT-3.5	$\Pr(r_{i+1} r_i)$.95	.96	.93	.95	1.0/.57	.95/.79	1.0/1.0	1.0/.94
	$\Pr(r_{i+1} \neg r_i)$.04	.03	.06	.00	.10/.00	.15/.00	.33/.04	.31/.02
GPT-4	$\Pr(r_{i+1} r_i)$.96	.96	.97	.95	.99/.96	.99/.97	1.0/.98	1.0/.97
	$\Pr(r_{i+1} \neg r_i)$	n/a	n/a	n/a	n/a	.35/.00	.27/.00	.52/.00	.55/.00
Llama2	$\Pr(r_{i+1} r_i)$.92	.93	.97	.92	1.0/.85	1.0/.81	1.0/.95	1.0/.93
	$\Pr(r_{i+1} \neg r_i)$.00	.00	.00	.00	.35/.00	.18/.00	.24/.02	.20/.04

(Table cropped for clarity; full results available in the paper)



ERBench enables new kind of analyses (e.g., step-by-step debugging)

Takeaways

- ERBench can be a powerful tool for evaluating LLM hallucinations
 - Key idea: use relational databases and design theory
- ERBench provides unique benefits compared to existing benchmarks
 - Automatic rationale verification using Functional Dependencies
 - Easy multi-hop question generation using Foreign Key Constraints
 - High usability by allowing everyone to convert their own databases

Thank you 😊
Please come to our **poster session** for more details!

Appendix

More details and Future Improvements

Additional Techniques

- Chain-of-Thought
 - Designed demonstration prompts to have balanced distribution w.r.t. the answer type (e.g. Yes/No)
- Fine-tuning
 - Finetuned GPT-3.5 with 3,000 Q&A pairs for each dataset

PROMPTS FOR MOVIE DATASET

Q: Is there a movie, released in 1975, starring Al Pacino where Sidney Lumet is the director?

A: Yes, the movie, among various movies Al Pacino starred in 1975, Sidney Lumet directed the movie "Dog Day Afternoon".

Q: Is it true that there are no movies, released in 2005, starring Ben Affleck where Kevin Smith is the director?

A: Yes, it is true. However, there is a movie, "Chasing Amy" that Kevin Smith directed, where Ben Affleck starred in, which was released in 1997.

Q: Is it true that there are no movies, released in 2019, starring Kangho Song where Bong Joon-ho is the director?

A: No, it is not true. The movie "Parasite", which was released in 2019 was directed by Bong Joon-ho, where Kangho Song starred in the movie.

Q: Is there a movie, released in 1997, starring Jason Biggs where Paul Weitz is the director?

A: No, however there is a movie, "American Pie" that Paul Weitz directed, where Jason Biggs starred in, which was released in 1999.

Q: Is there a movie, released in 1979, starring George Lazenby where Peter R. Hunt is the director?

A: No, however there is movie, "On Her Majesty's Secret Service" that George Lazenby starred in, where Peter R. Hunt is the director, which was released in 1969.

Q: Is it true that there are no movies, released in 2005, starring Leonardo DiCaprio where Martin Scorsese is the director?

A: Yes, it is true. However, there is a movie, "Gangs of New York" that Martin Scorsese directed and Leonardo DiCaprio starred in, which was released in 2002.

Q: Is it true that there are no movies, released in 1952, starring Robert Taylor where Richard Thorpe is the director?

A: No, it is not true. The movie, "Ivanhoe" that Richard Thorpe directed, where Robert Taylor appeared in the movie was released in 1952.

Q: Is there a movie, released in 1982, starring Dustin Hoffman where Sydney Pollack is the director?

A: Yes, the movie, among various movies that Sydney Pollack directed in 1982, Dustin Hoffman starred in the movie, "Tootsie".

Metrics

- *Answer Accuracy* (**A**) [11]: Portion of LLM responses that are correct.
- *Rationale Accuracy* (**R**): Portion of responses whose rationales contain the FD-inferred values.
- *Answer-Rationale Accuracy* (**AR**): Portion of responses that are not only correct, but also contain FD-inferred values in their rationales.
- *Hallucination Rate* (**H**) [11]: Portion of responses that are incorrect, excluding those where LLMs admit uncertainty in their responses (e.g., *Unsure*). Specifically, $\mathbf{H} = 1 - \mathbf{A} - \mathbf{M}$, where **M** denotes the percentage of LLM responses that admit they cannot answer the given question (i.e., *missing rate*). A lower **H** value is better.

Experiment Results

New Findings

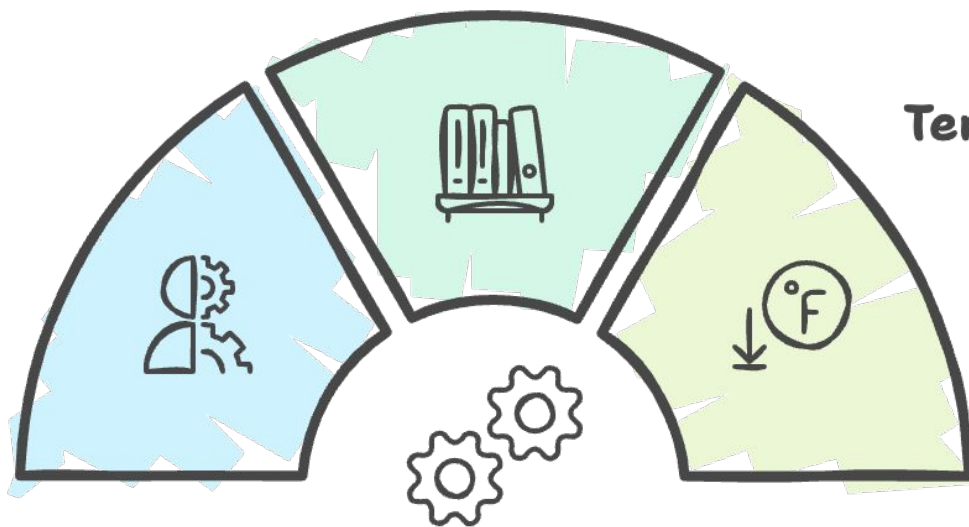
Experimental Setups

5 Datasets

Movie, Soccer, Airport,
Music, Book

6 LLMs

GPT-3.5 / GPT-4 /
LLama2-70B-Chat /
Gemini-Pro /
Claude-3-Sonnet /
Mistral-7B-Instruct



Temperature = 0

LLMs become
deterministic

Single-hop Questions

Model	Metric	Movie			Soccer		
		BN _(Y)	BN _(N)	MC	BN _(Y)	BN _(N)	MC
GPT-4	A	.58	.45	.96	.45	.18	.81
	R	.76	.69	.95	.64	.46	.58
	AR	.57	.44	.95	.38	.16	.57
	H (↓)	.42	.53	.04	.38	.03	.06
Llama2	A	.02	1.0	.85	.01	1.0	.63
	R	.31	.82	.92	.07	.36	.41
	AR	.02	.82	.81	.00	.36	.32
	H (↓)	.98	.00	.15	.99	.00	.37
Mistral	A	.29	1.0	.69	.50	.99	.40
	R	.28	.39	.71	.06	.13	.20
	AR	.13	.39	.59	.04	.13	.17
	H (↓)	.71	.00	.31	.50	.01	.60

(zoomed in)

- LLMs can show answer bias
 - While showing similar rationale accuracy, answer accuracy is much better on BN_(N)



Focusing solely on answers is insufficient

- Providing both the correct answer and rationale can be challenging
 - **AR** is often lower than **A**



Considering rationale is important

Multi-hop Questions

- Subsequent hops can be **highly dependent** on the previous hop
 - e.g., $\Pr(r_{i+1}|r_i)$: Previous hop is correct and the following hop is also correct

Model	Metric	Movie & Director				Soccer & Olympic			
		w/o CoT		w/ CoT		w/o CoT		w/ CoT	
		BN _(Y)	BN _(N)	BN _(Y)	BN _(N)	BN _(Y)	BN _(N)	BN _(Y)	BN _(N)
GPT-3.5	$\Pr(r_{i+1} r_i)$.95	.96	.93	.95	1.0/.57	.95/.79	1.0/1.0	1.0/.94
	$\Pr(r_{i+1} \neg r_i)$.04	.03	.06	.00	.10/.00	.15/.00	.33/.04	.31/.02
GPT-4	$\Pr(r_{i+1} r_i)$.96	.96	.97	.95	.99/.96	.99/.97	1.0/.98	1.0/.97
	$\Pr(r_{i+1} \neg r_i)$	n/a	n/a	n/a	n/a	.35/.00	.27/.00	.52/.00	.55/.00
Llama2	$\Pr(r_{i+1} r_i)$.92	.93	.97	.92	1.0/.85	1.0/.81	1.0/.95	1.0/.93
	$\Pr(r_{i+1} \neg r_i)$.00	.00	.00	.00	.35/.00	.18/.00	.24/.02	.20/.04

(zoomed in)



ERBench enables new kind of analyses (e.g., step-by-step debugging)

Single-hop Questions

Model	Metric	Movie			Soccer		
		BN _(Y)	BN _(N)	MC	BN _(Y)	BN _(N)	MC
GPT-4	A	.58	.45	.96	.45	.18	.81
	R	.76	.69	.95	.64	.46	.58
	AR	.57	.44	.95	.38	.16	.57
	H (↓)	.42	.53	.04	.38	.03	.06
Llama2	A	.02	1.0	.85	.01	1.0	.63
	R	.31	.82	.92	.07	.36	.41
	AR	.02	.82	.81	.00	.36	.32
	H (↓)	.98	.00	.15	.99	.00	.37
Mistral	A	.29	1.0	.69	.50	.99	.40
	R	.28	.39	.71	.06	.13	.20
	AR	.13	.39	.59	.04	.13	.17
	H (↓)	.71	.00	.31	.50	.01	.60

(zoomed in)

- LLMs can show answer bias
 - Answer accuracy is much better on BN_(N), while showing similar rationale accuracy.

Single-hop Questions

Model	Metric	Movie			Soccer		
		BN _(Y)	BN _(N)	MC	BN _(Y)	BN _(N)	MC
GPT-4	A	.58	.45	.96	.45	.18	.81
	R	.76	.69	.95	.64	.46	.58
	AR	.57	.44	.95	.38	.16	.57
	H (↓)	.42	.53	.04	.38	.03	.06
Llama2	A	.02	1.0	.85	.01	1.0	.63
	R	.31	.82	.92	.07	.36	.41
	AR	.02	.82	.81	.00	.36	.32
	H (↓)	.98	.00	.15	.99	.00	.37
Mistral	A	.29	1.0	.69	.50	.99	.40
	R	.28	.39	.71	.06	.13	.20
	AR	.13	.39	.59	.04	.13	.17
	H (↓)	.71	.00	.31	.50	.01	.60

(zoomed in)

- LLMs struggle to provide both the correct answer and rationale
 - **AR** is often lower than **A**

Model	Metric	Movie			Soccer		
		BN _(Y)	BN _(N)	MC	BN _(Y)	BN _(N)	MC
GPT-4	A	.58	.45	.96	.45	.18	.81
	R	.76	.69	.95	.64	.46	.58
	AR	.57	.44	.95	.38	.16	.57
	H (↓)	.42	.53	.04	.38	.03	.06
Llama2	A	.02	1.0	.85	.01	1.0	.63
	R	.31	.82	.92	.07	.36	.41
	AR	.02	.82	.81	.00	.36	.32
	H (↓)	.98	.00	.15	.99	.00	.37
Mistral	A	.29	1.0	.69	.50	.99	.40
	R	.28	.39	.71	.06	.13	.20
	AR	.13	.39	.59	.04	.13	.17
	H (↓)	.71	.00	.31	.50	.01	.60