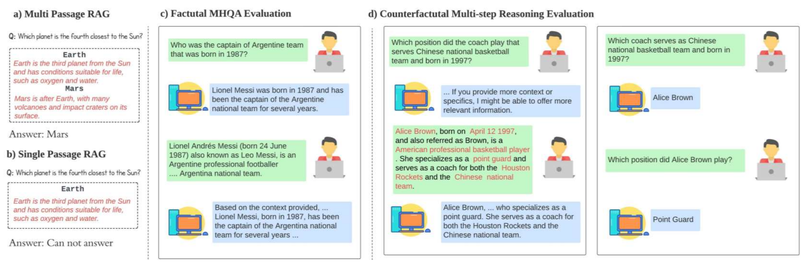
CofCA: A Step-Wise Counterfactual Multi-hop QA benchmark

Jian Wu *, Linyi Yang*, Zhen Wang, Manabu Okumura, Yue Zhang

* Equal contribution, wu.j.as@m.titech.ac.jp

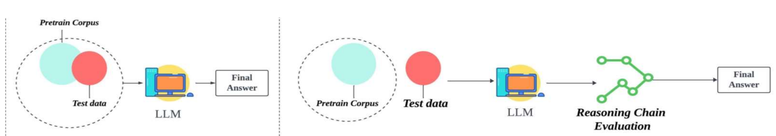
Overview of CofCA

LLMs sometimes generate answers that rely on internal memory rather than retrieving evidence and reasoning in the given context, which brings concerns about the evaluation quality of real reasoning abilities.



In this paper, CofCA makes several key contributions: **(1)** the first introduce **counterfactual data** in to multi-step reasoning evaluation; **(2) Sub-QA annotation** for reasoning chain evaluation; (3) **an automated annotation framework with human in the loop**; and (4) **extensive experiments on Wikipedia-based benchmarks and CofCA**, demonstrating an obvious performance gap.

CofCA Evaluation Task



Benchmark characteristics:

1) **No overlap** with existence knowledge

2) Human Annotated **Sub-Questions** and **Counterfactual Context**

3) Increased **Difficulty Levels**

4) Reasoning Chain Evaluation by **Sub-QA**

Open-source Availability:

We are open to discussing about any questions on CofCA, please contact:

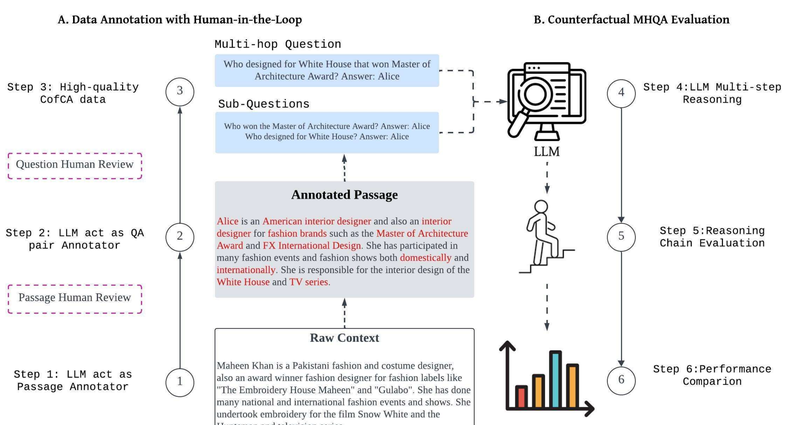
wu.j.as@m.titech.ac.jp

Comparison to Existing Benchmarks

Benchmarks	Data Type	Data Source	Task	Reasoning Chain
HotpotQA	Factual	Wikipedia	Multi-hop/final-QA	✗
2WikiMultihopQA	Factual	Wikipedia	Multi-hop/final-QA	✗
MusiQue	Factual	Wikipedia	Multi-hop/final-QA	✗
DisentQA	Counterfactual & Factual	Natural Questions	Single-hop/final-QA	✗
IfQA	Counterfactual & Factual	Crowdsourcing	Open Domain/final-QA	✗
CofCA(Ours)	Counterfactual & Factual	Rewriting Wikipedia	Multi-hop/final-QA/sub-qa	✓

Automatic Data Annotation Framework with Human in the Loop

The automatic annotation framework is 1) replace the **noun phrases**, **synonym** and **named entities** of the given passage; 2) **translation back** (Eng-to-CN and CN-to-Eng); 3) generate **new multi-hop questions** to fit the written passages; 4) Check **grammar issues** and **answerabilities** of the generated question.



Stage	Passages
HotpotQA	Elliott Lester is an English film and television director, best known for directing the film "Blitz". He made his directing debut in 2006 with "Love Is the Drug", and his latest film, "Aftermath", was released on April 4, 2017.
Stage 1	Marvin Kellaway is a British movie and radio helmsman, foremost noted for helming the movie 'Flashstrike'. He commenced his helming genesis in 2007 with 'Affection Is the Potion', and his most recent movie, 'Repercussion', was unveiled on May 3, 2018.
Stage 2	Marvin Kellaway, a prominent British director in both film and radio, is best known for directing the popular movie 'Flashstrike'. He began his directing career in 2007 with 'Affection Is the Potion', and his latest film, 'Repercussion', was released on May 3, 2018.

Different productions in different rewriting stages.

Evaluation Metrics

Counterfactual Multi-hop QA Evaluation: F1, EM and LLM-as-Judge Partial Match

Reasoning Chain Evaluation: Calculate the proportion of different reasoning chains.

Experimental Results

Performance of LLMs on Wikipedia-based multi-hop QA datasets. The performance is measured by EM and F1 scores with a zero-shot setting. PM† indicates the partial match of LLMs' outputs evaluated with GPT-4-turbo with the same prompt.

Datasets	Wikipedia								
	HotpotQA			2Wiki			MuSiQue		
Metrics	EM	F1	PM †	EM	F1	PM †	EM	F1	PM †
Proprietary LLMs									
GPT-4	69.9±1.5	82.3±1.3	74.8 ±1.2	59.7±1.4	67.4±2.7	64.8±0.9	57.3±1.9	65.4±2.9	63.9±1.4
GPT-3.5	58.6±0.9	69.1±1.1	62.8±0.7	56.3±0.9	67.6±0.8	59.4±0.9	49.3±0.8	63.2±1.5	53.1±0.6
GEMINI-pro	58.2±1.3	68.4±1.3	63.5±0.9	48.5±1.6	58.5±0.9	54.7±1.2	41.3±1.5	54.5±0.7	46.9±1.3
text	50.3±0.9	61.4±0.8	54.9±0.8	42.3±1.4	53.9±1.5	46.7±0.7	40.2±0.9	51.0±1.5	44.6±0.4
Bing Chat	68.1±0.6	78.3±1.2	72.1±1.2	58.9±0.5	69.9±0.5	63.4±0.8	49.6±1.1	64.1±0.8	52.3±0.7
O1-preview	72.2 ±0.6	82.7 ±0.7	76.9 ±1.1	68.7 ±0.9	79.8 ±0.8	72.3 ±1.2	63.9 ±0.9	72.4 ±0.5	67.9±0.8
Open Source LLMs									
Llama 2-7b	34.5±1.2	41.3±1.1	38.5±0.3	30.6±1.1	34.7±1.1	33.8±0.9	31.7±0.8	35.6±1.2	34.2±0.9
Mistral-7b	30.6±1.5	37.2±1.4	34.9±0.5	27.4±0.6	29.8±0.9	31.4±0.8	25.2±0.7	28.9±0.8	29.2±0.7
Qwen 2-7b	36.2±1.5	43.5±1.3	39.3±0.4	31.7±1.0	35.8±0.8	36.8±0.5	28.2±1.1	31.2±1.2	33.5±0.4

LLMs show a performance gap between the **Wikipedia-based datasets** and **CofCA**. Taking **GPT-4** as an example, GPT-4 achieves high EM and F1 scores (69.9 and 82.3, respectively), while in CofCA, EM and F1 scores are **sharply declined** (53.1 and 62.8).

Datasets	CofCA								
	2-hop			3-hop			4-hop		
Metrics	EM	F1	PM †	EM	F1	PM †	EM	F1	PM †
Proprietary LLMs									
GPT-4	53.1±1.5	62.8±1.3	57.6±1.1	44.5±1.3	56.4±1.7	49.5±1.2	42.3±0.6	53.5±0.9	48.8±1.1
GPT-3.5	40.6±0.7	56.7±0.5	43.7±0.9	37.7±0.5	50.9±1.1	42.1±1.3	32.5±1.2	44.6±0.8	36.2±1.1
GEMINI-pro	35.0±0.7	45.3±1.6	38.2±0.8	29.6±0.5	42.7±0.9	31.9±0.6	26.1±1.1	35.3±1.2	29.8±1.1
text	32.6±0.9	48.5±0.8	37.4±0.9	27.8±0.9	46.3±0.8	33.5±1.2	24.8±0.8	44.1±0.9	27.6±0.7
Bing Chat	41.9±0.8	53.4±0.9	45.4±0.9	39.6±1.1	49.4±1.2	43.7±0.7	30.7±0.9	42.2±0.7	35.6±0.7
O1-preview	59.4 ±0.4	68.5±0.3	63.9±0.8	52.3 ±0.5	66.1±0.4	58.4±0.7	50.7 ±0.4	62.3±0.6	53.2±0.7
Open Source LLMs									
Llama 2-7b	26.1±0.9	34.3±1.2	30.5±0.4	22.6±1.1	26.7±1.3	25.8±0.8	24.9±1.2	28.7±1.1	30.5±0.9
Mistral-7b	24.7±0.9	29.5±0.7	28.5±0.3	20.8±1.1	25.3±1.1	24.8±0.7	18.6±1.1	22.2±1.3	23.5±0.5
Qwen 2-7b	30.8±1.1	38.2±1.4	34.1±0.4	27.2±1.1	31.7±1.3	31.8±0.4	25.2±0.8	28.6±0.9	29.2±0.5

The performance on **reasoning chain** tasks. A 2-hop question which requires 2 times reasoning, having 8 different reasoning chains. **Red rows: wrong reasoning chain, wrong intermediate steps but correct final answer; Green row: correct reasoning chain and final answer.**

q_{sub1}	q_{sub2}	q	O1-preview	GPT-4	GPT-3.5	GEMINI-pro	text	Bing Chat	GPT-4o
c	c	c	45.2	36.3	13.3	15.0	17.3	28.3	44.5
c	c	w	9.8	12.3	9.3	9.0	10.7	7.7	10.1
c	w	c	3.2	2.0	6.7	5.3	7.7	6.0	2.7
c	w	w	15.4	25.3	24.3	14.7	25.0	16.3	12.7
w	c	c	4.9	5.7	3.7	5.3	6.7	2.3	5.8
w	c	w	4.1	3.7	3.7	5.3	3.7	3.0	4.7
w	w	c	0.8	0.3	7.3	13.3	8.7	5. 0	1.4
w	w	w	16.6	14.3	31.7	32.3	30.3	31.3	18.1

Inflated performance based on **wrong reasoning chain**.

Key Insights

Our findings highlight that, although LLMs performed relatively well on QA tasks, the performance dropped on multi-hop questions that were based on **new, counterfactual knowledge**. In addition, their high performances are inflated and benefit from a **high proportion of incorrect reasoning chains**.