

Empowering Language Models with Active Inquiry for Deeper Understanding

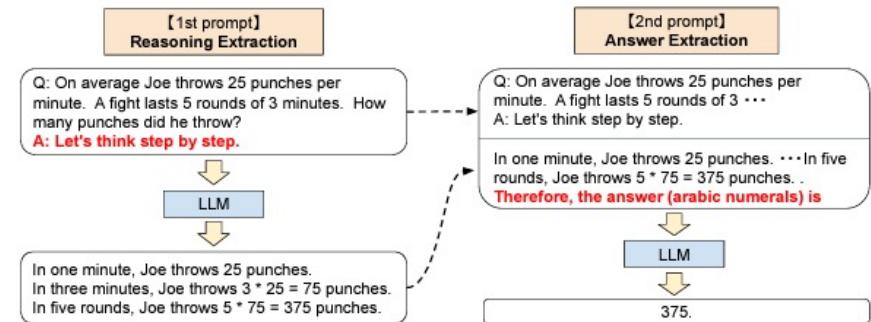
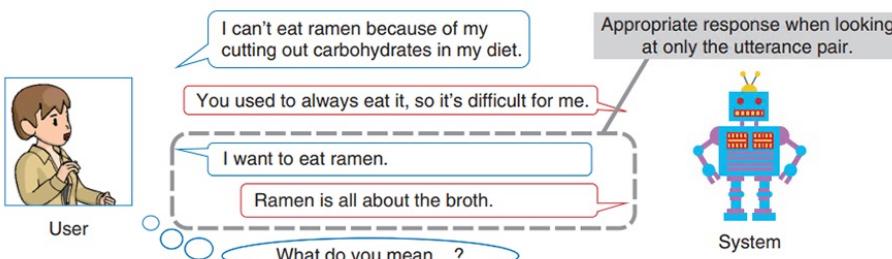
Jing-Cheng Pang^{1,3,*}, Heng-Bo Fan^{2,*}, Pengyuan Wang^{1,3,*}, Jia-Hao Xiao^{2,*}, Nan Tang^{1,3},
Si-Hang Yang^{1,3}, Chengxing Jia^{1,3}, Sheng-Jun Huang^{2,△} Yang Yu^{1,3}

March 19th, 2024



Background

LLMs demonstration proficiency in solving textual problems



Dialog

$$\lim_{x \rightarrow 1} \frac{e^{9x} - e}{2\ln x - 3} \quad P = r^2 \pi \quad \ln = \sqrt{ax^b}$$

$$B \sum_{i=0}^{n-1} h_i = \int_0^n y dx \quad e = 2,79$$

$$y = 2x^2 + 3x \quad P = \frac{8}{3} \pi$$

$$A - C = \frac{x^2}{C}$$

$$(x+y)^2 = \left(\frac{y}{2}\right)^2 = x^2 + 2ax + a^2$$

$$\tan(2a) - \frac{2\tan(a)}{1-\tan^2(a)} \quad 15 \Delta t = T - \frac{3a}{x}$$

$$+ y^2 = z \quad \frac{\Delta x}{\Delta y} = \lim_{n \rightarrow \infty} \frac{\Delta x + 2}{\Delta y - 1}$$

$$X_{1/2} = \frac{b \pm \sqrt{a-c}}{2}$$

Math

Reasoning

Input Article

Marseille, France (CNN) The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane. Marseille prosecutor Brice Robin told CNN that "so far no videos were used in the crash investigation." He added, "A person who has such a video needs to immediately give it to the investigators." Robin's comments follow claims by two magazines, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings Flight 9525 as it crashed into the French Alps. All 150 on board were killed. Paris Match and Bild reported that the video was recovered from a phone at the wreckage site. ...

Abstractive summarization

Generated summary
Prosecutor : " So far no videos were used in the crash investigation "

Text Summarization Models

Extractive summary
marseille prosecutor brice robin told cnn that " so far no videos were used in the crash investigation . " robin 's comments follow claims by two magazines , german daily bild and french paris match , of a cell phone video showing the harrowing final seconds from on board germanwings flight 9525 as it crashed into the french alps . paris match and bild reported that the video was recovered from a phone at the wreckage site .

Reading comprehension

Background

LLMs often misinterpret user queries and answer the questions directly.



How many 'm's are in the word 'Weather'?



There is one 'm' in the word 'Weather'.



Are you sure?



Apologies for the oversight. You're right; there are no 'm's in the word 'Weather'. Thank you for pointing it out.

Background

LLMs often misinterpret user queries and answer the questions directly.

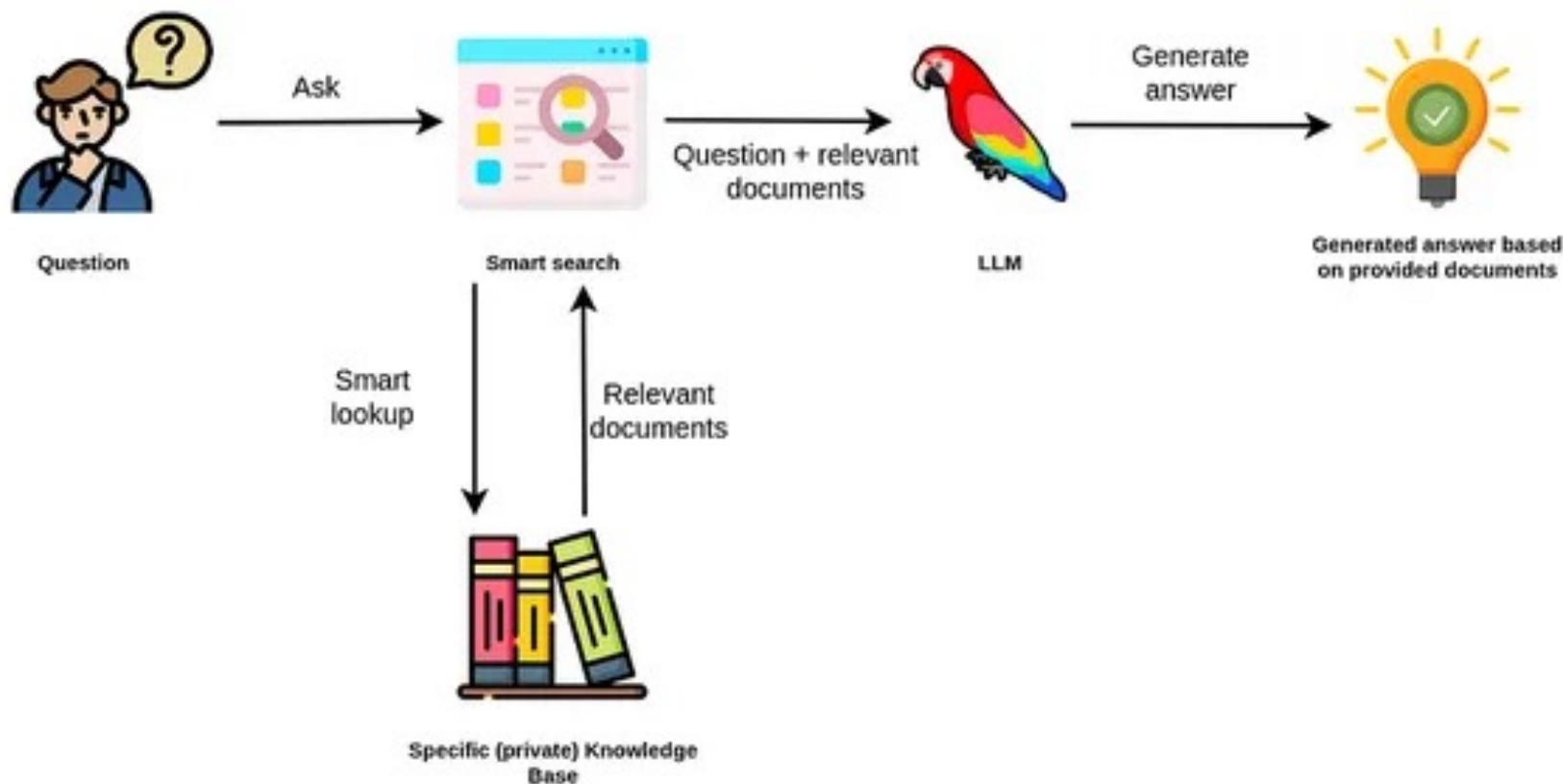


LLM hallucinations, also known as Late-Life Migraine Accompaniments, are a specific type of hallucination that can occur in individuals experiencing migraines, typically later in life. These hallucinations are visual in nature and can be quite vivid and intricate. While LLM hallucinations are most commonly associated with migraines, they can also be linked to other medical conditions, such as Charles Bonnet Syndrome or certain neurological disorders.



Related work

To supply information for LLMs...



Related work

To supply information for LLMs...

Asking clarifying questions

Examples of clarifying questions

what county is that?

are you enrolled at Keiser?

Well, you need to use your local library, so where do you live?

What library are you referring to?

Are you looking for books in your local library?

Who was the mayor of what city?

I can find you a tel # or e-mail for Orange County, would that help?

Would you like me to show you their website and we can take a look?

Do you have the titles of either book?

Where have you already looked?

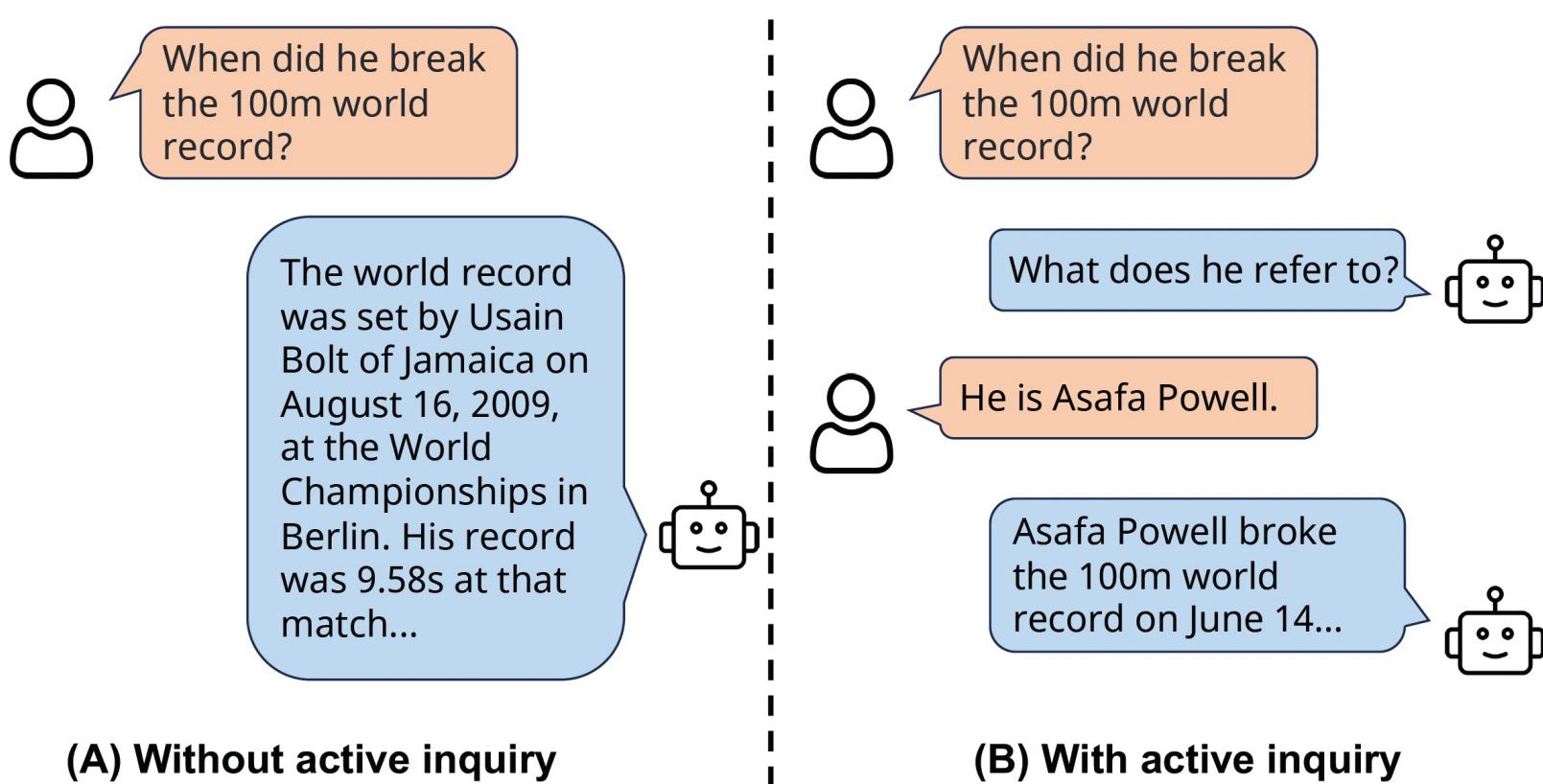
Did you ask your grandmother? If she lives there, she might know.

Can I help you with anything else today?

Is it helpful at all?

Background

Human can actively inquire for information for clarification.



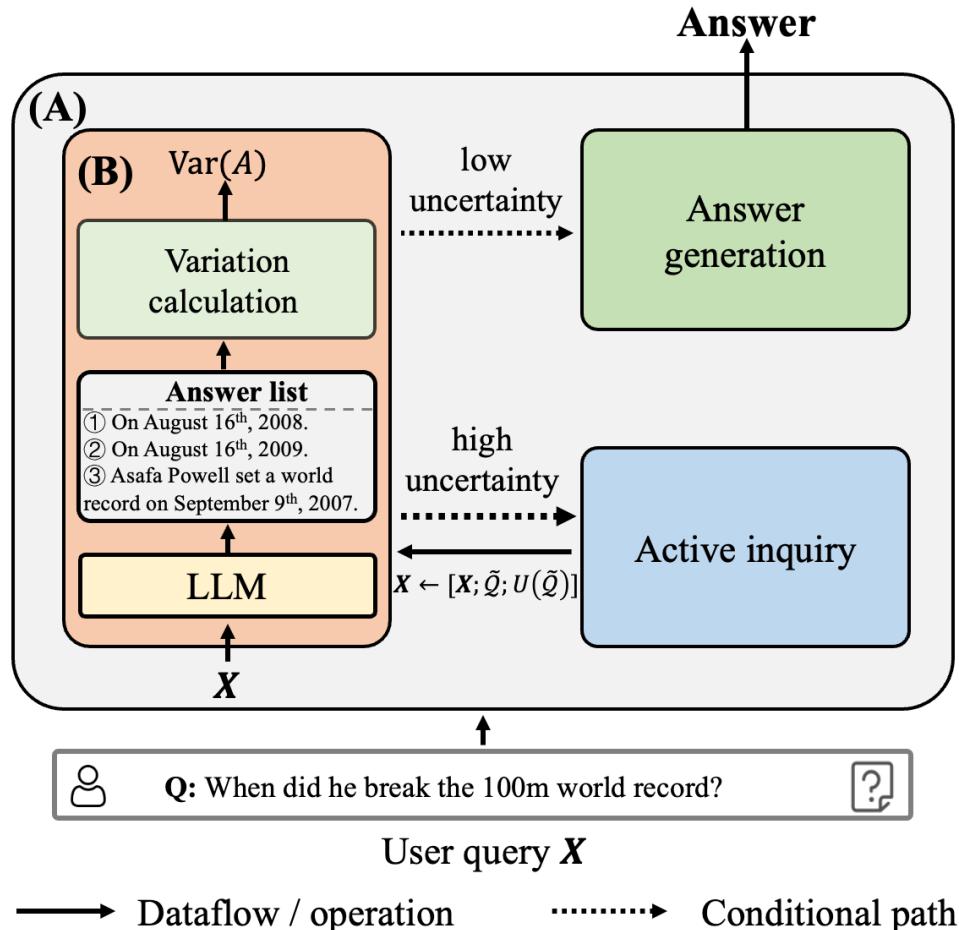
Our method: LaMAI

Empower LLMs with the ability of active inquiry:

1. When should LLMs actively inquire?
2. What questions do LLMs ask?
3. How to leverage user feedback?

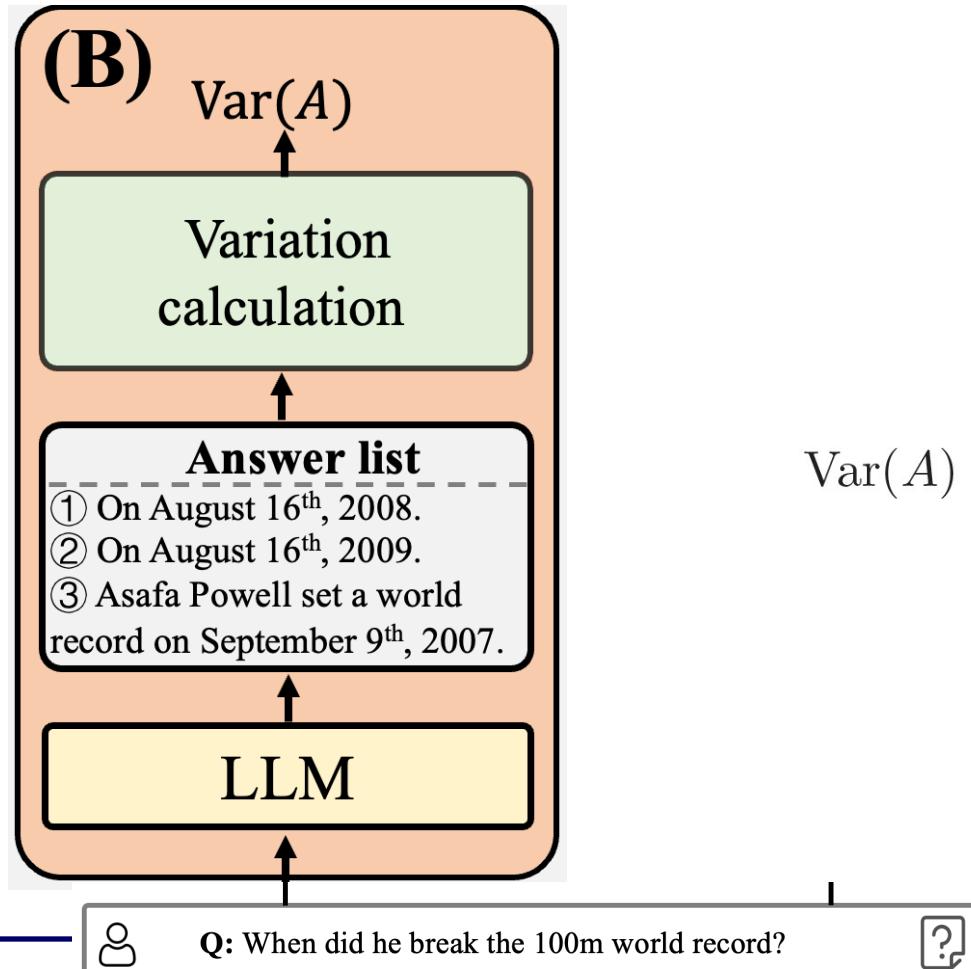
Our method: LaMAI

Empower LLMs with the ability of active inquiry



Uncertainty estimation

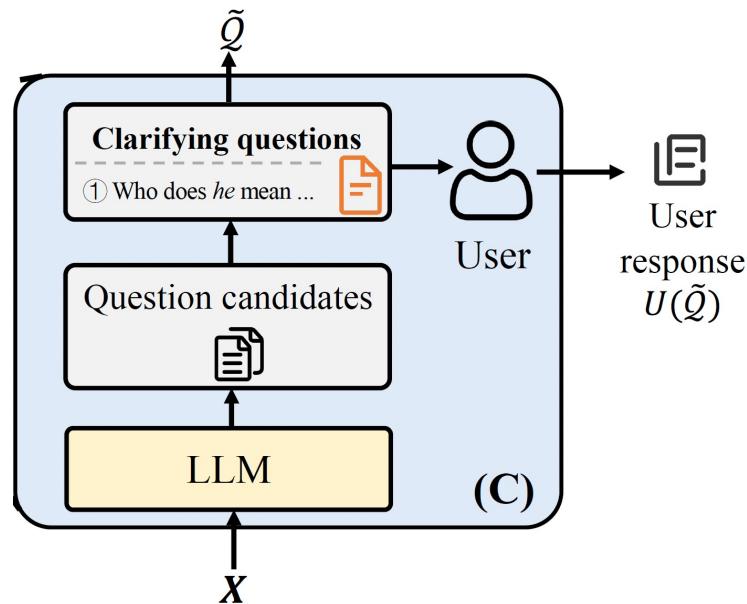
Estimate LLM's uncertainty about an user query X:



$$\text{Var}(A) = \frac{1}{K} \sum_{k=1}^K \left(\frac{1}{T-1} \sum_{i=1}^T (E_i^k - \bar{E}^k)^2 \right)$$

Active inquiry

Actively ask user with clarifying questions

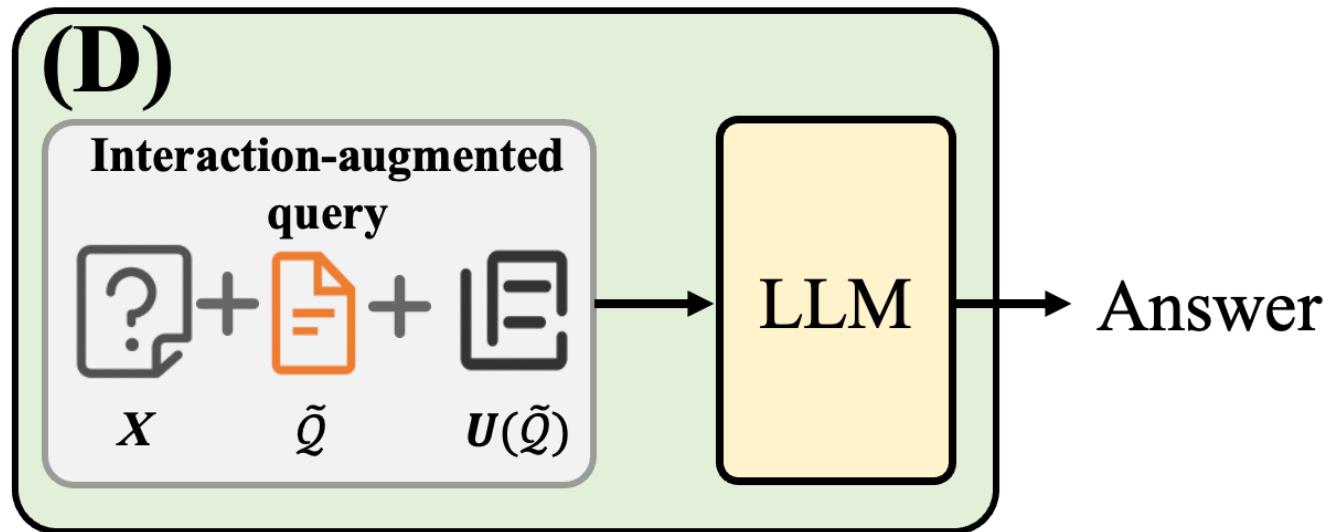


Active learning for question selection:

1. Similarity
2. Diversity
3. Random

Answer generation

Answer generation with augmented query:



Original user query.

Hints:

Q1: Do you mean that...

A1: Yes...

Experiment setting

Dataset: HotpotQA, StrategyQA, 2WikiMultiHopQA, MuSiQue, IIRC, QMSum. Questions with context, facts or knowledge.

	HotpotQA	StrategyQA
User query	Were Up and The Watercolor released in the same year?	Are more people today related to Genghis Khan than Julius Caesar?
Supporting facts	Up and The Watercolor are two films. Up was released in ...	Compare the number of their offspring. Julius Caesar had three children. Genghis Khan had sixteen children ...
Label	Yes	True

Experiment setting

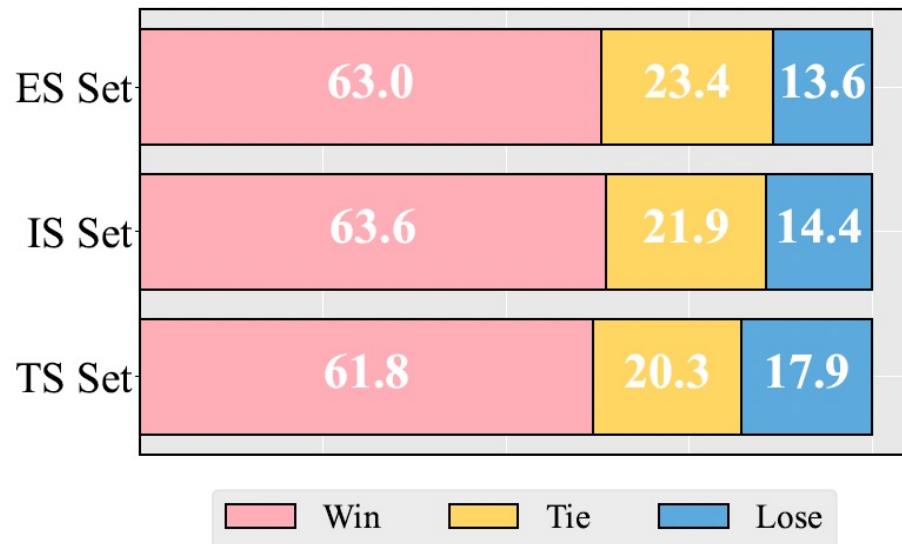
User response: GPT-4 (access to supporting facts), as the pseudo-human interlocutor.

Baselines: DG, CoT, Self-ask, RAG (web), Oracle

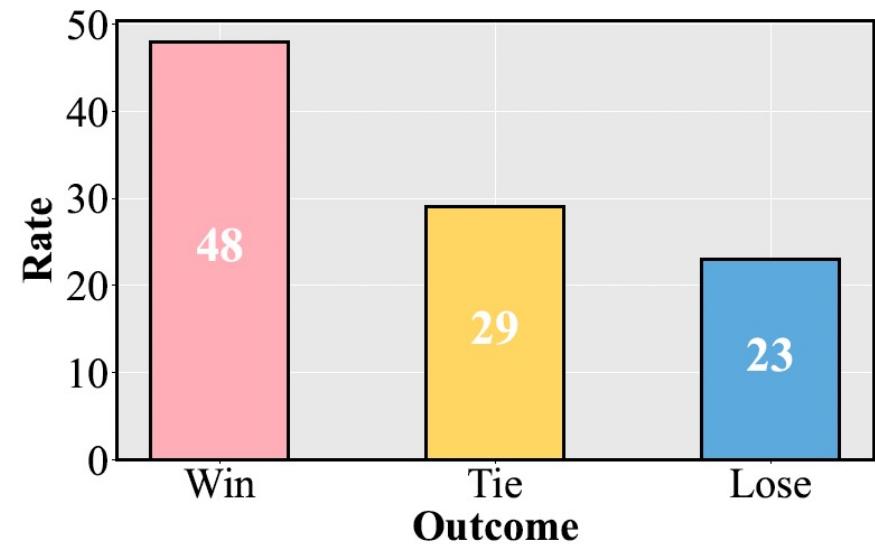
Results

Method \ Dataset	HotpotQA			StrategyQA			2WikiMultiHopQA			MuSiQue			IIRC		
	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc
DG	29.1	38.3	44.4	57.2	57.6	57.6	17.3	22.9	43.3	3.8	14.0	20.5	14.7	18.1	20.8
CoT	32.6	41.6	48.1	66.7	66.9	66.9	31.0	32.9	49.6	5.5	12.8	19.8	17.5	22.0	24.6
Self-ask	27.7	38.0	58.6	34.3	63.2	34.3	42.5	49.2	54.0	15.0	27.0	28.5	10.1	30.1	16.8
RAG (web)	16.2	25.4	47.9	33.1	63.3	33.1	28.7	36.9	51.1	10.0	20.2	27.8	6.0	10.7	25.6
LaMAI	47.5	58.7	68.1	66.3	66.4	66.4	42.8	52.0	71.3	16.5	25.9	30.5	34.1	42.6	51.1
LaMAI+CoT	49.1	59.2	69.6	71.7	71.7	71.7	49.8	61.1	73.0	18.5	27.7	31.5	27.2	36.6	45.1
Oracle	59.4	86.9	72.4	78.2	78.2	78.1	60.5	72.0	83.5	21.0	33.5	36.0	29.4	62.5	42.6

Results



(a) GPT-4 Evaluation

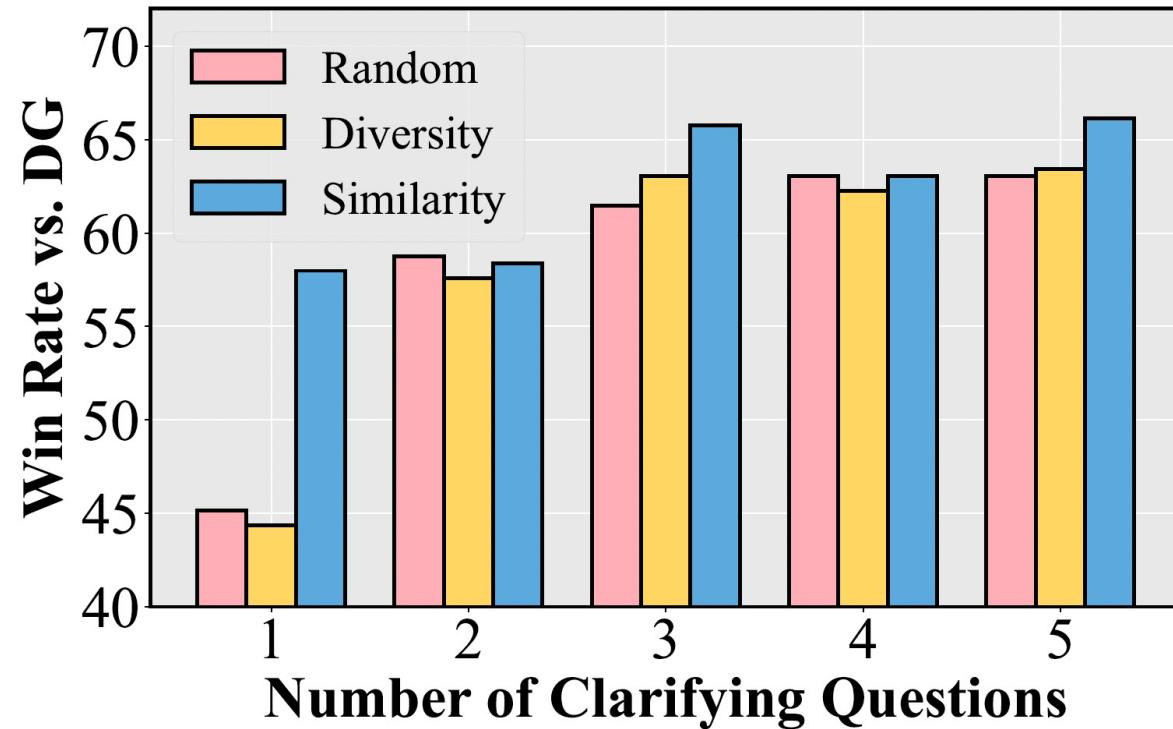


(b) Human-participated Experiment

Results on Vicuna-13B

	2WikiMultiHopQA			MuSiQue		
	EM	F1	Acc	EM	F1	Acc
DG	10.0	18.2	32.3	9.0	8.1	15.3
CoT	30.8	35.8	40.5	11.5	12.1	17.2
LaMAI	12.3	24.1	58.5	10.0	13.5	33.0
LaMAI+CoT	32.5	41.1	63.3	12.0	18.8	30.2

Ablation study



Thanks!

Q & A

Jing-Cheng Pang
2024.3.19