SuRe: Summarizing Retrievals using Answer Candidates for Open-domain QA of LLMs

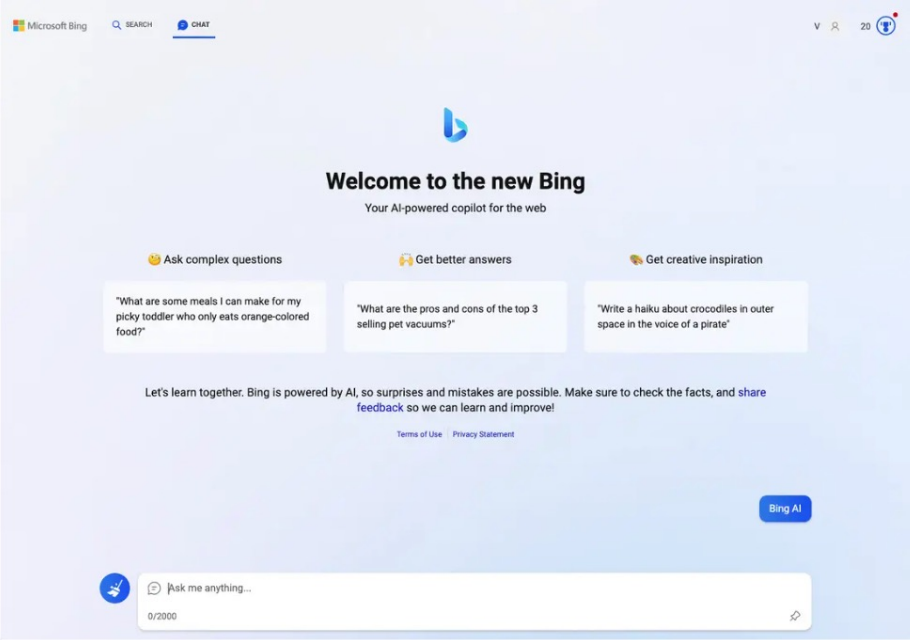
Jaehyung Kim[1], Jaehyun Nam[2], Sangwoo Mo[3], Jongjin Park[2],
Sang-Woo Lee[2,5], Minjoon Seo[2], Jung-Woo Ha[4,5], and Jinwoo Shin[2]

ICLR 2024

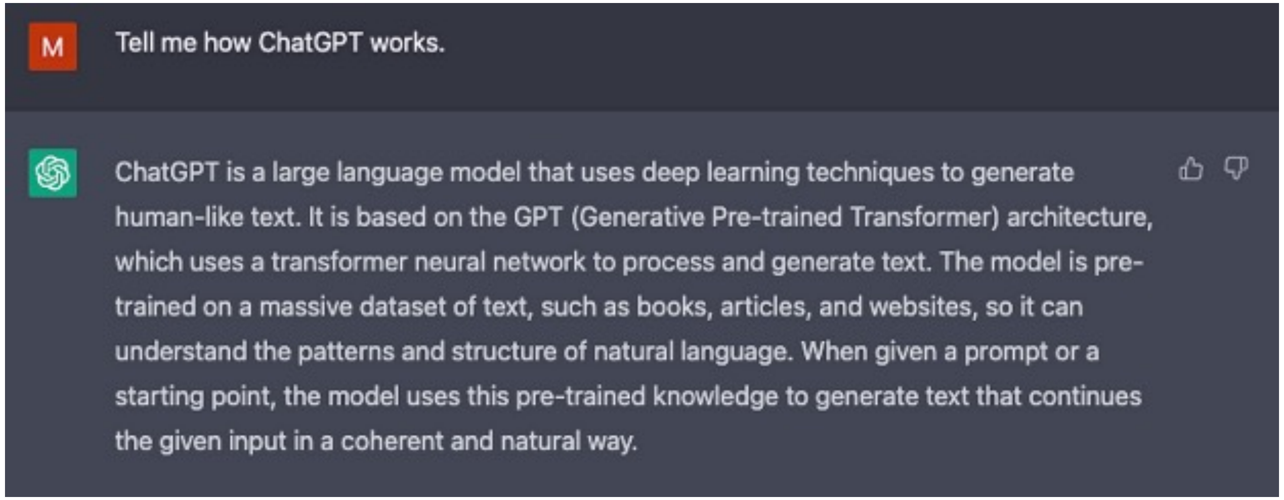
[1] CMU, [2] KAIST, [3] University of Michigan, [4] Naver AI Lab , [5] Naver Cloud

Success of Large Language Models (LLMs)

- LLMs become **core technology** in many real-world application



Search engine (Bing AI)[1]

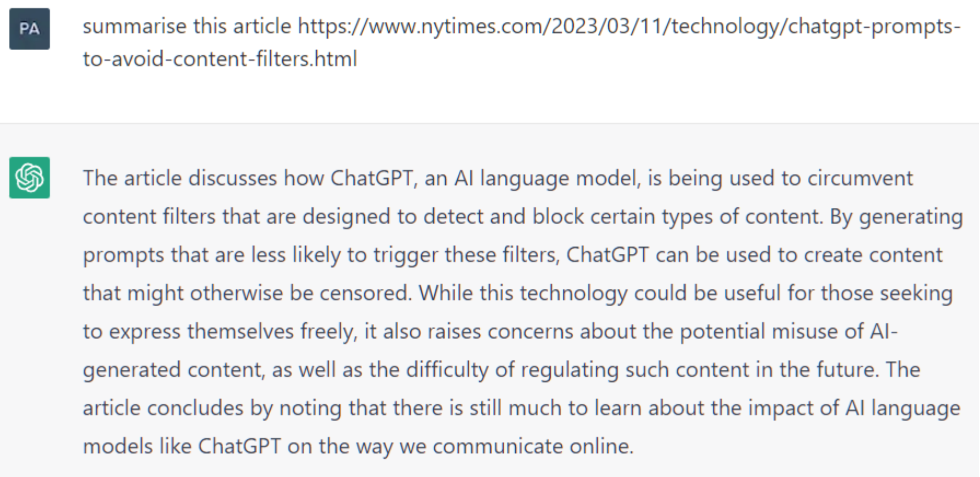


Chatbot (ChatGPT)[2]

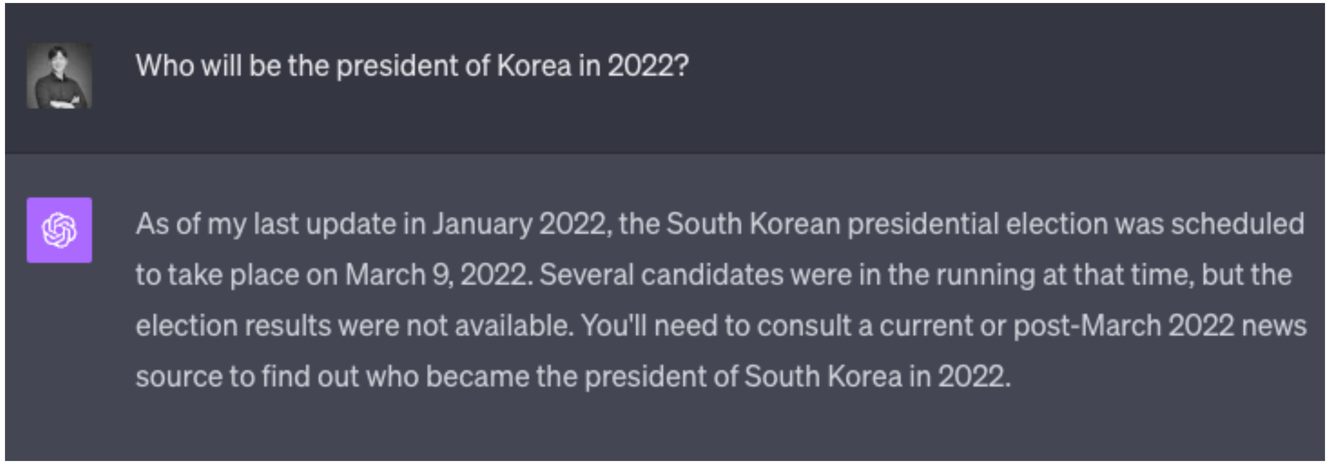
[1] Quy et al., Evaluation of Chatgpt and Microsoft Bing ai Chat Performances on Physics Exams of Vietnamese National High School Graduation Examination., arXiv:2306
[2] OpenAI., https://openai.com/blog/chatgpt., 2022

Challenges of Large Language Models

- Despite its great success, LLMs still suffer from **critical limitations**
  - Non-factual but seemingly plausible generation, *i.e.*, hallucinations[1]
  - Difficulty in integrating up-to-date knowledge



Hallucination of LLMs[2]



Struggle to incorporate up-to-date knowledge

[1] Welleck et al., Neural Text Generation with Unlikelihood Training., ICLR 2020
[2] https://en.wikipedia.org/wiki/Hallucination_%28artificial_intelligence%29

Retrieval Augmentation

- **Retrieving** and incorporating relevant knowledge → promising solution
  - *Retrieval* is effective way to improve QA system, *i.e.*, open-domain QA (ODQA)[1,2]

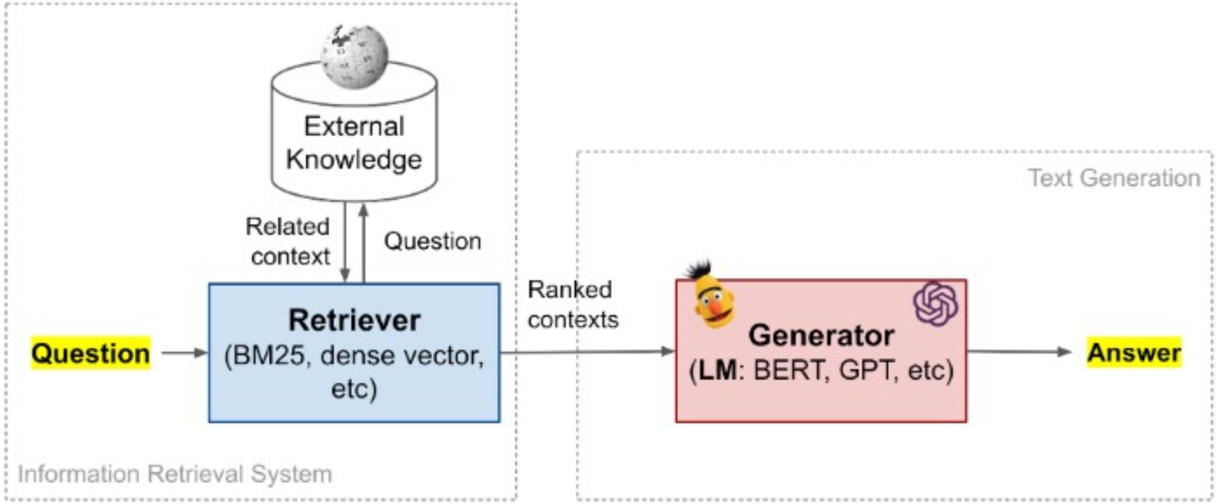
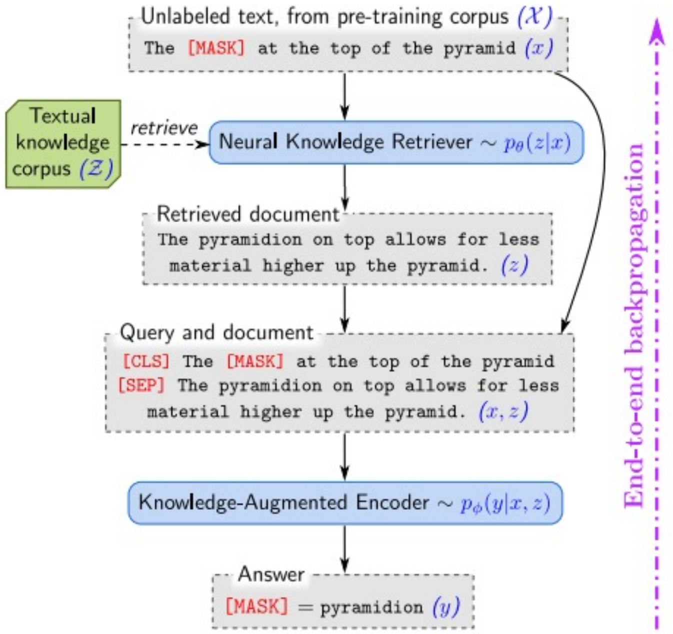


Illustration of retrieve-and-read system for ODQA[1]

[1] https://lilianweng.github.io/posts/2020-10-29-odqa/
[2] Karpukhin et al., Dense Passage Retrieval for Open-Domain Question Answering., EMNLP 2020

Retrieval Augmentation

- **Retrieving** and incorporating relevant knowledge → promising solution
  - *Retrieval* is effective way to improve QA system, *i.e.*, open-domain QA (ODQA)[1,2]
  - *Retrieval-augmented LLMs* outperforms ordinary LLMs[3,4]



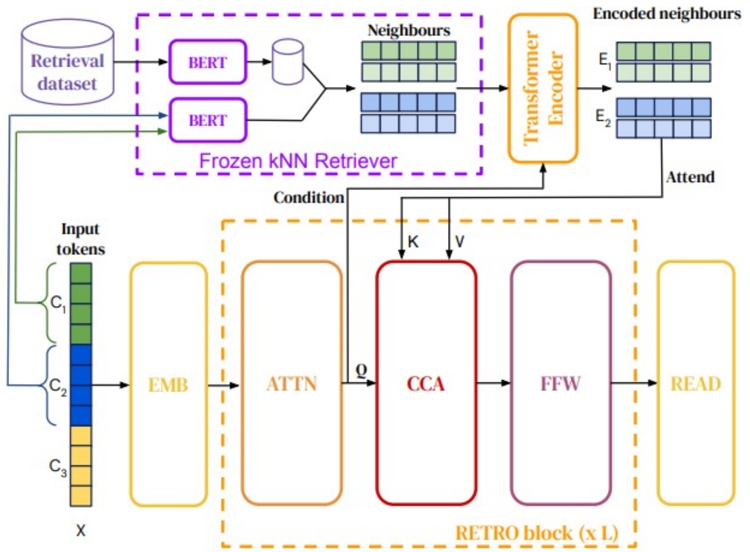
Name	NQ (79k/4k)	WQ (3k/2k)	CT (1k /1k)	# params
BERT-Baseline (Lee et al., 2019)	26.5	17.7	21.3	110m
T5 (base) (Roberts et al., 2020)	27.0	29.1	-	223m
T5 (large) (Roberts et al., 2020)	29.8	32.2	-	738m
T5 (11b) (Roberts et al., 2020)	34.5	37.4	-	11318m
DrQA (Chen et al., 2017)	-	20.7	25.7	34m
HardEM (Min et al., 2019a)	28.1	-	-	110m
GraphRetriever (Min et al., 2019b)	31.8	31.6	-	110m
PathRetriever (Asai et al., 2019)	32.6	-	-	110m
ORQA (Lee et al., 2019)	33.3	36.4	30.1	330m
Ours ($\mathcal{X}$ = Wikipedia, $\mathcal{Z}$ = Wikipedia)	39.2	40.2	**46.8**	330m
Ours ($\mathcal{X}$ = CC-News, $\mathcal{Z}$ = Wikipedia)	**40.4**	**40.7**	42.9	330m

Illustration of REtrieval-Augmented Language Model (REALM)[3]

[1] https://lilianweng.github.io/posts/2020-10-29-odqa/
[2] Karpukhin et al., Dense Passage Retrieval for Open-Domain Question Answering., EMNLP 2020
[3] Guu et al., REALM: Retrieval-Augmented Language Model Pre-Training., ICML 2020
[4] Borgeaud et a., Improving Language Models by Retrieving from Trillions of Tokens., ICML 2022

Challenges with Retrieval Augmented LLMs

- Retrieval-augmented LLMs are typically constructed with **finetuning**[1,2]
  - But, recent large scale & black-box nature make it less attractive & infeasible

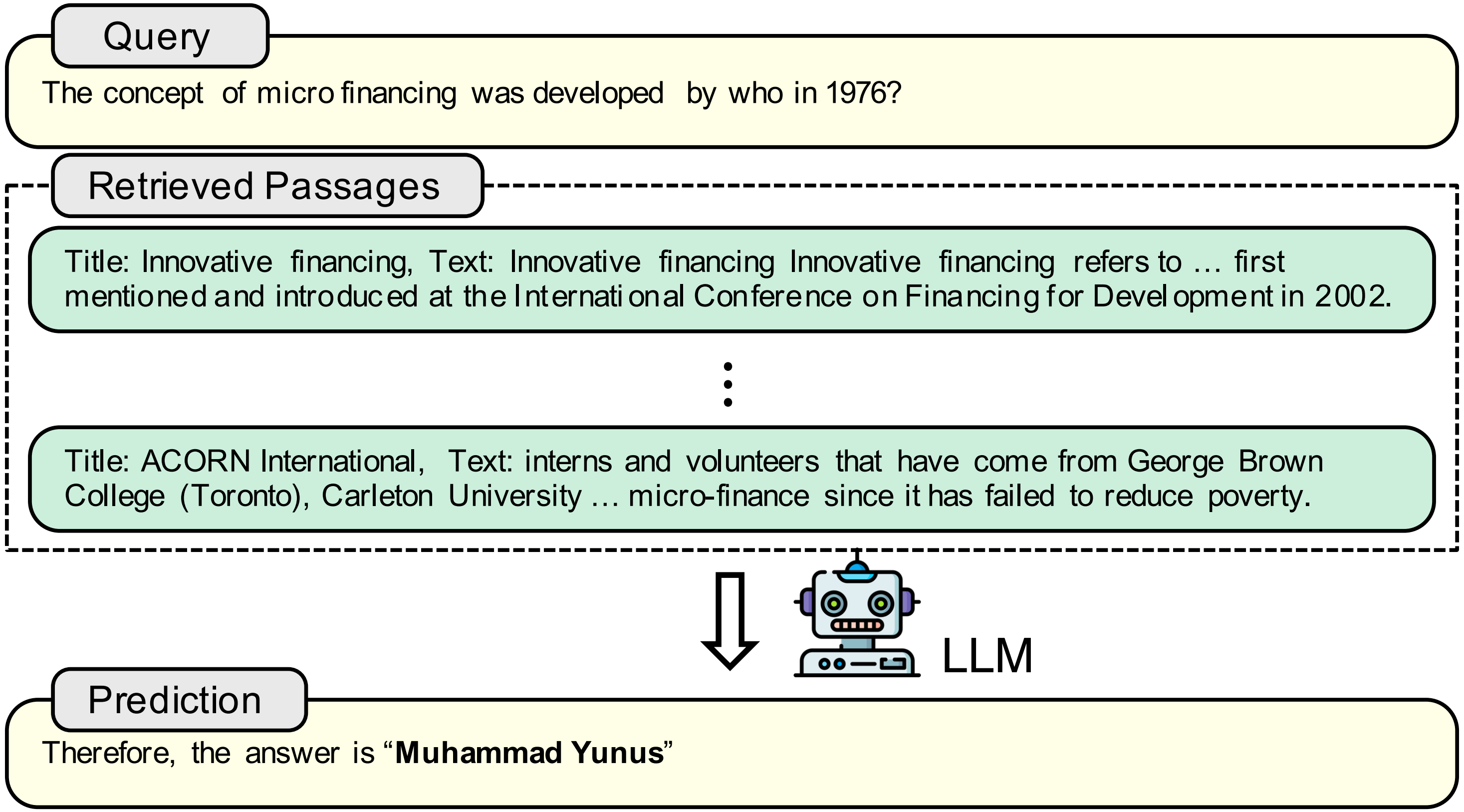


Modification of architecture and training to learn retrieval augmentation[2]

[1] Guu et al., REALM: Retrieval-Augmented Language Model Pre-Training., ICML 2020
[2] Borgeaud et a., Improving Language Models by Retrieving from Trillions of Tokens., ICML 2022

Challenges with Retrieval Augmented LLMs

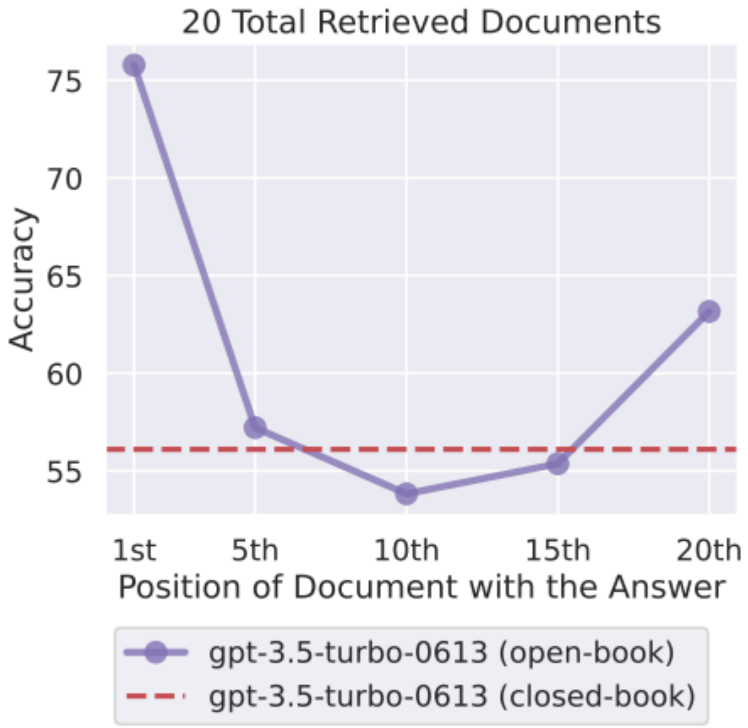
- **Prompting** is simple and efficient way to augment retrieved passages[1]



[1] Shi et al., REPLUG: Retrieval-Augmented Black-Box Language Models., arXiv:2301
[2] Liu et al., Lost in the Middle: How Language Models Use Long Contexts., arXiv:2307

Challenges with Retrieval Augmented LLMs

- **Prompting** is simple and efficient way to augment retrieved passages[1]
  - Naïve approach (*e.g.*, appending) could be limited in fully exploiting the retrieval[2]



LLMs struggle to handle long retrieved passages[2]

[1] Shi et al., REPLUG: Retrieval-Augmented Black-Box Language Models., arXiv:2301
[2] Liu et al., Lost in the Middle: How Language Models Use Long Contexts., arXiv:2307

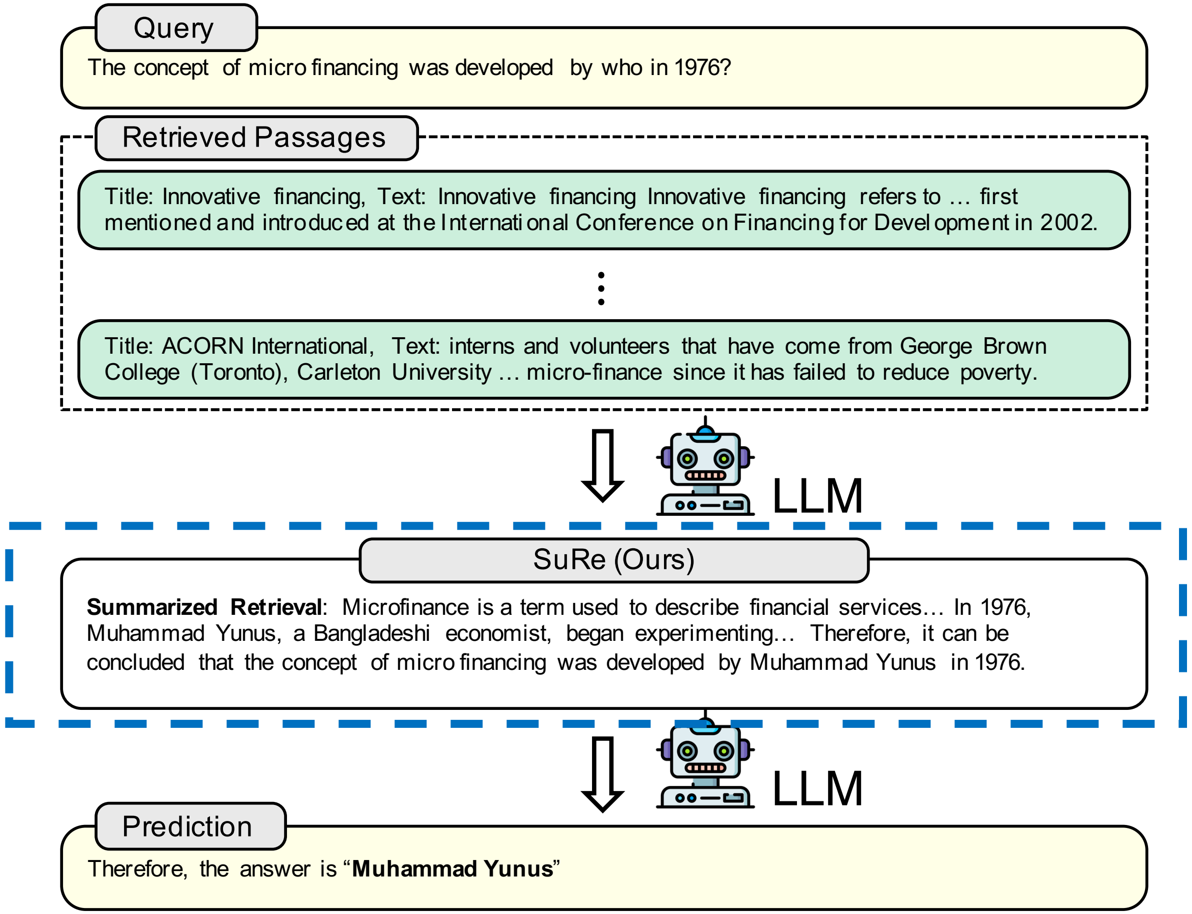
Research Goal

- **Prompting** is simple and efficient way to augment retrieved passages[1]
  - Naïve approach (*e.g.*, appending) could be limited in fully exploiting the retrieval[2]

Research Goal: a simple yet effective framework based on prompting to improve ODQA with LLMs

Proposed Method: SuRe

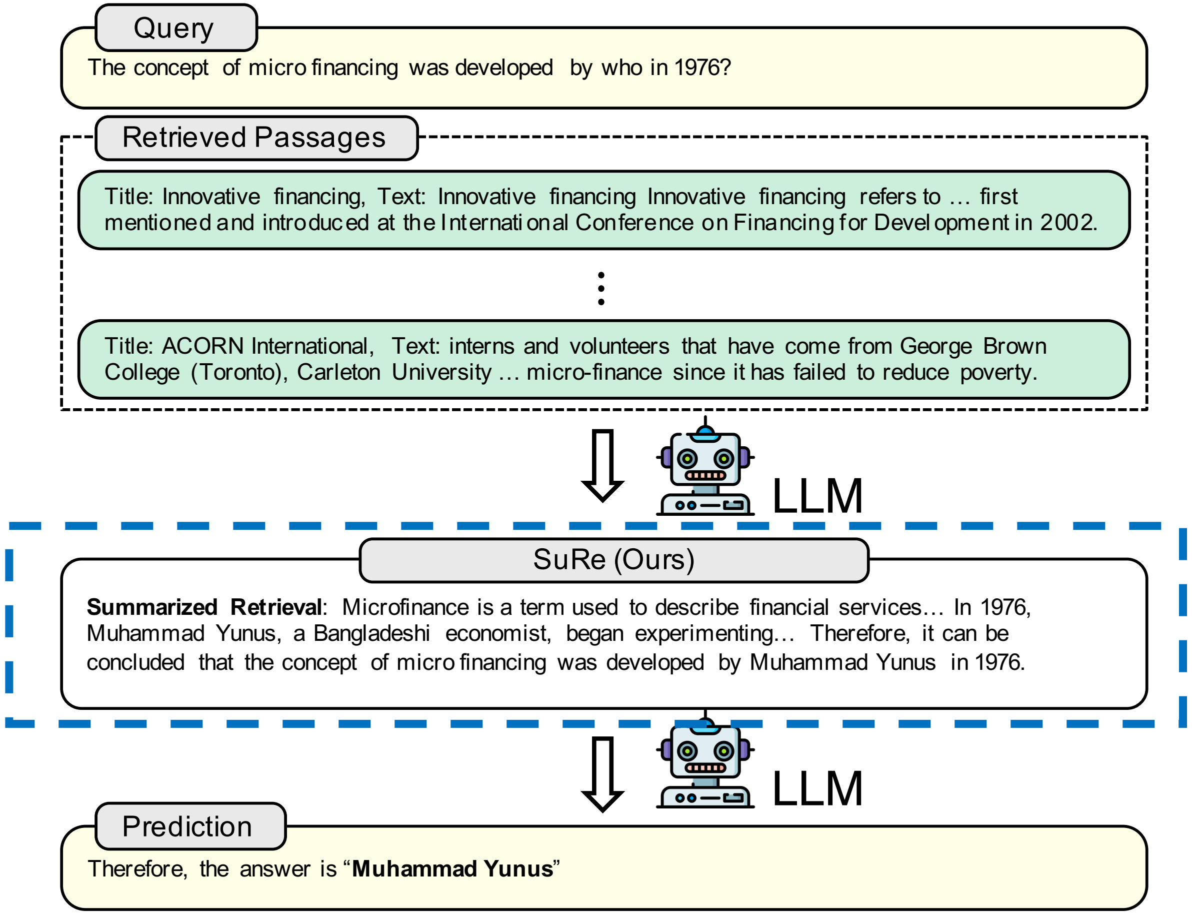
- **SuRe**: Summarizing Retrievals using Answer Candidates



Example of Retrieval Augmentation of LLMs via Proposed SuRe Framework

Proposed Method: SuRe

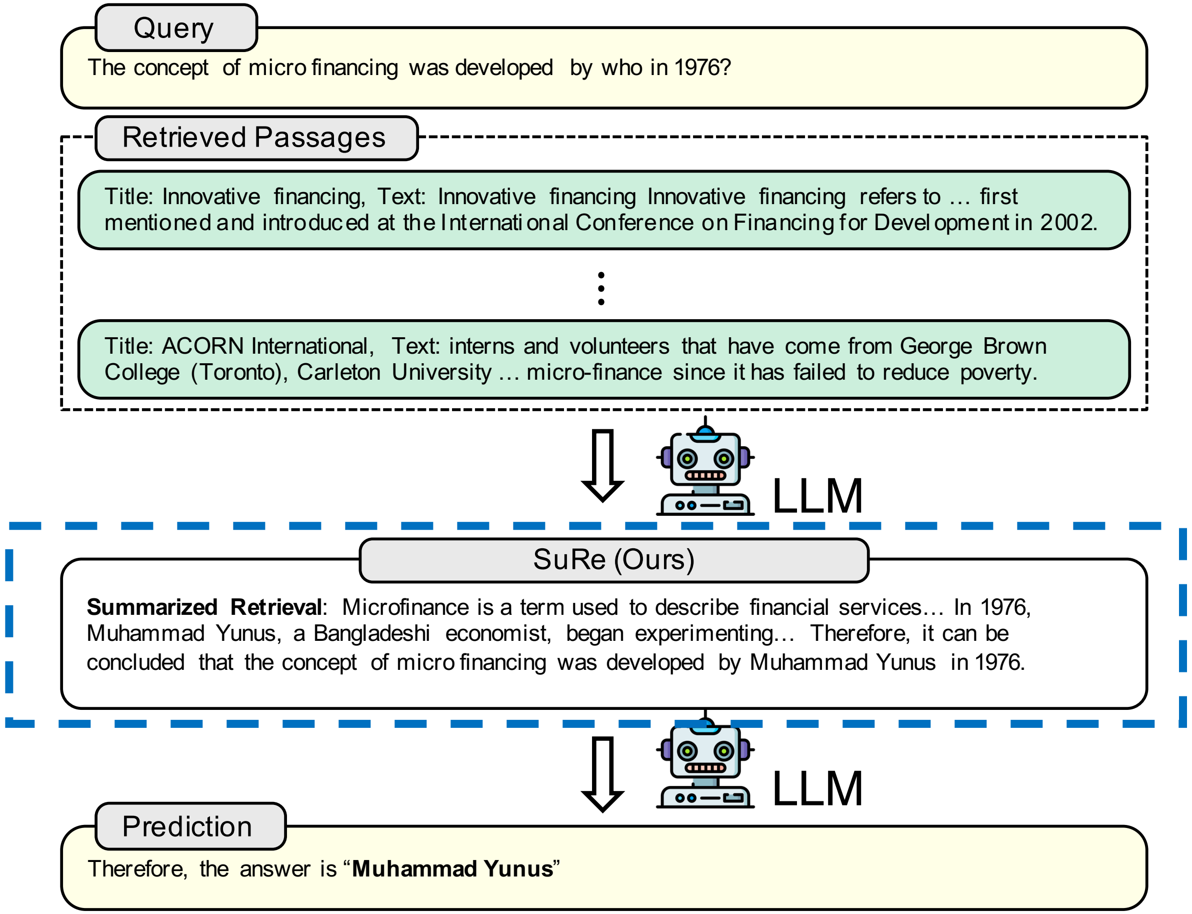
- **SuRe**: Summarizing Retrievals using Answer Candidates
  - Key idea. Select answer that is well-supported by summarization of retrievals (explicit rationale)



Example of Retrieval Augmentation of LLMs via Proposed SuRe Framework

Proposed Method: SuRe

- **SuRe**: Summarizing Retrievals using Answer Candidates
  - Key idea. Select answer that is well-supported by summarization of retrievals



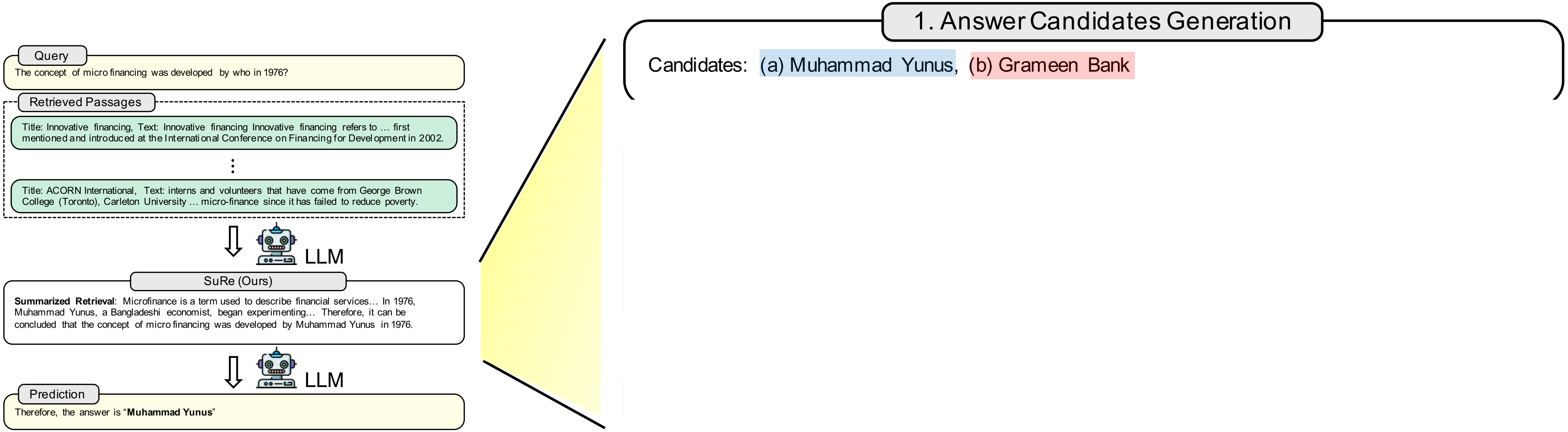
Example of Retrieval Augmentation of LLMs via Proposed SuRe Framework

Proposed Method: SuRe

- SuRe conducts **three steps** sequentially (via zero-shot prompting)

Proposed Method: SuRe

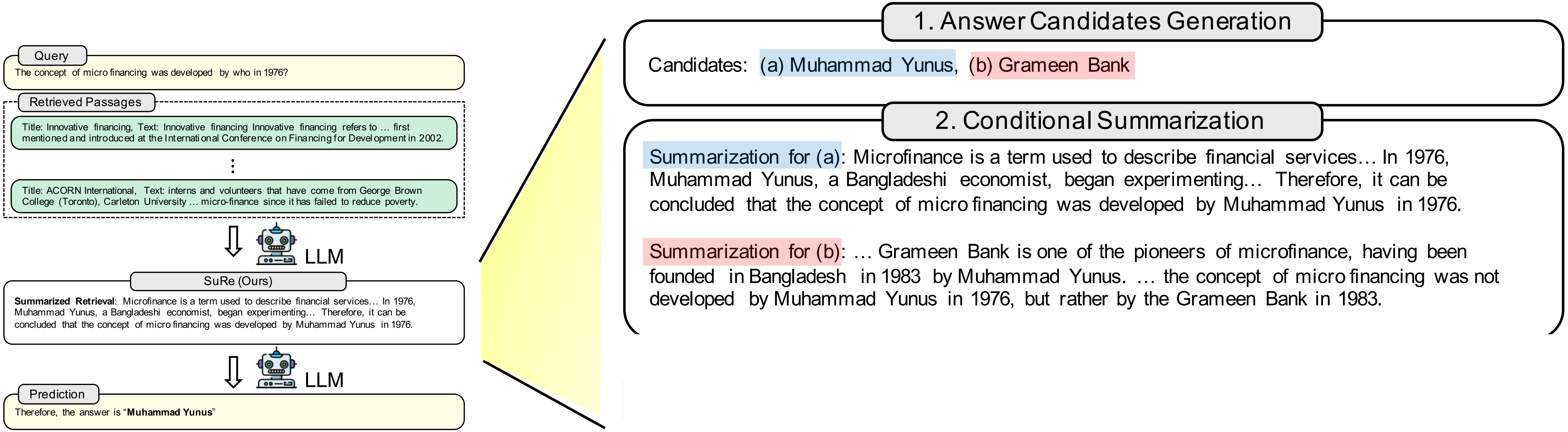
- SuRe conducts **three steps** sequentially (via zero-shot prompting)
  1. Answer candidates generation



Overall illustration of SuRe Framework

Proposed Method: SuRe

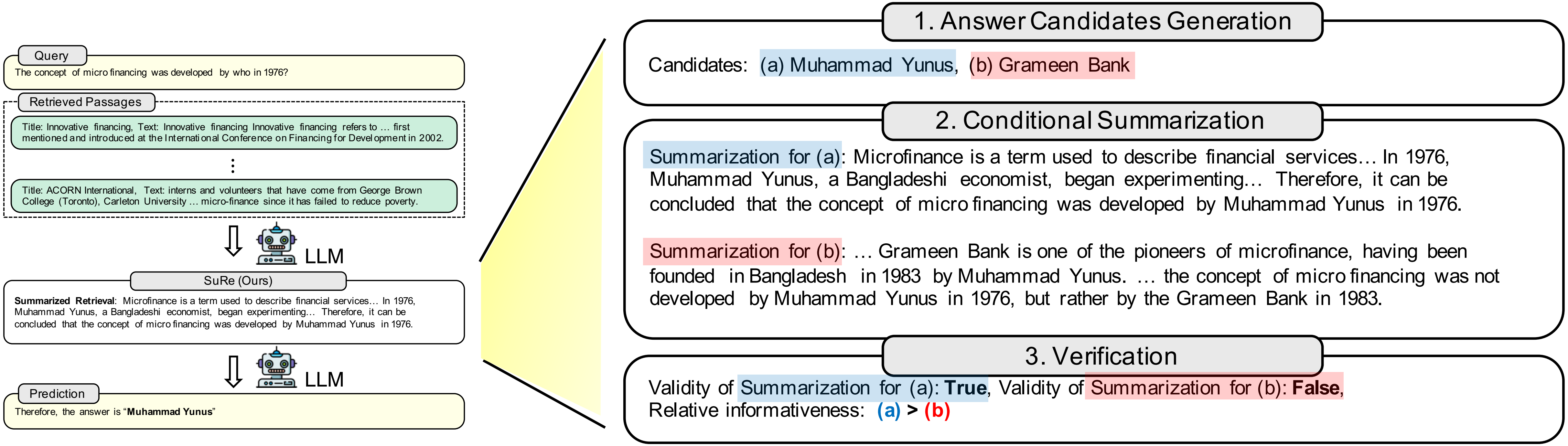
- SuRe conducts **three steps** sequentially (via zero-shot prompting)
  1. Answer candidates generation
  2. Conditional summarization



Overall illustration of SuRe Framework

Proposed Method: SuRe

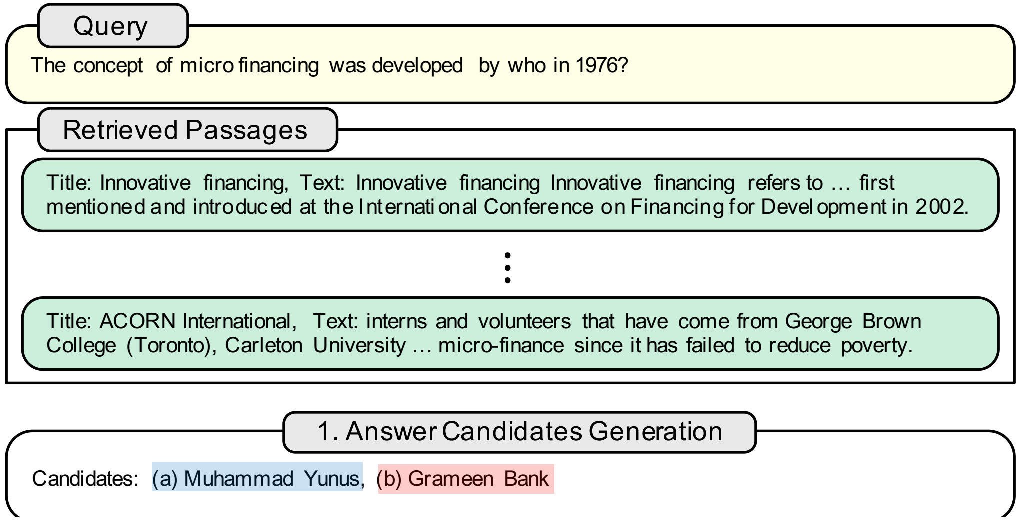
- SuRe conducts **three steps** sequentially (via zero-shot prompting)
  1. Answer candidates generation
  2. Conditional summarization
  3. Selection via verification (validity & ranking)



Overall illustration of SuRe Framework

SuRe: Answer Candidates Generation

- SuRe first generate K answer candidates from question and retrieval
  - Compare to random sampling, we observe *direct prompting is more effective*
  - We use fixed K=2 during experiments, as K > 2 is quite inefficient

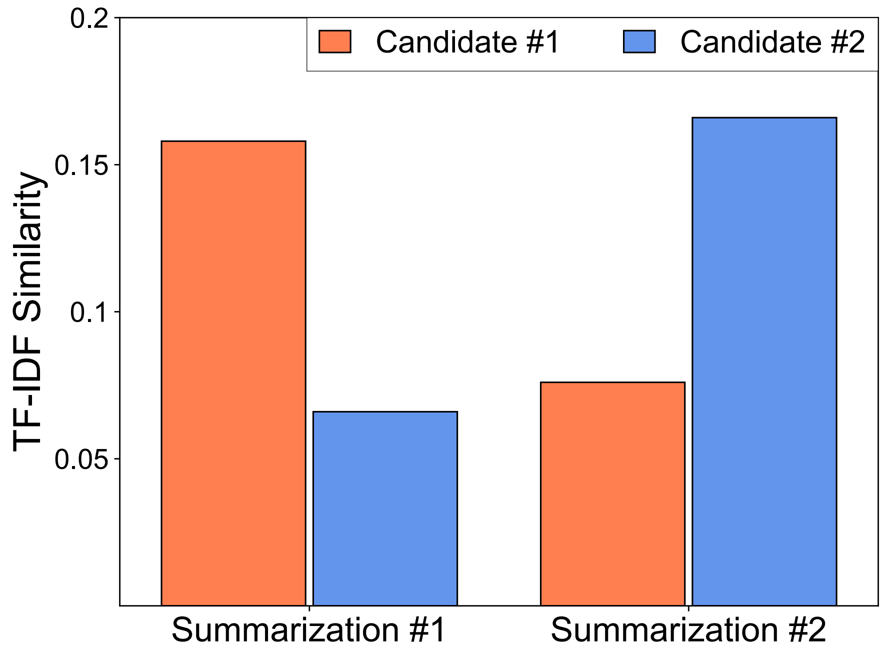


Listing 1 Prompt for answer candidates generation.

```
f'''
Below are N passages related to the question at the end. After reading
↪ the passages, provide two correct candidates for the answer to the
↪ question at the end. Each answer should be in the form: (a) xx, (b)
↪ yy, and should not exceed 3 words for each candidate.

Passage #1 Title: {Passage #1 Title}
Passage #1 Text: {Passage #1 Text}

...

Passage #N Title: {Passage #N Title}
Passage #N Text: {Passage #N Text}

Question: {Question}

Answer:
'''
```

SuRe: Conditional Summarization

- SuRe then summarize retrieval conditioned on each candidate
  - *i.e.*, question, retrieved passages, and answer candidate → summarization
  - Conditional summarization includes *specific contexts* supporting given candidate



TF-IDF similarity between candidates and conditional summarizations

SuRe: Selection via Verification

- Lastly, SuRe verify summarizations and select most plausible candidate

1. Answer Candidates Generation

Candidates: (a) Muhammad Yunus, (b) Grameen Bank

2. Conditional Summarization

Summarization for (a): Microfinance is a term used to describe financial services… In 1976, Muhammad Yunus, a Bangladeshi economist, began experimenting… Therefore, it can be concluded that the concept of micro financing was developed by Muhammad Yunus in 1976.

Summarization for (b): … Grameen Bank is one of the pioneers of microfinance, having been founded in Bangladesh in 1983 by Muhammad Yunus. … the concept of micro financing was not developed by Muhammad Yunus in 1976, but rather by the Grameen Bank in 1983.

3. Verification

Validity of Summarization for (a): **True**, Validity of Summarization for (b): **False**,
Relative informativeness: **(a) > (b)**

SuRe: Selection via Verification

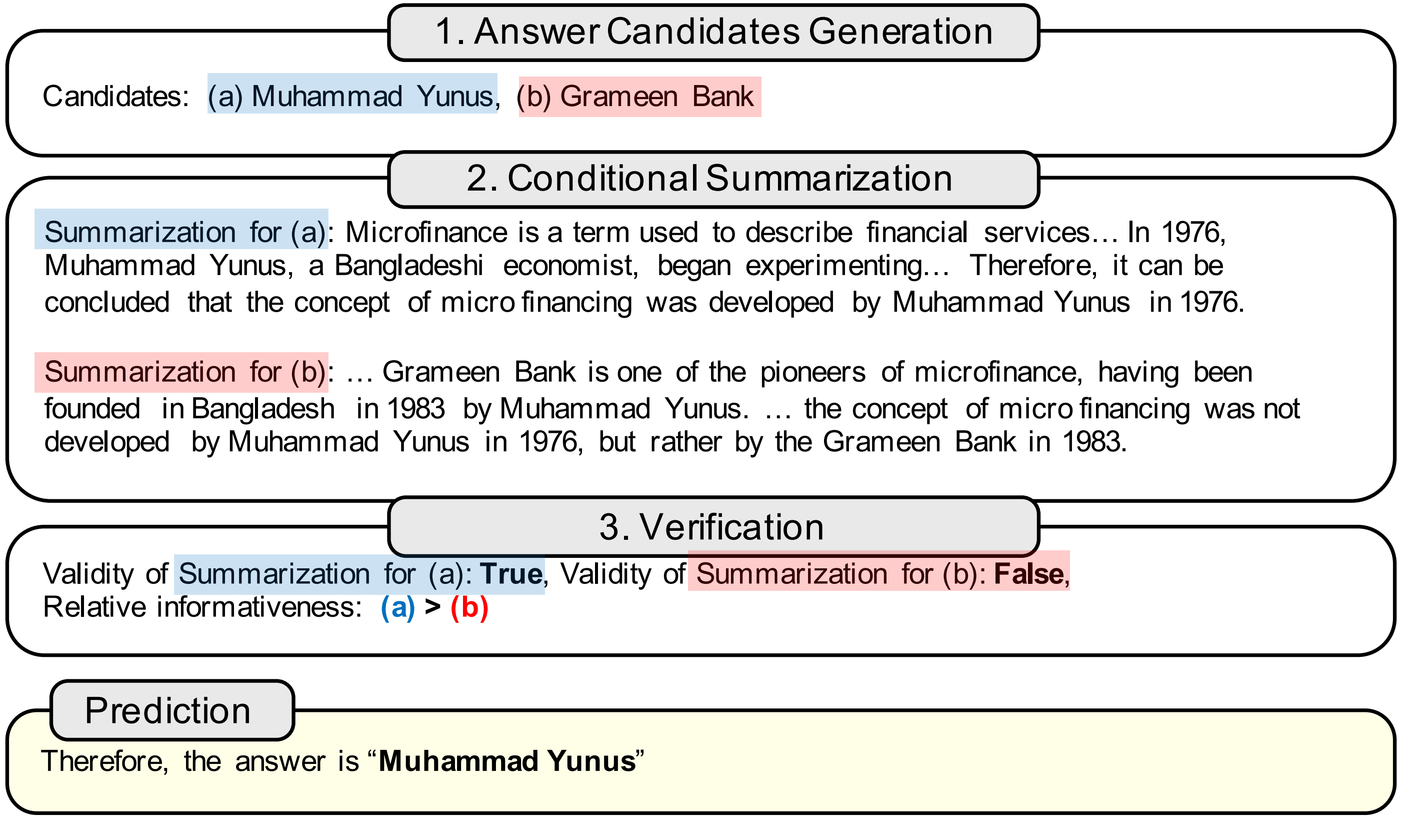
- Lastly, SuRe verify summarizations and select most plausible candidate
  - Two scores: (1) Validity of summarization

Listing 3 Prompt for instance-wise validation.

```
f'''
Question: {Question}

Prediction: {Prediction}

Passage: {Passage}

Does the passage correctly support the prediction? Choices: [True,
↪  False]. Answer:
'''
```

SuRe: Selection via Verification

- Lastly, SuRe verify summarizations and select most plausible candidate
  - Two scores: (1) Validity of summarization & (2) Ranking between summarizations
  - Candidate with highest scored summarization → **Answer**



Experiments: Comparison with Baselines

- **ChatGPT** with 10 retrieved passages from **BM25**
  - Augmenting retrieved passages with prompting (*Base*) is effective
  - Other baselines are ineffective due to challenging setup (zero-shot, black-box API)
  - SuRe significantly improve *Base* with a large margin (4.6% in EM & 4.0% in F1)

Methods / Datasets	NQ	WebQ	2Wiki	HotpotQA	Average
No retrieval	27.6 / <u>39.0</u>	<u>25.0</u> / **38.8**	21.4 / 24.8	22.2 / 31.9	24.1 / 33.6
Base	<u>28.4</u> / 38.8	19.6 / 32.5	<u>27.4</u> / <u>32.8</u>	30.8 / 40.3	<u>26.6</u> / <u>36.1</u>
Rerank	24.8 / 33.9	18.8 / 30.6	23.0 / 28.4	27.8 / 37.4	23.6 / 32.6
RePlug	26.0 / 35.3	18.8 / 31.5	23.6 / 28.5	28.0 / 37.9	24.1 / 33.3
Selection-inference	24.3 / 32.8	17.3 / 28.6	22.6 / 29.5	30.8 / 39.6	23.8 / 32.6
Chain-of-thoughts	22.3 / 31.4	15.2 / 27.8	19.6 / 22.5	25.6 / 31.8	20.7 / 28.4
Self-verification	25.2 / 35.4	16.1 / 28.5	23.2 / 30.5	<u>31.6</u> / <u>41.8</u>	24.0 / 34.1
SuRe (Ours)	**33.5 / 42.3**	**25.1** / <u>36.6</u>	**32.8 / 38.1**	**33.2 / 43.4**	**31.2 / 40.1**

Experiments: Compatibility of SuRe

- Evaluation on **different LLMs and retrieval methods**
  - Naïve prompting (Base) is compared as baseline
  - SuRe is consistently effective for all setups with 4.6% EM improvement in average
  - SuRe successfully improves LLaMA2-chat → Usefulness for open sourced LLMs!

	ChatGPT						GPT-4		LLaMA2-chat	
Datasets	BM25	+ SuRe	DPR	+ SuRe	Contriever	+ SuRe	BM25	+ SuRe	BM25	+ SuRe
NQ	28.4	**33.5**	36.1	**41.0**	35.8	**40.4**	30.2	**32.4**	18.6*	**30.4***
WebQ	19.6	**25.1**	23.2	**27.3**	22.5	**28.7**	21.5	**21.7**	16.0*	**24.0***
2Wiki	27.4	**32.8**	19.2	**21.4**	27.2	**32.6**	34.8	**38.2**	20.2	**27.8**
HotpotQA	30.8	**33.2**	25.6	**27.4**	32.2	**33.6**	34.8	**40.6**	24.0	**28.0**
Average	26.6	**31.2**	26.0	**29.3**	29.4	**33.8**	30.3	**33.2**	19.7	**27.6**

Experiments: Additional Advantages of SuRe

- Not only answer, but SuRe also generates **conditional summarization**
  - Then, is this conditional summarization really a *good rationale for answer*?

Question

What's the movie called with Justin Timberlake and time?

Answer

In Time

Generic Summarization

Justin Timberlake has had a successful career not only in music but also in film. He has appeared in several movies, including "The Social Network", "Friends with Benefits", and "Trolls". However, there is one movie that stands out as a collaboration between Timberlake and the concept of time: "In Time". Released in 2011, the movie is set in a future where people stop aging at 25 and must work to buy themselves more time. Timberlake plays the lead role of Will Salas, a man who is falsely accused of murder and must go on the run with a hostage, played by Amanda Seyfried. The movie received mixed reviews but was praised for its unique concept and Timberlake's performance. Therefore, "In Time" is the movie that features Justin Timberlake and the concept of time.

Candidate-conditioned Summarization (Ours)

Justin Timberlake starred in the 2011 science fiction action film "In Time" alongside Amanda Seyfried. The film is set in a future where people stop aging at 25 and must work to buy themselves more time, with time becoming the new currency. Timberlake plays Will Sales, a man falsely accused of murder who goes on the run with a wealthy hostage, played by Seyfried. The two must navigate a dangerous world where time is everything and corruption is rampant. The film received mixed reviews but was praised for its unique concept and stylish visuals.

Qualitative comparison between SuRe's summarization and generic (i.e., unconditional) summarization

Experiments: GPT-4 & Human Evaluations

- We ask GPT-4 to select more informative & plausible sum. as rationale[1]

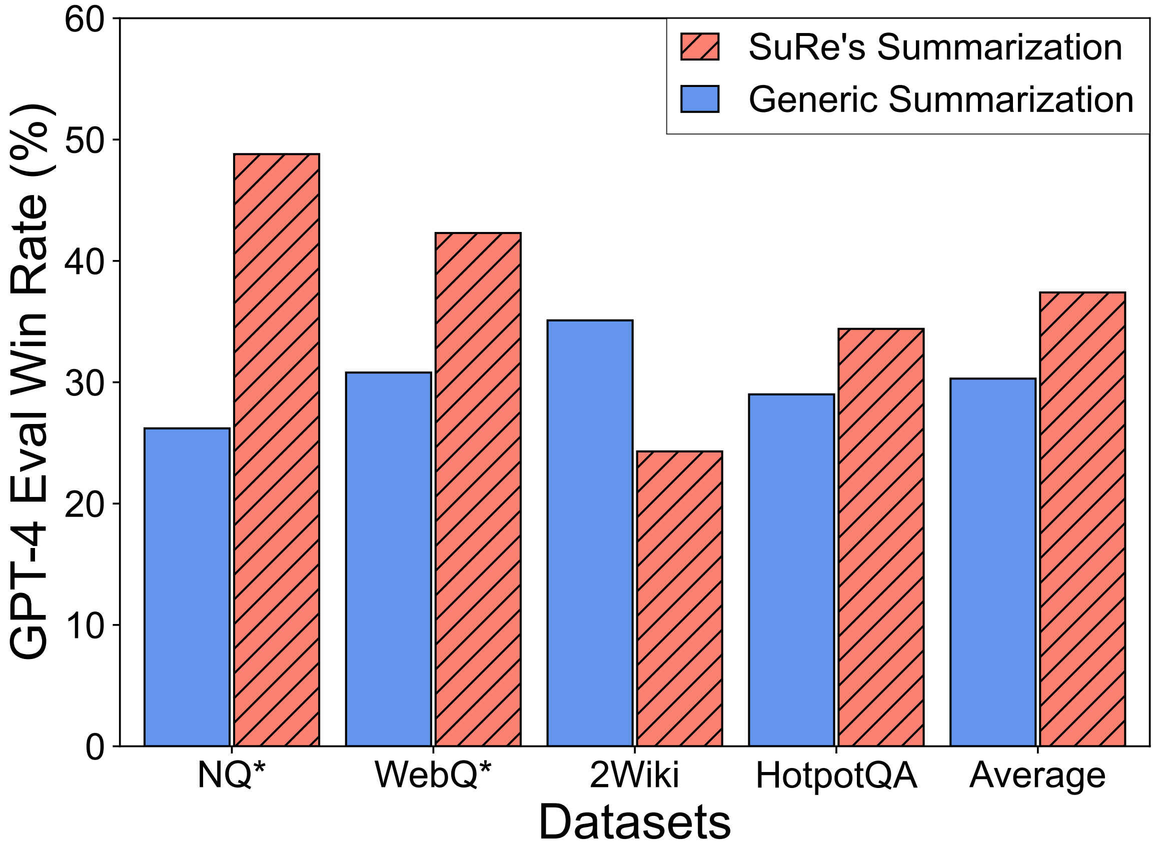
```
f'''
Question: Given the following summaries for the target question,
↪   determine which one is more informative and plausible as rationale
↪   to support a given target question-answer pair.

Summary 1: {Summary 1}

Summary 2: {Summary 2}

Target Question: {Question}

Target Answer: {Answer}

Your Task:
Identify which summary (Summary 1 or Summary 2) is more informative and
↪   plausible as rationale to support a given answer at hand. Choices:
↪   [Summary 1, Summary 2].

Answer:
'''
```

Designed Prompt for GPT-4 Evaluation between Summarizations

[1] Liu et al., G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment., arXiv:2303

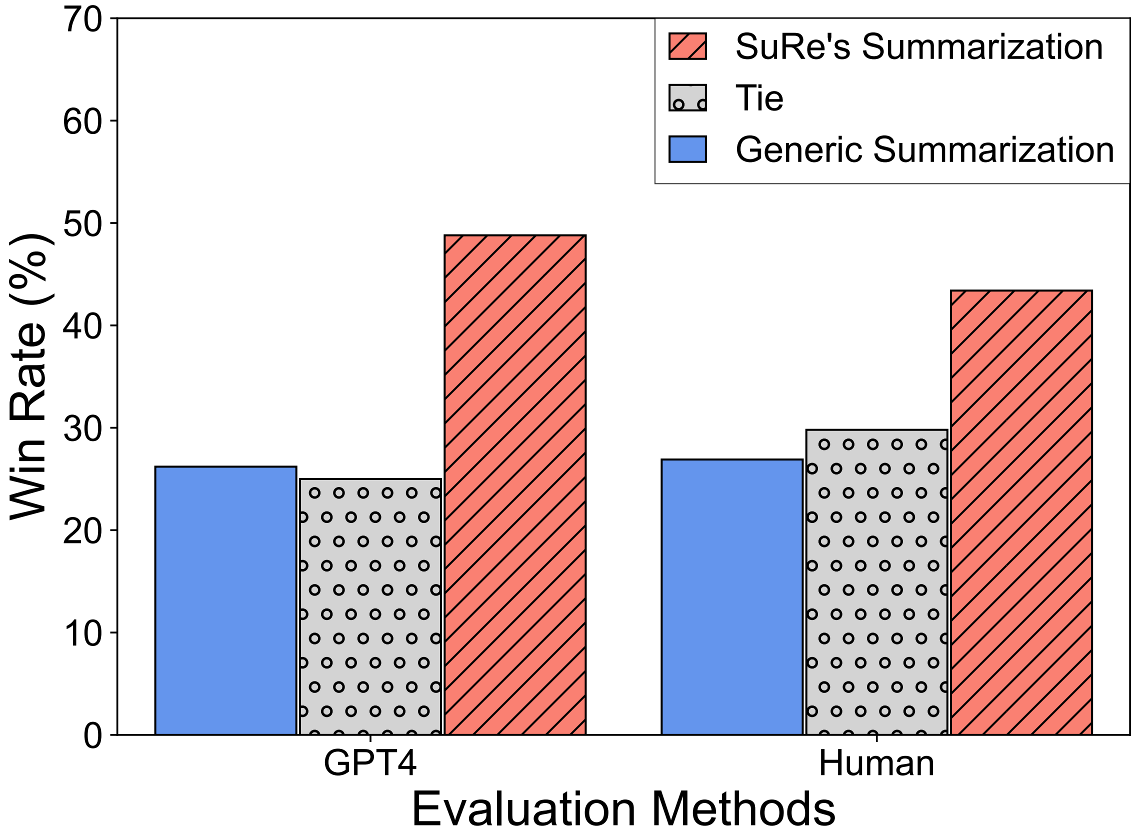
Experiments: GPT-4 Evaluation

- **GPT-4 prefers SuRe's** summarization than generic summarization
  - Setup. We only use questions that both summarizations correctly predict
  - Result. Average: **37.4%** (SuRe) vs 30.3 % (Generic)



Experiments: Human Evaluation

- **Human also prefers SuRe's** summarization than generic one
  - Setup. 84 NQ questions that both summarizations correctly predict (7 experts)
  - Result. **43.4%** (SuRe) vs 26.9 % (Generic)



Thank you for attention 🙏