中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

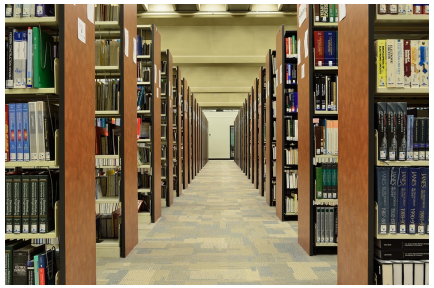
Smarter Retrieval for Smarter Generation

--When and How to Retrieve for Retrieval-Augmented Generation

Keping Bi

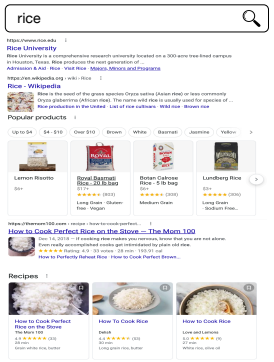
Institute of Computing Technology
Chinese Academy of Sciences
2025/11/14

The Evolution of Information Access



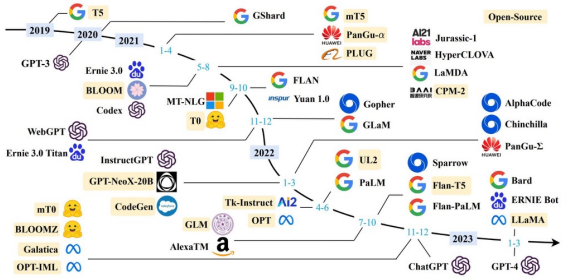
Traditional Library Era

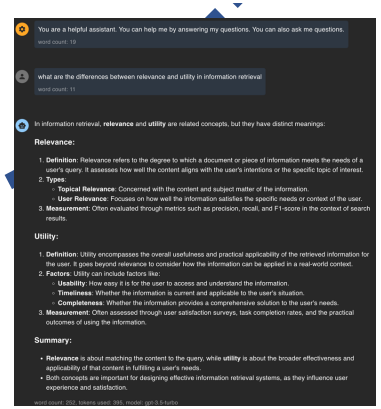
1960s



Web Search Era

1990s



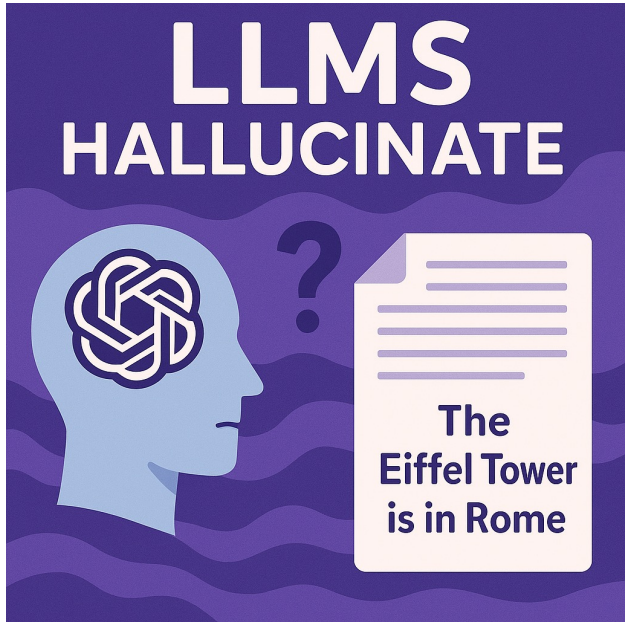


Generative AI Era

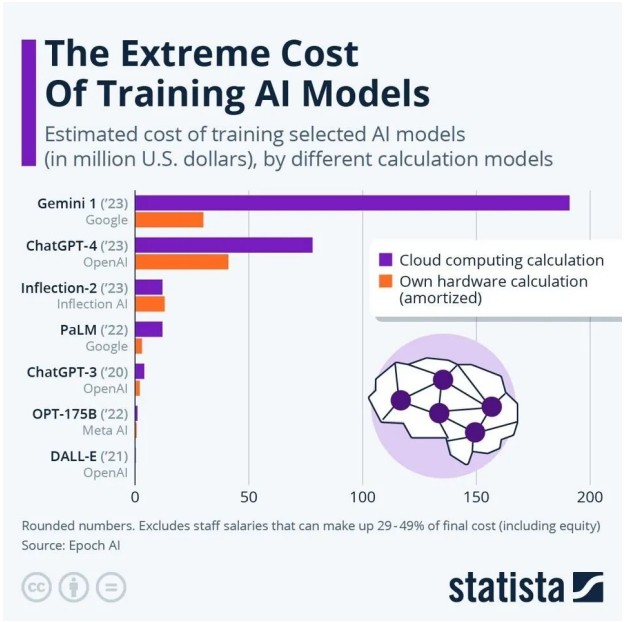
2020s

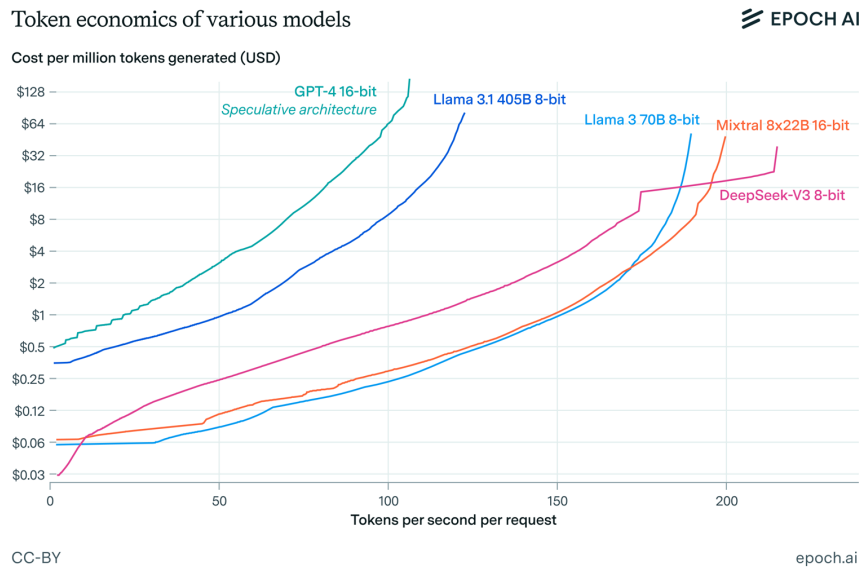
Limitations of LLMs

Hallucination & Outdated information

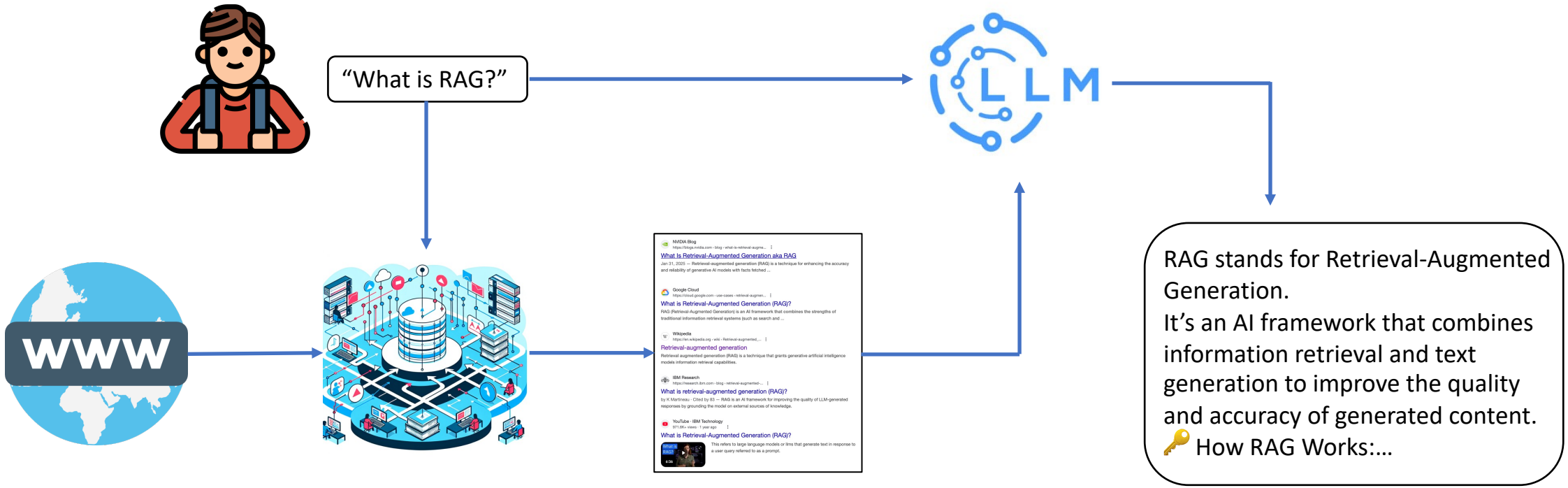


High Training and Inference Cost





Retrieval-Augmented Generation (RAG)



- Provide factual evidence
- Continuously update the index

LLM-Generated Responses vs. RAG

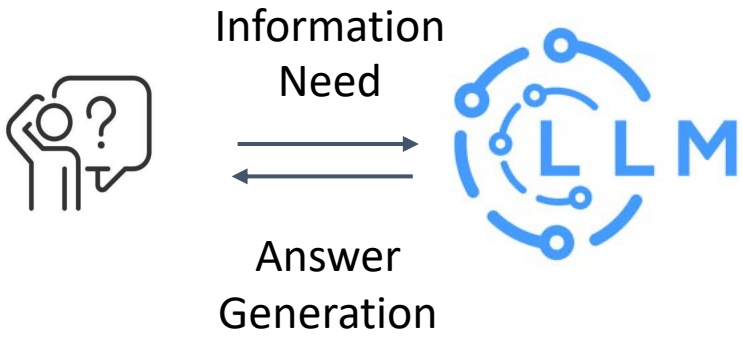
📕 Closed-Book

- Answers questions only using the model's internal knowledge
- Information sources are sealed and unknown
- Smaller inference cost

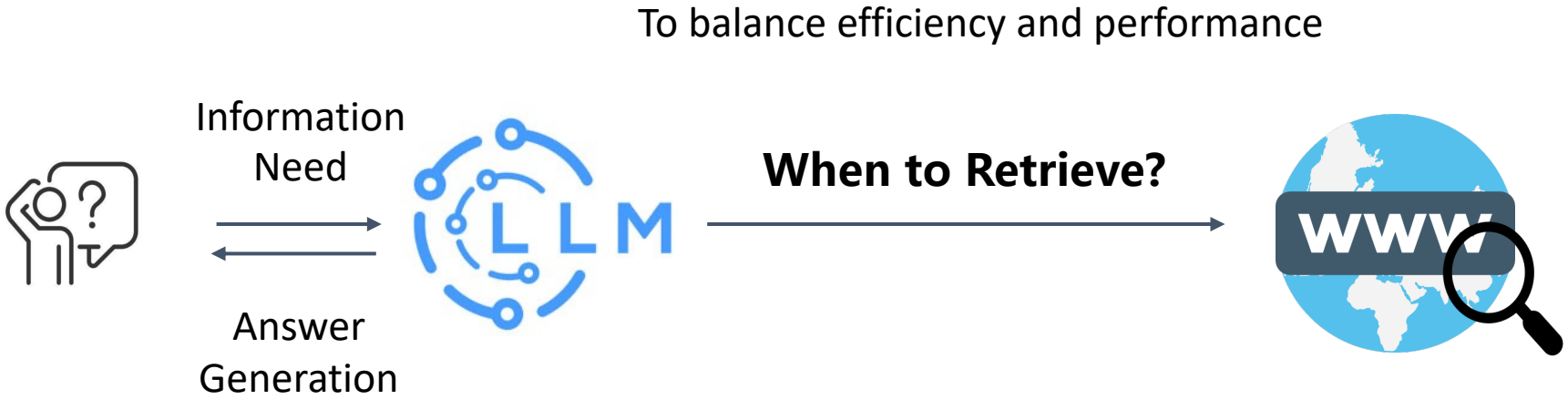
📖 Open-Book

- Uses external knowledge (retrieval results) to assist LLM generation
- Information sources are **open** and **explicit**
- **Larger inference cost**: retrieval + longer input sequences
- Performance **relies on retrieval quality**

Smarter Retrieval for Smarter Generation

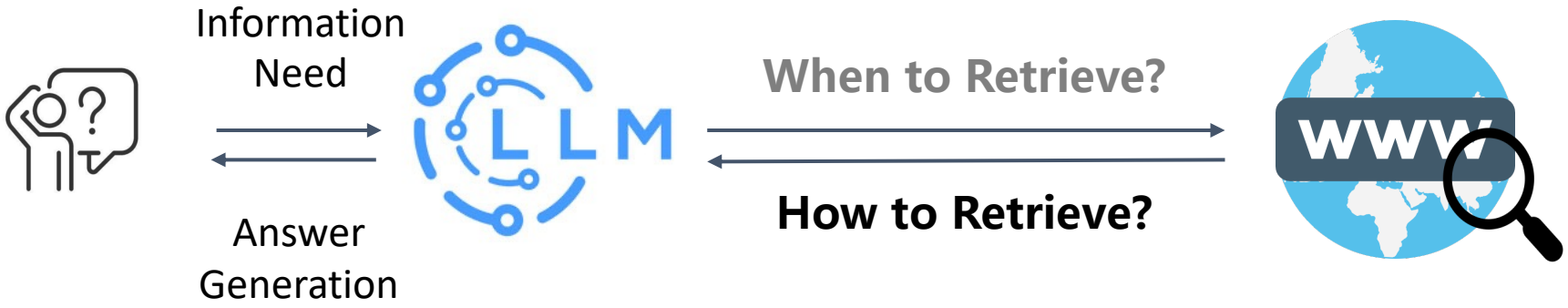


Smarter Retrieval for Smarter Generation

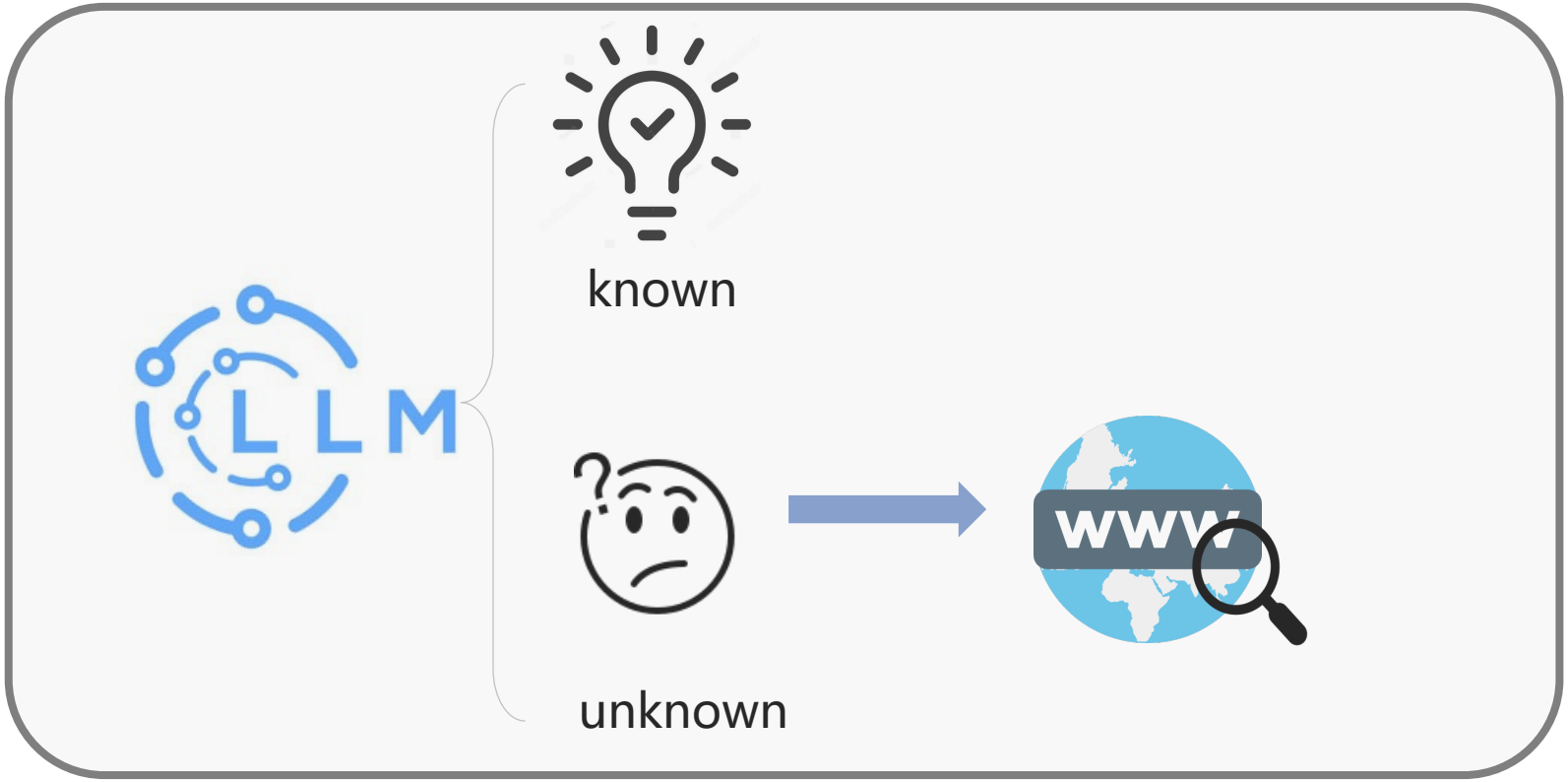


Smarter Retrieval for Smarter Generation

To balance efficiency and performance



To provide results beneficial for answer generation



When to Retrieve for RAG?

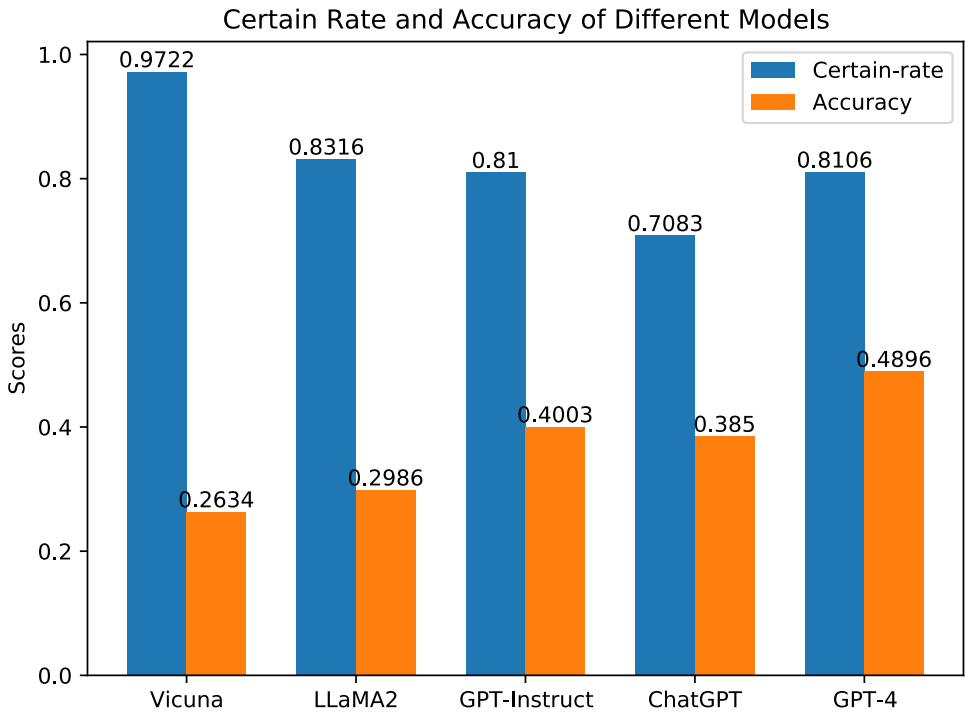
LLM Knowledge Boundary Perception

Goals

- Effectively determine whether LLMs know/unknow the knowledge
  - ➢ Invoke retrieval augmentation **only when the model does not know**, reducing unnecessary RAG overhead
  - ➢ Judge the reliability of model outputs

Key Challenges

- LLMs tend to be **overconfident**
- Signals indicating knowledge boundaries are **unclear**



Confidence Estimation

Confidence Estimation

Token-Prob-Based

Response: Albert Einstein received the 1921 Nobel Prize in Physics for ... So the answer is Albert Einstein.

Confidence = Avg Token Prob

Continuous; inaccurate

Verbalized

Input: Answer the question and **give your confidence** on your answer.
Response: ... The answer is Albert Einstein and **my confidence is 100%.**

Discrete; natural; flexible

Self-Consistency-Based

... Albert Einstein

...Wilhelm Röntgen

⋮

... Albert Einstein

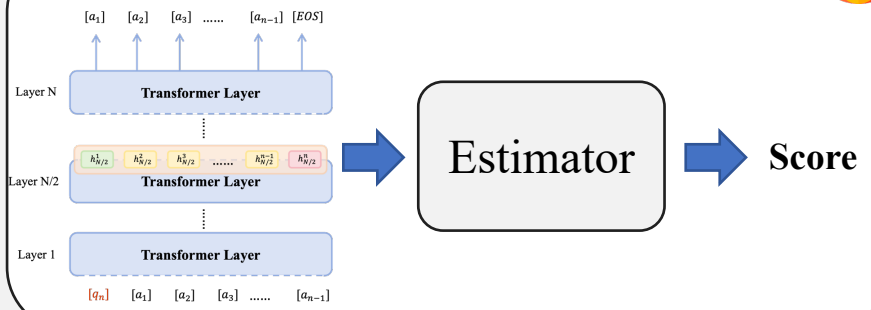
Response Sampling

Consistency Checking

Consistency Ratio

Discrete confidence; extensive sampling

Internal-State-Based



Continuous; limited generalization; requires ground-truth annotation

Verbalized Confidence

❑ Mitigating Overconfidence via Prompts

Ask LLMs to Be Prudent

Punish

Add "You will be punished if the answer is not right but you say certain"

Challenge

Challenge the correctness of the generated answer

Enhance QA Performance

Explain

Ask the model to explain the reason for its answer

Generate

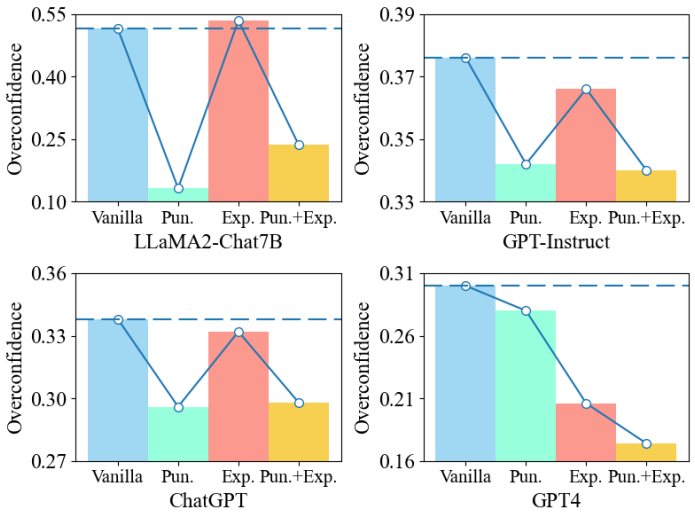
Ask the model to generate a short document that aids in answering the question

* These methods can be combined

When Do LLMs Need Retrieval Augmentation? Mitigating LLMs' Overconfidence Helps Retrieval Augmentation. ***ACL'2024***

Verbalized Confidence

Overconfidence Level after Mitigation



Effectiveness and Efficiency of Adaptive RAG

Model	Retrieval	NQ Static	NQ Vanilla	NQ Punish	NQ Explain	NQ Pun.+Exp.	HotpotQA Static	HotpotQA Vanilla	HotpotQA Punish	HotpotQA Explain	HotpotQA Pun.+Exp.
LLaMA2	RA Rate	100%	14.6%	71.4%	9.2%	51.8%	100%	44.8%	78.4%	46.2%	70.2%
	None	0.352	0.352	0.276	**0.382**	0.316	0.160	0.160	0.138	**0.186**	0.172
	Sparse	0.256	0.370	0.316	**0.390**	0.356	**0.334**	0.270	0.310	0.298	0.314
	Dense	**0.534**	0.414	0.522	0.418	0.494	0.288	0.244	0.276	0.276	**0.292**
	Gold	**0.774**	0.460	0.706	0.448	0.642	**0.516**	0.370	0.468	0.412	0.474
GPT-Instruct	RA Rate	100%	16.6%	21.4%	13.4%	16.8%	100%	18.0%	20.6%	12.0%	16.2%
	None	0.496	0.496	0.486	0.522	**0.528**	0.294	0.294	0.302	**0.378**	0.354
	Sparse	0.282	0.474	0.476	**0.516**	0.512	0.344	0.312	0.316	**0.390**	0.374
	Dense	0.538	0.518	0.520	0.538	**0.554**	0.324	0.306	0.306	**0.378**	0.362
	Gold	**0.816**	0.588	0.614	0.602	0.620	**0.568**	0.354	0.364	0.422	0.418
ChatGPT	RA Rate	100%	25.4%	33.2%	17.8%	22.6%	100%	52.6%	52.6%	39.4%	40.6%
	None	0.468	0.468	0.456	0.530	**0.536**	0.240	0.240	0.236	**0.326**	0.326
	Sparse	0.228	0.448	0.422	**0.510**	0.504	0.276	0.300	0.300	**0.360**	0.346
	Dense	0.506	0.490	0.488	0.550	**0.556**	0.238	0.276	0.266	**0.344**	0.336
	Gold	**0.800**	0.602	0.630	0.616	0.646	0.406	0.352	0.350	0.404	**0.412**
GPT-4	RA Rate	100%	13.6%	21.0%	20.8%	33.2%	100%	26.8%	35.0%	38.2%	51.8%
	None	0.592	0.592	0.538	**0.666**	0.650	0.404	0.404	0.414	**0.516**	0.484
	Sparse	0.572	0.610	0.600	**0.664**	0.634	0.546	0.464	0.478	0.566	**0.568**
	Dense	**0.698**	0.622	0.624	0.688	0.676	0.510	0.458	0.464	**0.540**	0.528
	Gold	**0.866**	0.676	0.680	0.756	0.764	**0.644**	0.500	0.530	0.616	0.620

Punish+Explain significantly reduces overconfidence by 8.8–55%.

Using Punish+Explain, triggering retrieval on **10–50%** of queries will **matching or outperforming** full RAG.

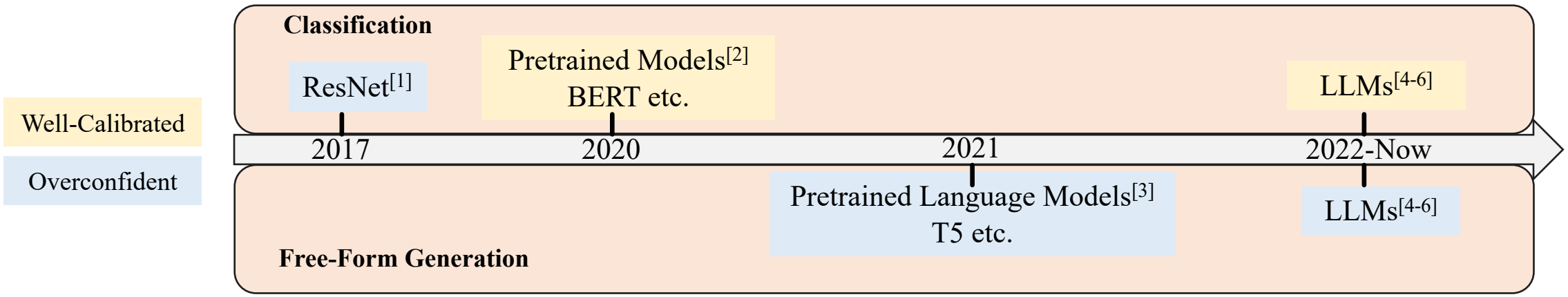
When Do LLMs Need Retrieval Augmentation? Mitigating LLMs' Overconfidence Helps Retrieval Augmentation. ***ACL'2024***

Token-Prob-Based Confidence

- How to measure
  - Classification: Probability of the generated choice/class
  - Free-form Generation: Probability of the generated tokens in the response

$$c = \exp\left(\frac{1}{T}\sum_{t=1}^{T}\log p_\theta^\pi(\tilde{r}_t \mid q, \tilde{r}_{<t})\right)$$

- Confidence value is continuous; Need binarization to determine when to retireve

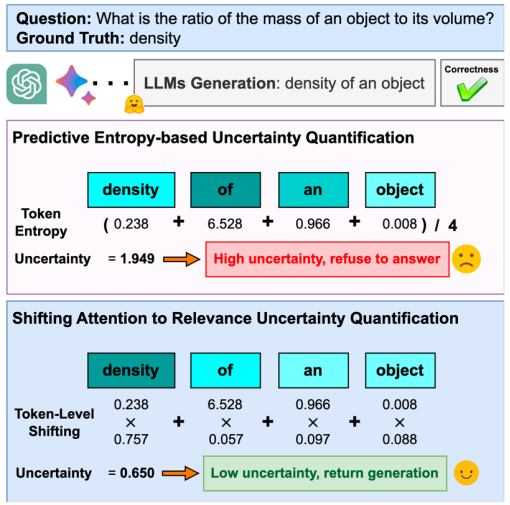


- Pre-training will make the classification or multi-choice QA less overconfident; but not for ResNet (probably due to the large amount of parameters)
- Generative models are all overconfident in free-form generation tasks.

[1] On Calibration of Modern Neural Networks. Chuan Guo et.al. ICML 2017
[2] Calibration of Pre-trained Transformers. Shrey Desai et.al. EMNLP 2020
[3] How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. Zhengbao Jiang et.al. TACL 2021
[4] Language Models (Mostly) Know What They Know. Saurav Kadavath et.al. Arxiv 2022
[5] Prompting GPT-3 To Be Reliable. Chenglei Si et.al. ICLR 2023
[6] Are Large Language Models More Honest in Their Probabilistic or Verbalized Confidence? Shiyu Ni et.al. CCIR 2024

Token-Prob-Based Confidence

- Shifting attention to important tokens[2]



- ➢ Important tokens contribute more to the sentence's semantics
- ➢ Assigns token weights as the self-NLI score when token is removed.

- Confidence calibration: temperature-scaling[1] & combined with external features[3]

Token-prob-based confidence is affected by both the **form** and **semantics** of the response.

However, only semantics matter!

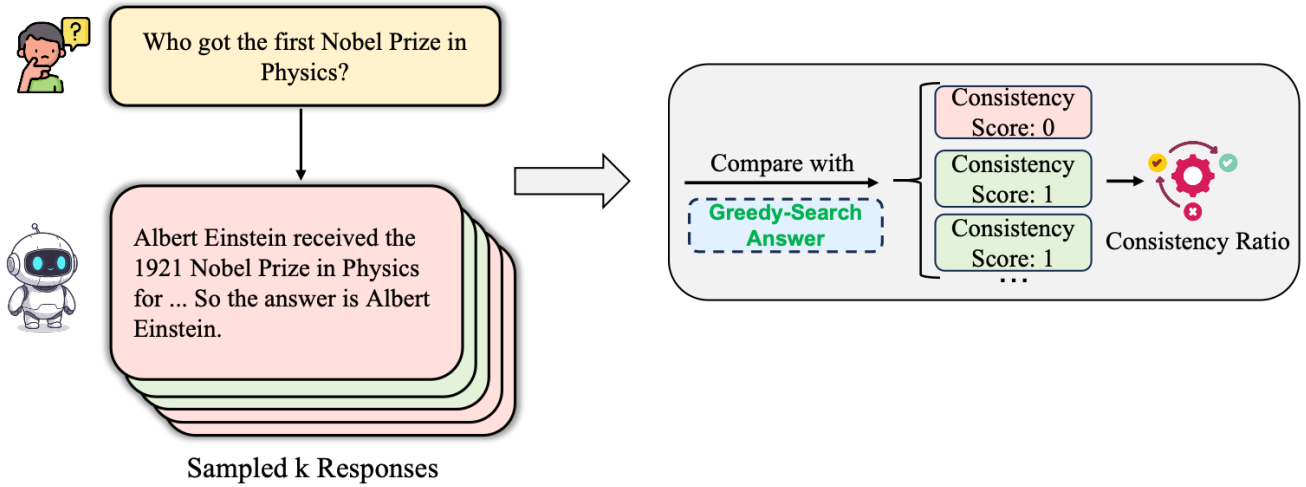
[1] On Calibration of Modern Neural Networks. ICML 2017

[2] Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models. ACL 2024

[3] How Knowledge Popularity Influences and Enhances LLM Knowledge Boundary Perception. Arxiv 2025

Self-Consistency-Based Confidence

❑ How to measure: Consistency across multiple sampled responses



- Similarity measurement
  - Lexical similarity[1]: Token overlap
  - Semantic similarity[2][3]: Measured by NLI models or LLMs

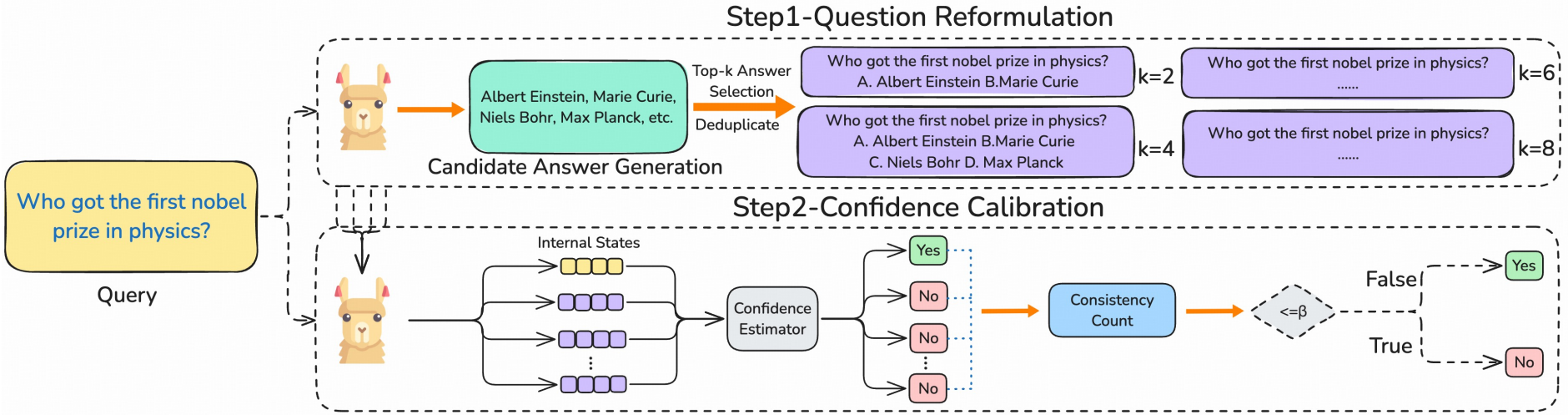
[1] Unsupervised Quality Estimation for Neural Machine Translation. TACL 2020

[2] Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. ICLR 2023

[3] SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. EMNLP 2023

Self-Consistency-Based Confidence

- **C**onfidence **C**onsistency-based **C**alibration (C^3)[2]
  - Converts open-ended questions into multiple-choice questions with variable option counts
  - Consistency across different questions



- Consistency **across outputs of different models**[1]
- Consistency across various **languages input**[3]

[1] SAC3: Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency. Jiaxin Zhang et.al. ***EMNLP 2023***

[2] Towards Fully Exploiting LLM Internal States to Enhance Knowledge Boundary Perception. Shiyu Ni et.al. ***ACL'2025***

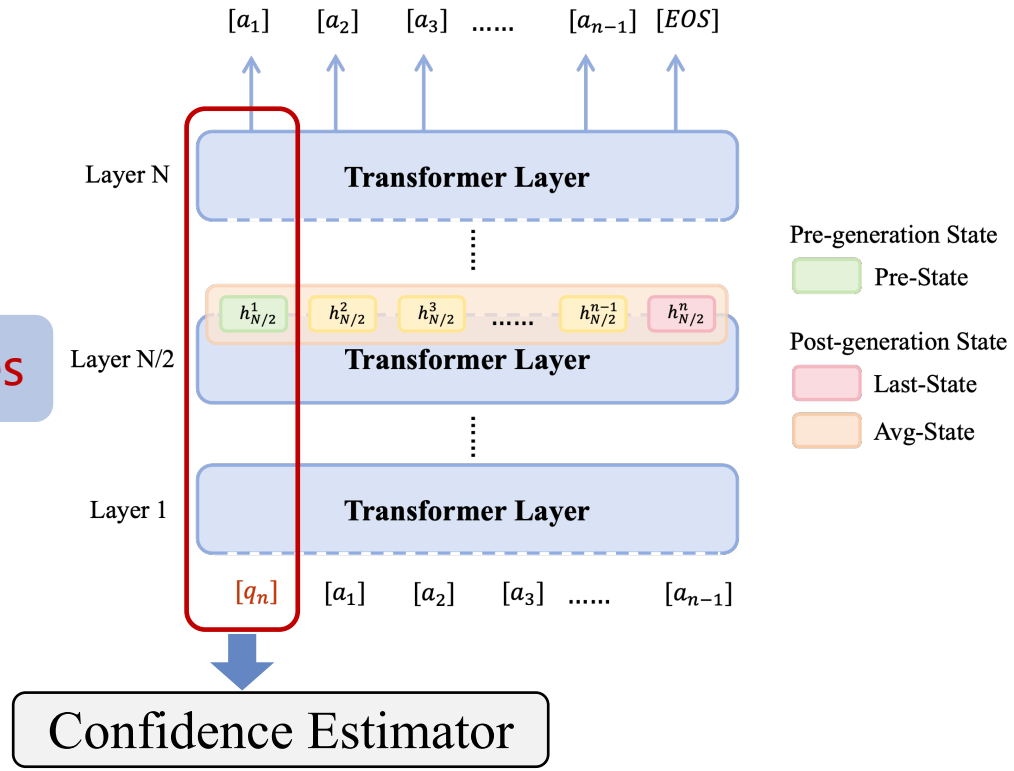
[3] Rowen: Adaptive Retrieval-Augmented Generation for Hallucination Mitigation in LLMs. Hanxing Ding et.al. ***SIGIR-AP 2025***

Internal-State-Based Confidence

- Train a confidence estimator based on internal states
  - a binary classifier
  - verbalized confidence

Training-based methods > zero-shot approaches

- Estimation before or after answer generation
  - Post-generation [1][2][3]
    - Costly to obtain
  - Pre-generation [4]
    - Only based on questions



Pre-generation vs. Post generation: **10% cost; 95% alignment performance**

[1] The Internal State of an LLM Knows When It's Lying. EMNLP 2023

[2] INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. ICLR 2024

[3] Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. ACL 2024

[4] Towards Fully Exploiting LLM Internal States to Enhance Knowledge Boundary Perception. ACL 2025

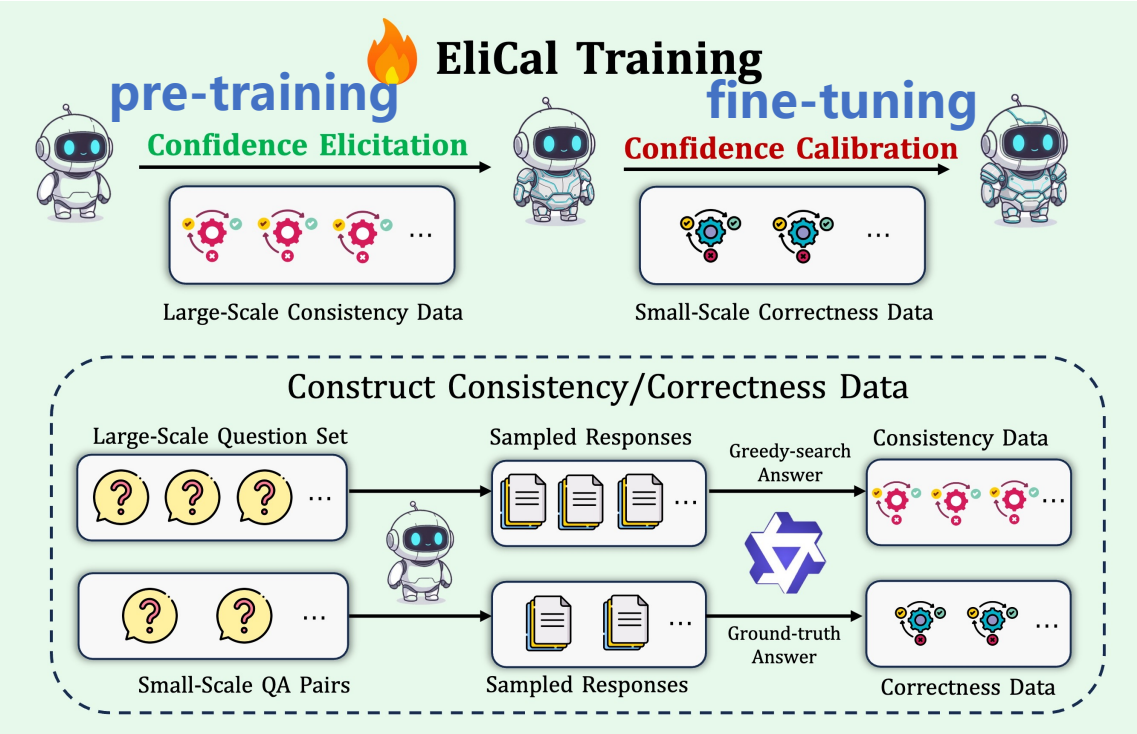
Comparisons Across Four Method Types

Confidence Types	Value	Need Training?	Need Ground-Truth Answer?	Performance Rank	Need Access to	Generalization
Verbalized	Discrete	No	No	4	Output Text	Good
Token-Prob-Based	Continuous	No	No	3	Output Logits	Good
Self-Consistency-Based	Discrete	No	No	2	Output Text	Good
Internal-State-Based	Continuous	Yes	Yes	1	Model Parameters	Bad

Pretraining for Confidence Calibration

Can we reduce the amount of ground-truth answers needed to obtain an effective and generalized confidence estimator?

Yes! **By incorporating an unsupervised pre-training phase.**



HonestyBench (560k training, 70k in-domain and out-of-domain evaluation merged from 10 representative QA datasets)

EliCal achieves ~**98%** of the upper bound performance using only 1k labeled samples (~**0.18%**)

Better generalization on other types of tasks (e.g., MMLU)

Annotation-Efficient Universal Honesty Alignment. ***Under submission to ICLR 2025***

Do LVLMs Know What They Know?

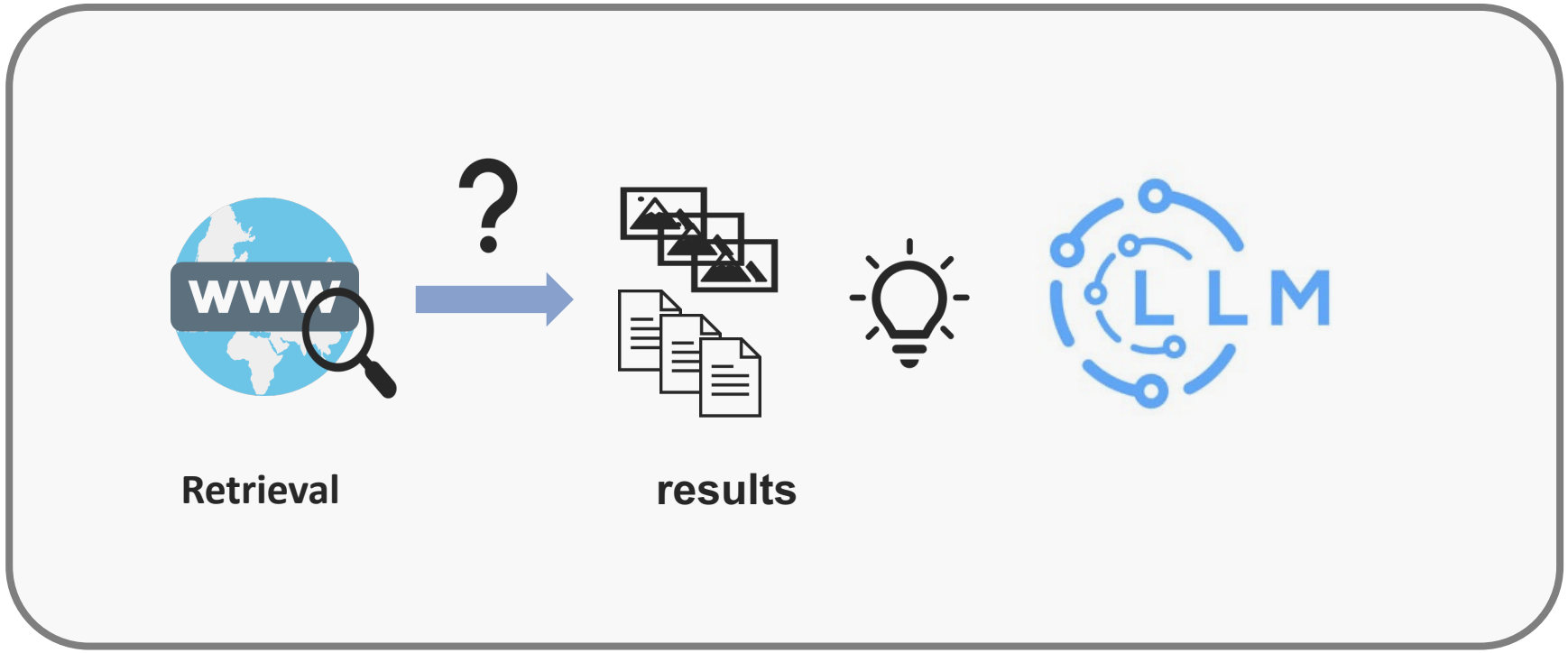
❑ *A Systematic Study of Knowledge Boundary Perception in LVLMs*

- Investigate three types of confidence (token-prob-based/verbalized/self-consistency)
- LVLMs can perceive their knowledge boundaries to some extent
- **Similarities to LLMs:**
  - Overconfident
  - Self-consistency/token-prob-based confidence > verbalized confidence
- **Differences from LLMs:**
  - More honest
    - difficult -> more uncertain
  - Some prompting methods do not work
    - Handling visual modality compromises their instruction-following ability

What episode of the cartoon marked the first time a deaf actor was cast on the show?

Response: The Heartbroke Kid

Answer: The Sound of Bleeding Gums



How to Retrieve for RAG?

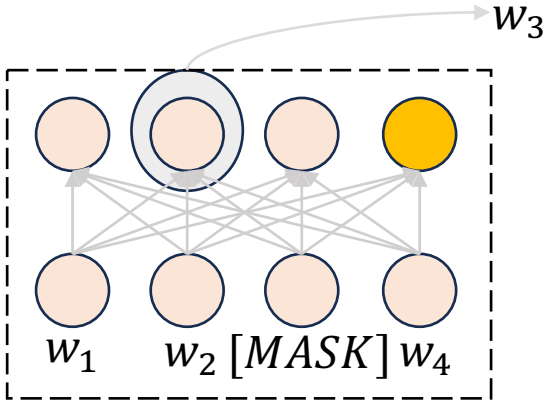
Better Retrieval Quality (Relevance)
Better for Generation (Utility)

LLMs as Retriever Backbones

- Pros: The retriever and generator based on the same backbone could facilitate better leverage of retrieval results by the generator [1].
- Pros: LLM-based retriever have higher potential in retrieving relevant results.
- Cons: The decoder-only structure makes later tokens invisible from former tokens.

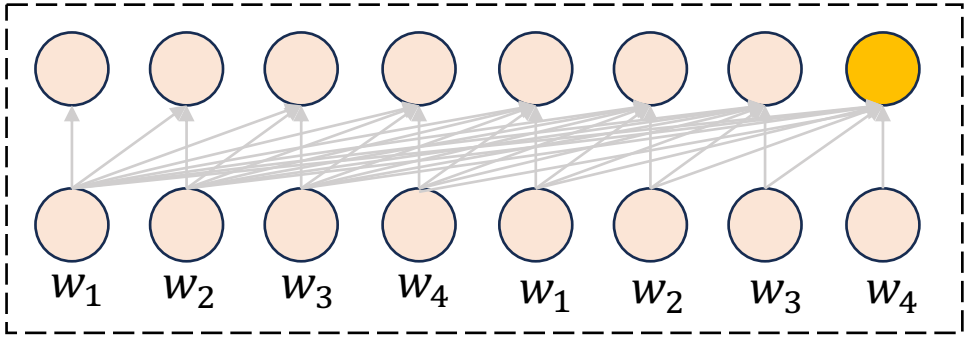
[1] SelfRACG: Enabling LLMs to Self-Express and Retrieve for Code Generation. EMNLP 2025

Text Retrieval



LLM2Vec[1]: Masked Next Token Prediction

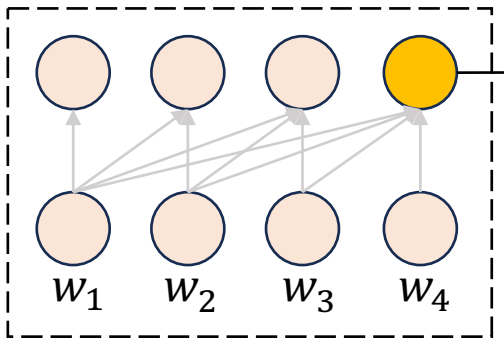
The original structure of the LLM was **changed**!



Echo[2]: Repetition improves language model embeddings

This is especially costly for documents.

Pre-training for IR



Projection

Multi-classification (Class Numbers: Vocabulary Size)

objective

Original sentence tokens Next sentence tokens

Llama2vec[3] → Larger data ~2.5M and ignoring the generation of the LLM itself

[1] Jacob Mitchell Springer, et al., Repetition improves language model embeddings. ICLR2025
[2] Parishad BehnamGhader, et al., LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. COLM2024
[3]Li, C., Liu, Z., Xiao, S., Shao, Y., & Lian, D. (2024, August). Llama2vec: Unsupervised adaptation of large language models for dense retrieval. ACL2024.

Unleashing the Power of LLMs in Dense Retrieval

LLMs are generative methods!!

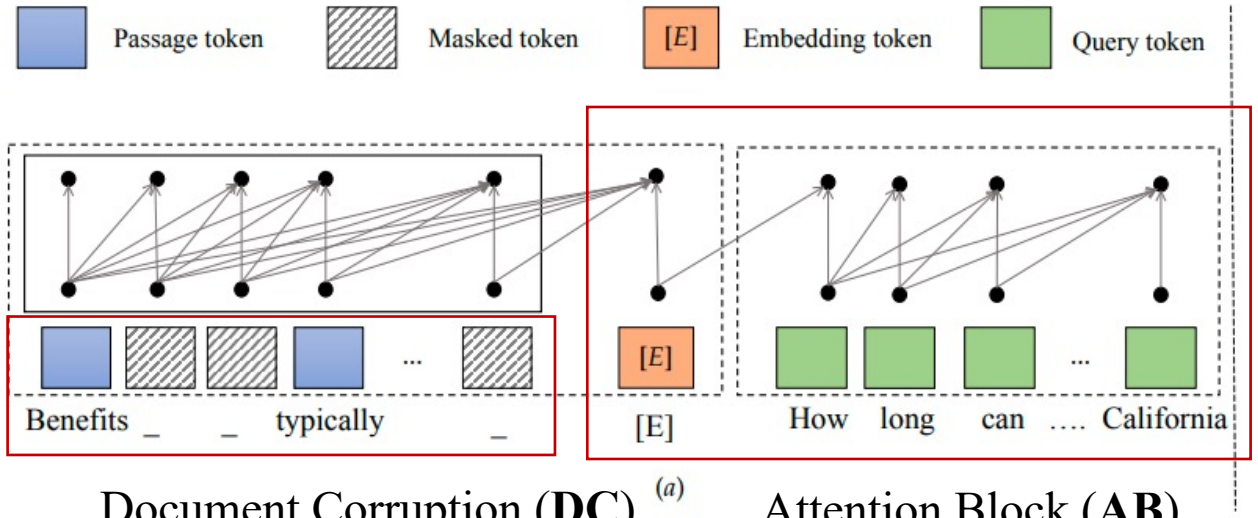
Query likelihood model:

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \propto P(q|d)$$

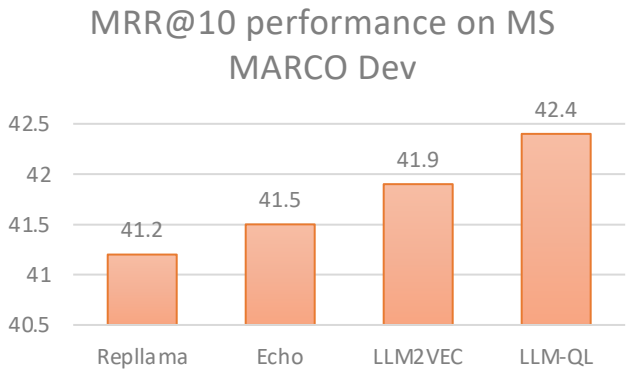
A generative method to estimate relevance between query and document.

Integrating **query likelihood as an auxiliary training task** for document representation

LLM-QL



Document Corruption (**DC**) $^{(a)}$ Attention Block (**AB**)



Significant performance improvement compared to bidirectional methods and repeating inputs.

Multi-Modal LLMs as Cross-Modal Retrievers

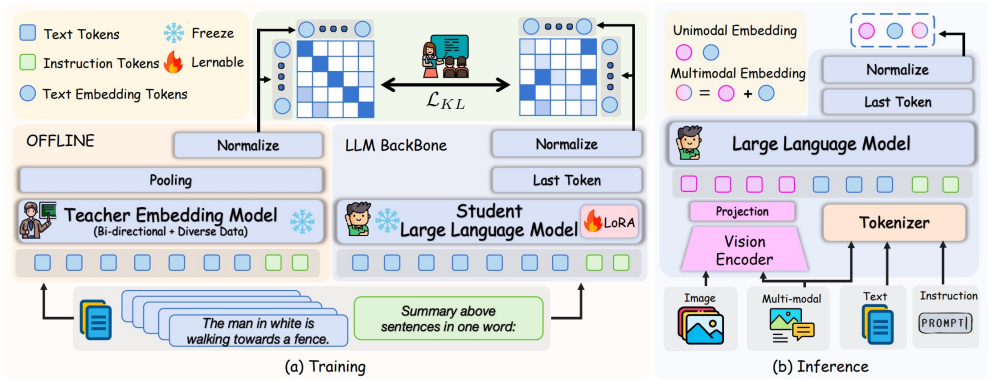
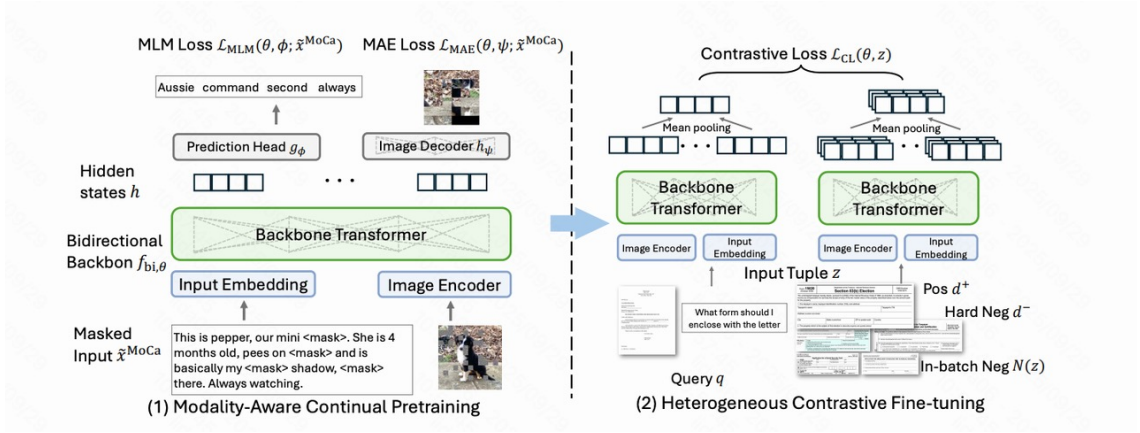


Figure 2: The framework of the Textual Discriminative Knowledge Distillation stage. We leverage the state-of-the-art LLM-based embedding model to enhance the discriminative capabilities of the MLLM's language component.



UniME [2]: Distillation from text-embedding LLMs

MoCa [3]: Decoder-only to **bidirectional**

May not necessarily generalize to other modalities

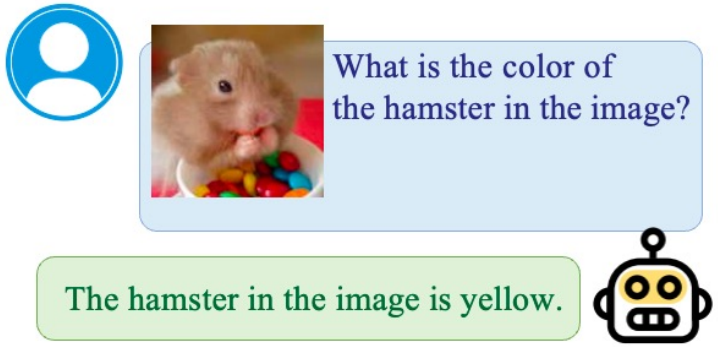
Requires a large amount of data: ~ 30B tokens

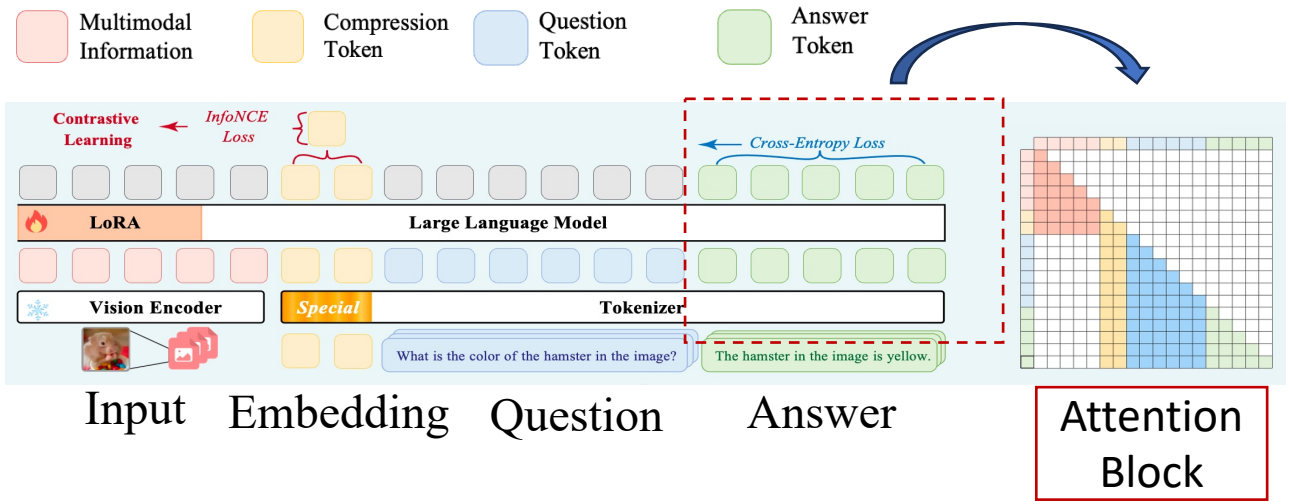
[1] Alec Radford, et al., Learning Transferable Visual Models From Natural Language Supervision. PMLR2021
[2] Haonan Chen, et al., Moca: Modality-aware Continual Pre-training Makes Better Bidirectional Multimodal Embeddings
[3] Gu, T., Yang, K., Feng, Z., Wang, X., Zhang, Y., Long, D., ... & Deng, J. (2025, October). Breaking the modality barrier: Universal embedding learning with multimodal llms. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 2860-2869).

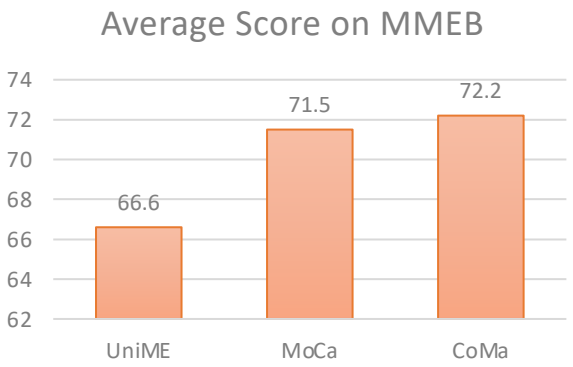
Pretraining Multi-Modal LLMs for Retrieval

CoMA

- Annotating diverse QA pairs covering different content of the images with MLLMs







With a small amount of pre-training data (0.3B tokens), CoMa can achieve better performance than MoCa (30B tokens.)

From Relevance to Utility - Utility-Focused Result Selection for RAG

Relevance **aboutness** **vs** Utility **value**

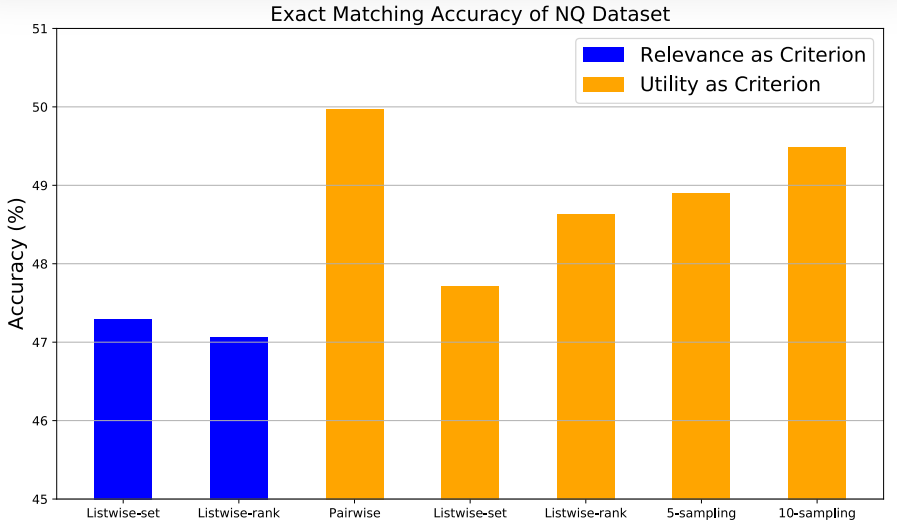
Effectiveness Measures

Two sets of measures for evaluating the effectiveness of a search have been used, based on the two most often used criteria:

1. *Relevance:* the degree of fit between the question and the retrieved item. The criteria of "*aboutness*" is used.
2. *Utility:* the degree of actual usefulness of answers to an information seeker. The criteria used is the *value* to the information seeker.

Saracevic, Tefko, et al. "A study of information seeking and retrieving. I. Background and methodology." *Journal of the American Society for Information science* 39.3 (1988): 161-176.

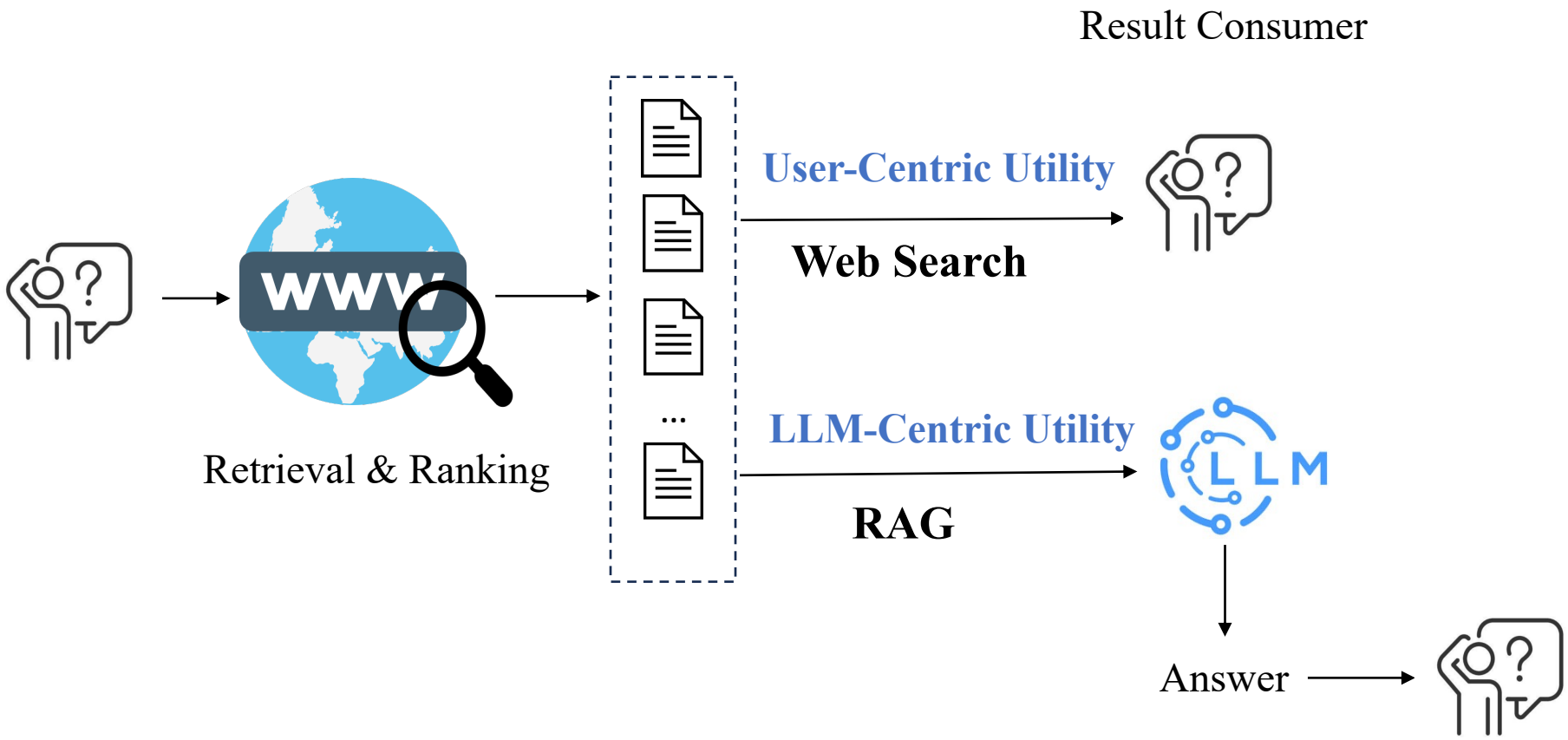
Utility in RAG: emphasizes the **usefulness** of a passage in facilitating the generation of an **accurate and comprehensive answer** to the question.



Zhang, Hengran, et al. "Are Large Language Models Good at Utility Judgments?." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024 (SIGIR'24). pp. 1941-1951

Using utility as criterion for selection yields **superior generation performance** over relevance

Utility-Focused Result Selection for RAG



LLM-Centric Utility-Based Result Selection for RAG

Utility Judgments

w.r.t. answer

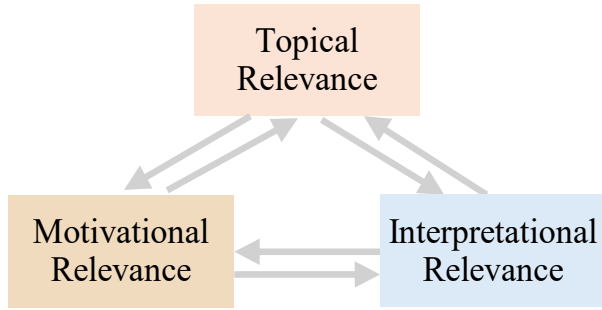
d_1 

d_2 

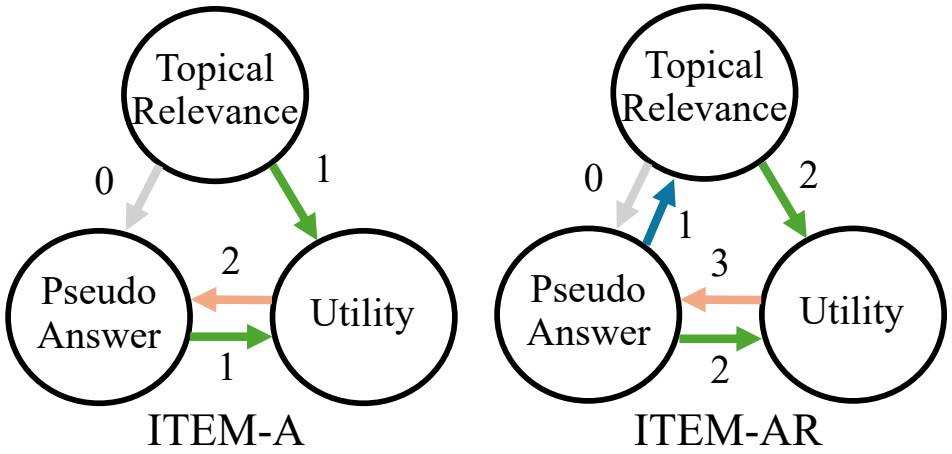
- **Judged by LLMs verbalized with pseudo answers** [1][2]
- Likelihood [3][5] of generating ground-truth answers
- Attention [4] to input documents

Costly; cannot scale to larger candidate sets

Relevance in Philosophy

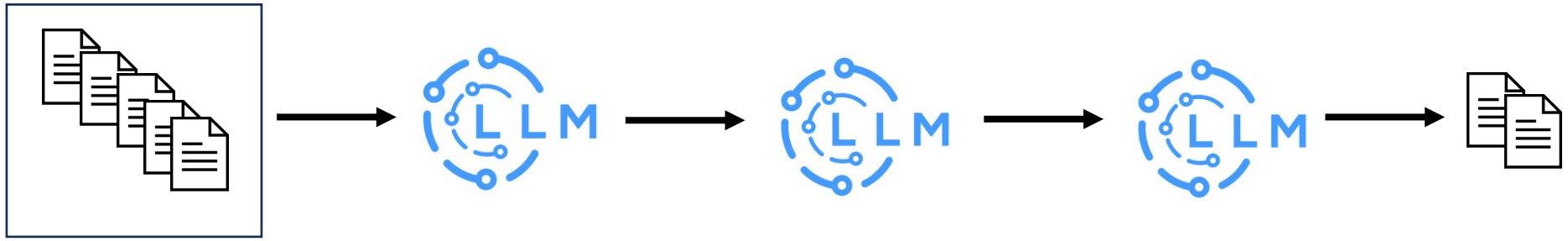


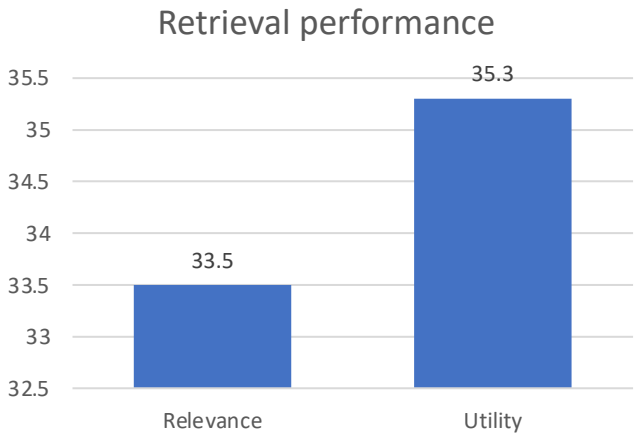
ITEM:Iterative utiliTy judgmEnt fraMework [2]



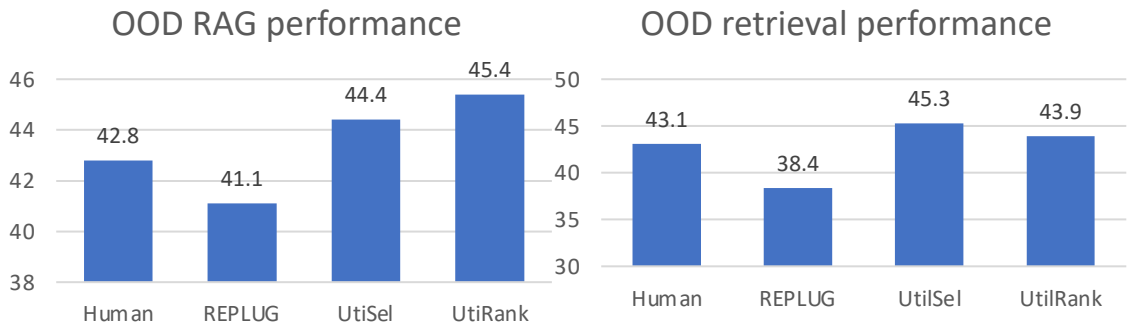
[1] Zhang, H., Zhang, R., Guo, J., de Rijke, M., Fan, Y., & Cheng, X. (2024, July). Are Large Language Models Good at Utility Judgments?. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1941-1951).
[2] Zhang, H., Bi, K., Guo, J., & Cheng, X. (2024). Iterative Utility Judgment Framework via LLMs Inspired by Relevance in Philosophy. *2024*.
[3] Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., ... & Yih, W. T. (2024, June). Replug: Retrieval-augmented black-box language models. NAACL2024
[4] Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., ... & Grave, E. (2023). Atlas: Few-shot learning with retrieval augmented language models. Journal of Machine Learning Research, 24(251), 1-43.
[5] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.

Training a Utility-Focused Retriever with LLM Judgments





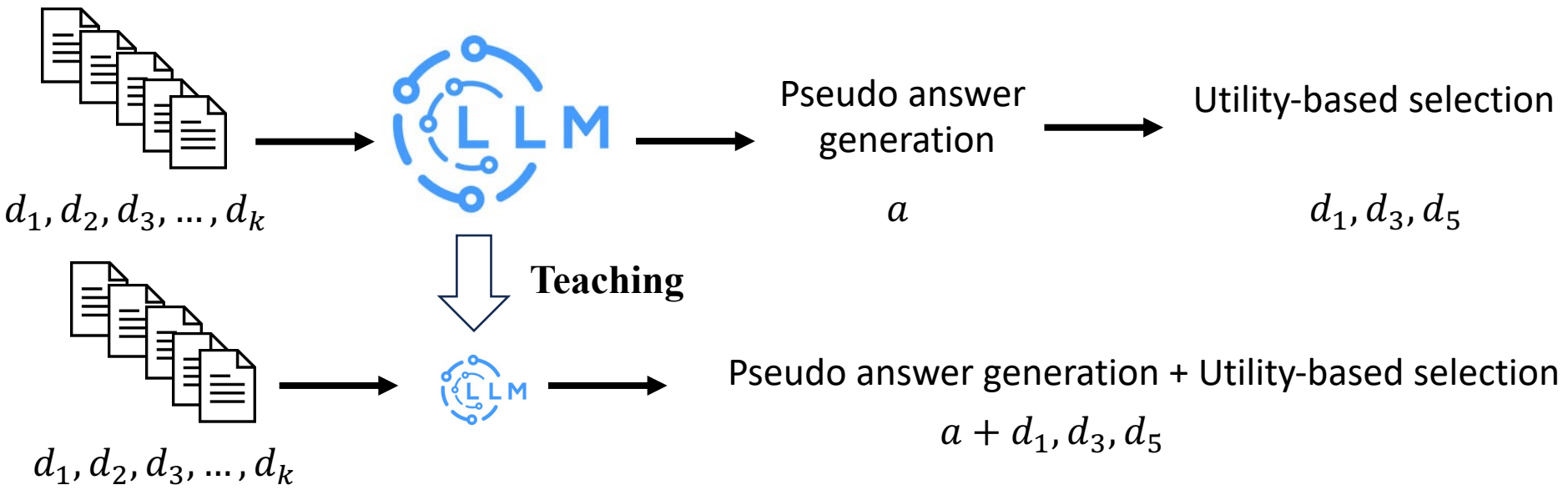
Utility-focused annotation is much better than relevance-focused annotation.



Utility-focused annotation has better generalization performance in retrieval and RAG

Zhang, H., Bi, K., Guo, J., Liu, S., Shi, D., Yin, D., & Cheng, X. (2025) Leveraging LLMs for Utility-Focused Annotation: Reducing Manual Effort for Retrieval and RAG. EMNLP2025

Distilling a Small Utility-Focused Selector for RAG

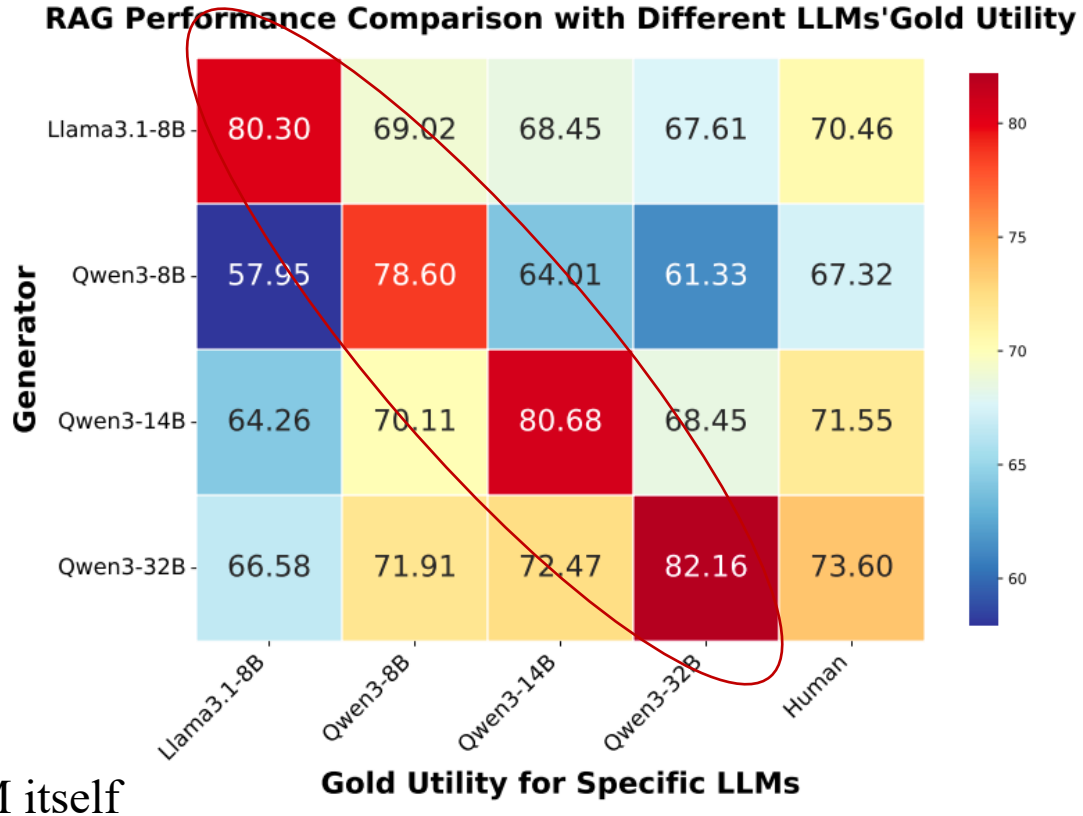


Scale utility-focused selection to a large number of candidate passages (from ~20 to 100+)

Zhang, H., Bi, K., Guo, J., Zhang, J., Wang, S., Yin, D., & Cheng, X. (2025). Distilling a Small Utility-Based Passage Selector to Enhance Retrieval-Augmented Generation. *SIGIR-AP25*

From General Utility to LLM-specific Utility

Gold utilitarian passage for specific LLM:
The inclusion of this passage leads to a performance gain in RAG performance.

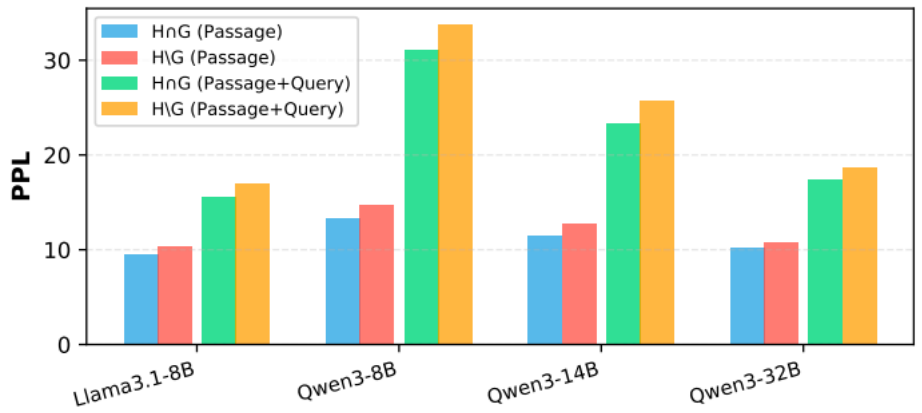


- RAG Performance Ranking:
  - 1st (Best): Gold utilitarian passages from the LLM itself
  - 2nd: Human-annotated positive passages
  - 3rd (Worst): Passages from other LLMs, especially other series

Zhang, H., Bi, K., Guo, J., Zhang, J., Wang, S., Yin, D., & Cheng, X. (2025). Distilling a Small Utility-Based Passage Selector to Enhance Retrieval-Augmented Generation. ArXiv. 2025

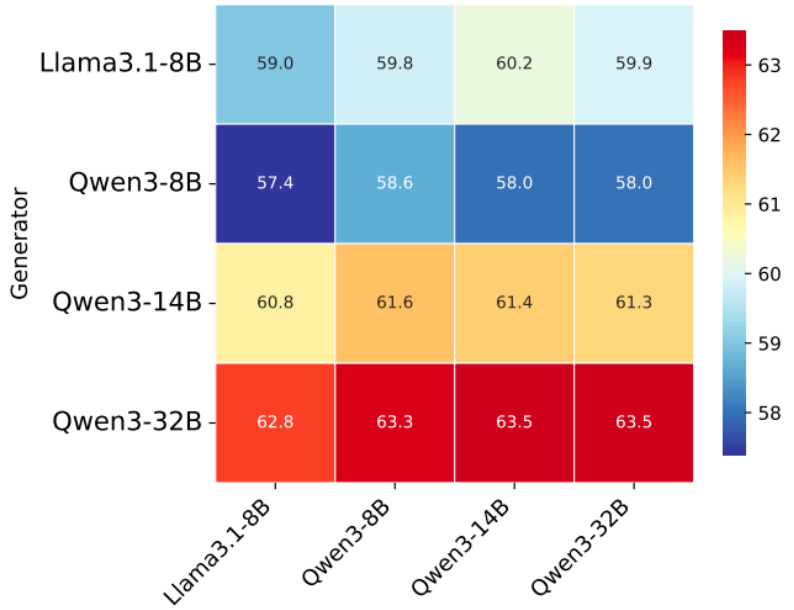
From General Utility to LLM-specific Utility

Among human-annotated positives:
LLM-specific ground-truth vs. others



- LLM cannot understand or take sufficient use of a passage that humans consider useful.
- PPL of gold passages < others

QA performance based on LLM-specific utility-based selection



- The utilitarian passages selected by existing methods do not appear to be sufficiently LLM-specific.
- We need to explore better methods for enhancing LLM-specific utility judgment performance!

Zhang, H., Bi, K., Guo, J., Zhang, J., Wang, S., Yin, D., & Cheng, X. (2025). Distilling a Small Utility-Based Passage Selector to Enhance Retrieval-Augmented Generation. ArXiv. 2025

Summary

- ❑ When should LLMs invoke retrieval augmentation?
  - ➢ When the model does not possess the internal knowledge
  - ➢ Enhance the model's awareness of its knowledge boundaries
    - Prompt the model to be more cautious
    - Consistency-based methods
    - Pretraining-based method (EliCal on HonestyBench)

- ❑ How to provide more beneficial results for LLM generation?
  - ➢ LLM-based retrievers
    - Modeling query-likelihood (text)
    - image QA data augmentation (multi-modal)
  - ➢ Utility-based result retrieval/selection
    - Iterative utility judgment framework
    - Utility-oriented retrievers/selectors for RAG
  - ➢ From General utility to LLM-specific Utility

Q&A

Contact US: **T**rustworthy **I**nfor**M**ation acc**E**ss (TIME) https://stay-hungry-time.github.io/