

Large Language Model Applications

Logistics

Our goals for II-766 students:

- Build critical thinking skills for figuring out which (if any) LLM is the right tool for any particular application
- Learn the different ways you can customize base LLM technology to different tasks and applications you care about
- Prepare you for industry jobs where you will be designing and building applications that rely on LLMs
 - (or prepare you for research positions where you'll be testing how choices in LLM implementation affect the performance/usefulness/etc. of different applications)

Whenever you use AI for this class, you should be asking yourself:

Is this helping me to learn more effectively, or is it helping me to get out of learning?

Our goals for II-766 students:

- Build critical thinking skills for figuring out which (if any) LLM is the right tool for any particular application
- Learn the different ways you can customize base LLM technology to different tasks and applications you care about
- Prepare you for industry jobs where you will be designing and building applications that rely on LLMs
 - (or prepare you for research positions where you'll be testing how choices in LLM implementation affect the performance/usefulness/etc. of different applications)

If your learning goals don't at least somewhat overlap with the ones on this slide, you will have trouble succeeding in II-766.

An Apology from Us

The template we provided for the project pitch did not contain a place for you to mark down acceptable use of AI. Because we failed to provide this, we will not be enforcing the following:

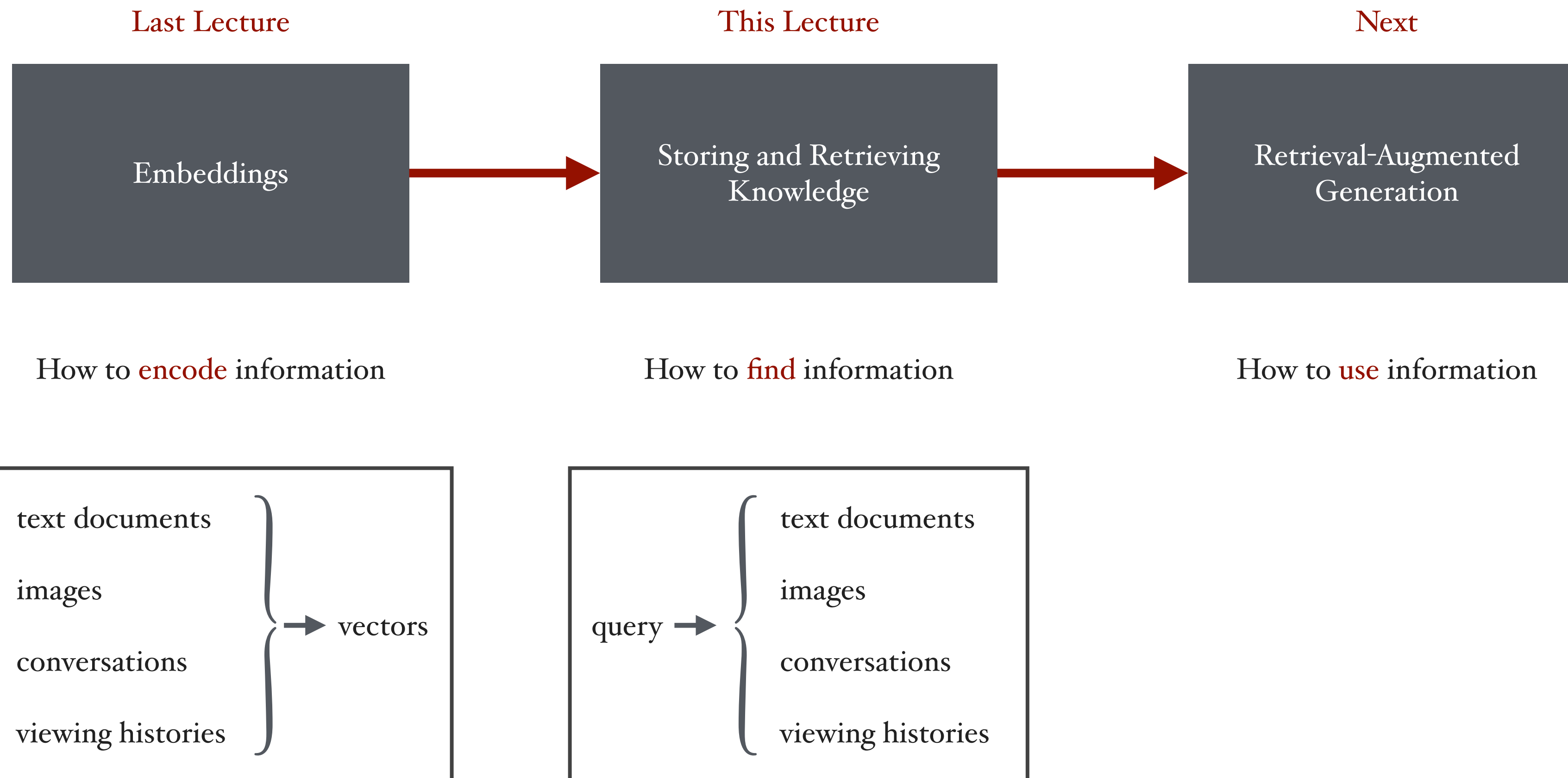
—

If you do use AI systems to help with the homework, please fill out the “Use of AI” question on the homework to help the instructors understand your usage. Use of AI without making an honest attempt at answering this question will be considered a violation of academic integrity.

Large Language Model Applications

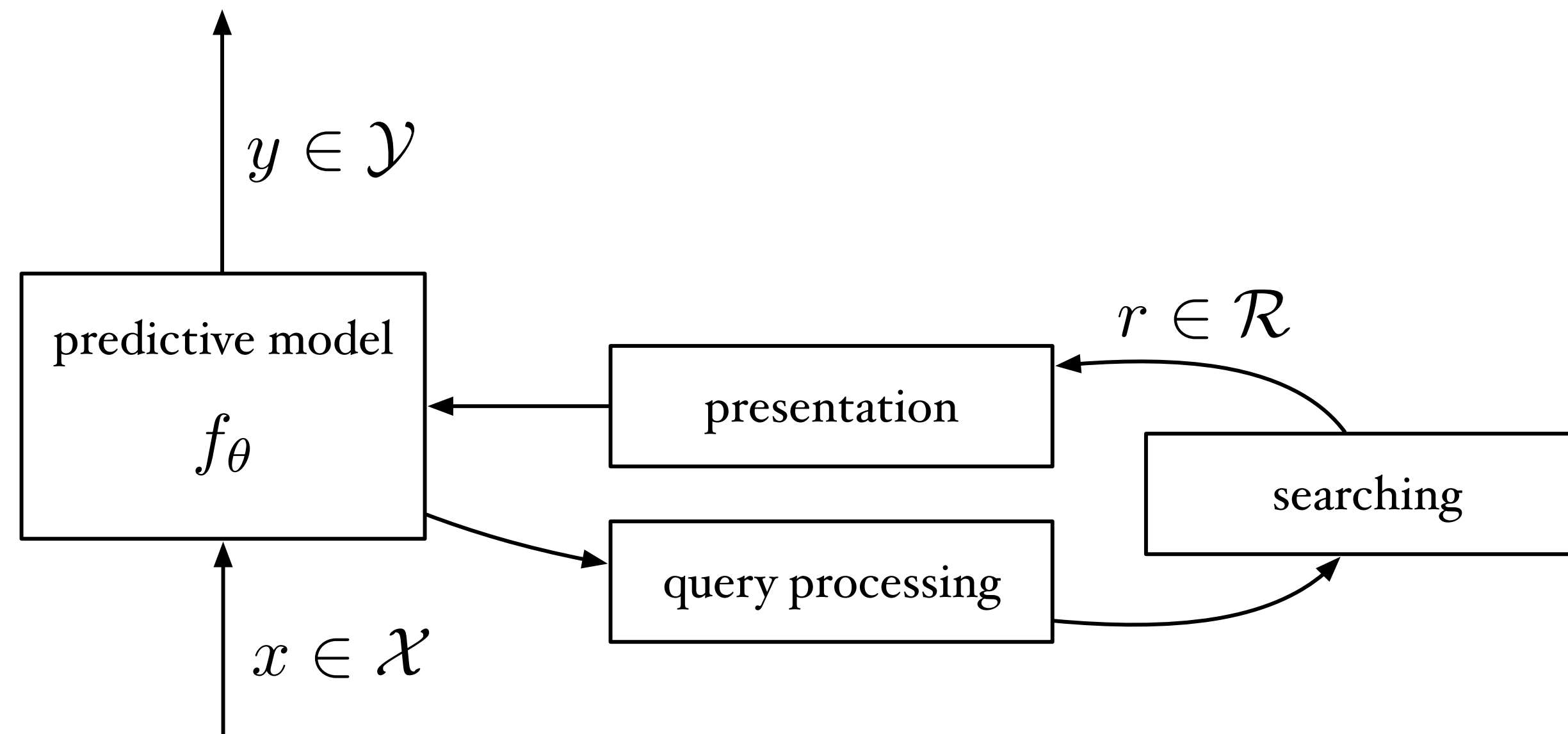
Retrieval Augmented Generation (2)
Deep Research

From Embeddings to Knowledge Access

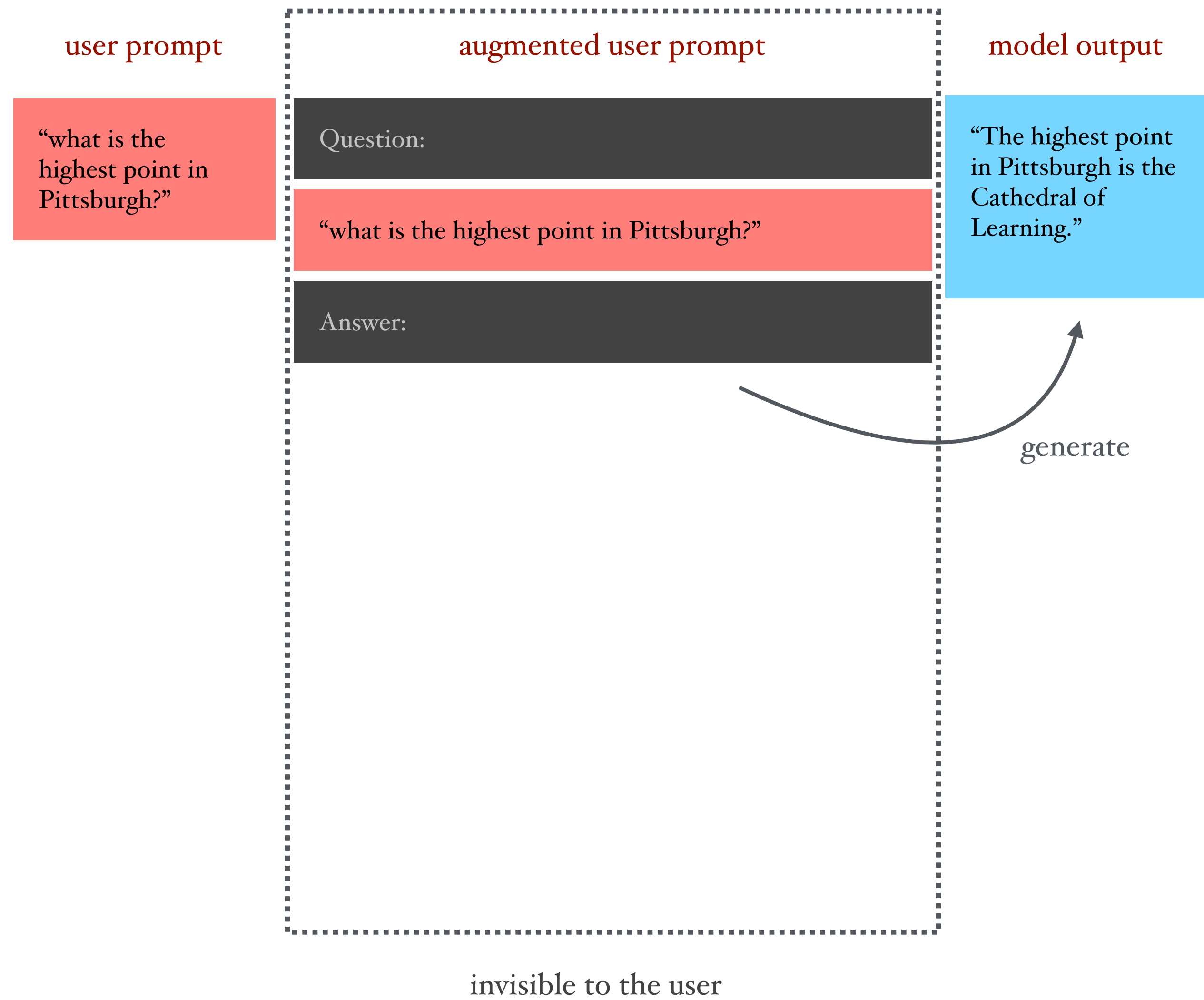


Retrieval-enhanced machine learning

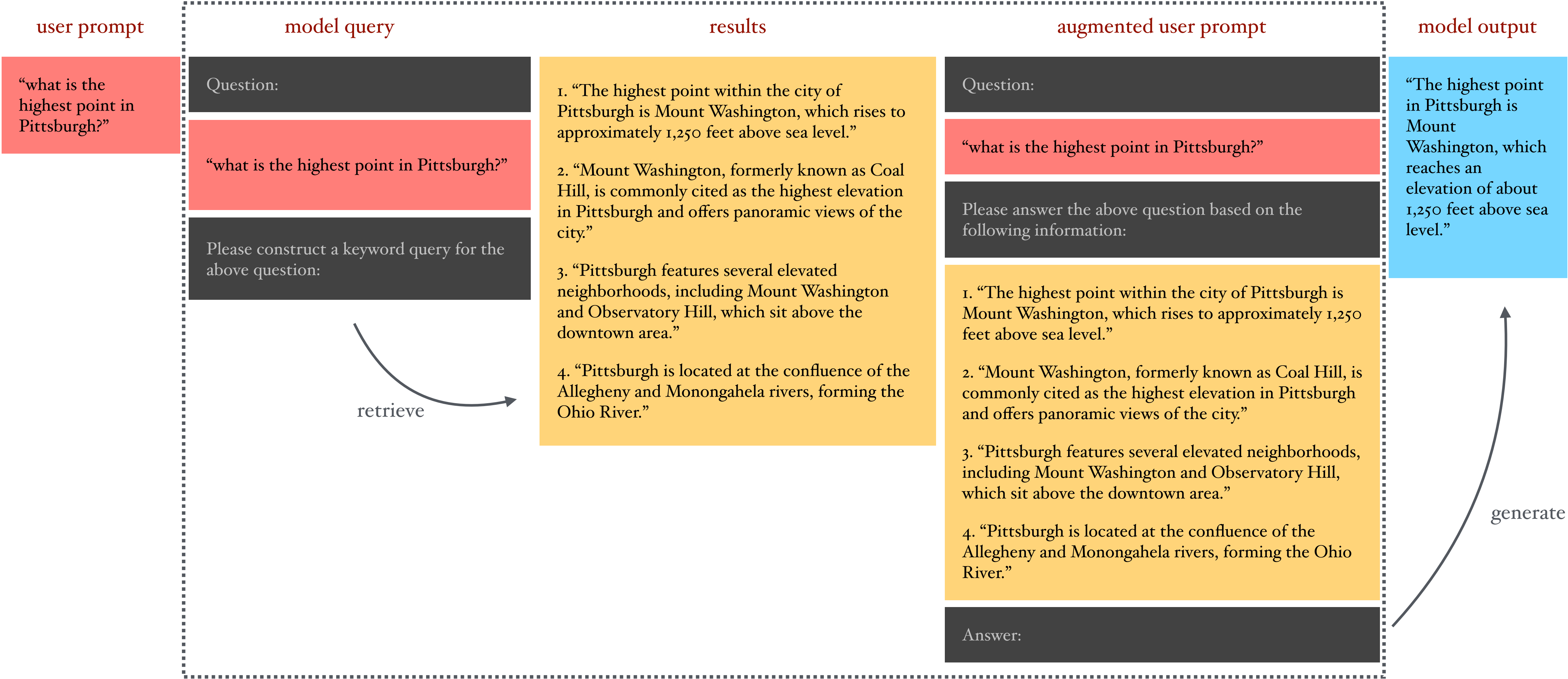
In LLMs the presented results are incorporated into the generator's prompt for the final answer



in-context RAG



Retrieval becomes prompt engineering: generator sees retrieved text in its context window

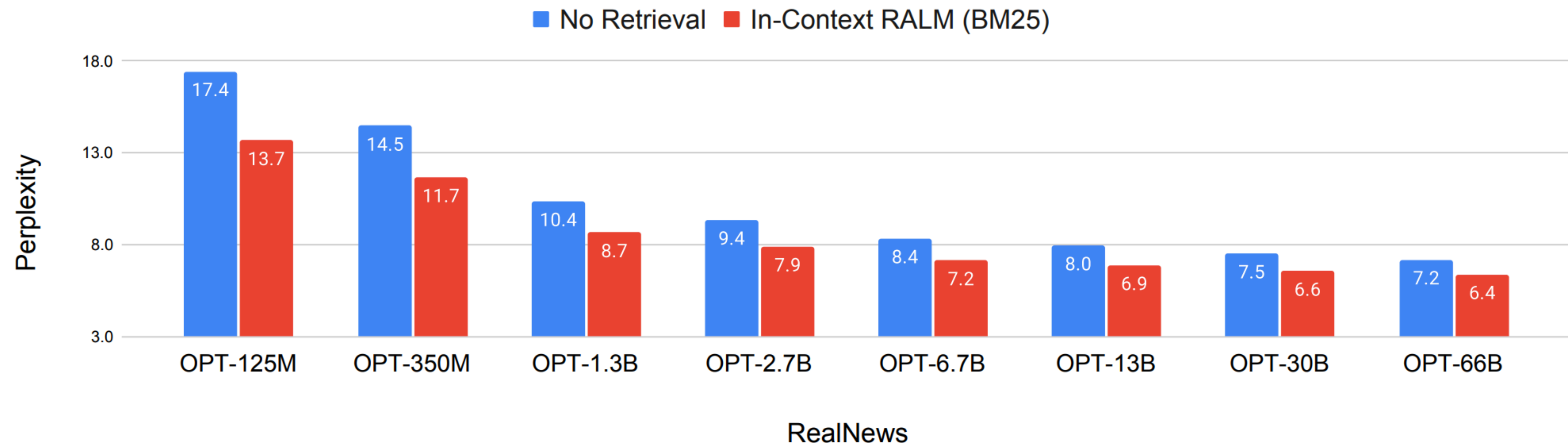




In-context RAG

Simply augmenting input-space of LMs give significant gain across different tasks

- In upstream language modeling task, simply adding retrieved context gives large gains, especially smaller models
- Similar significant gains in downstream tasks such as Question Answering

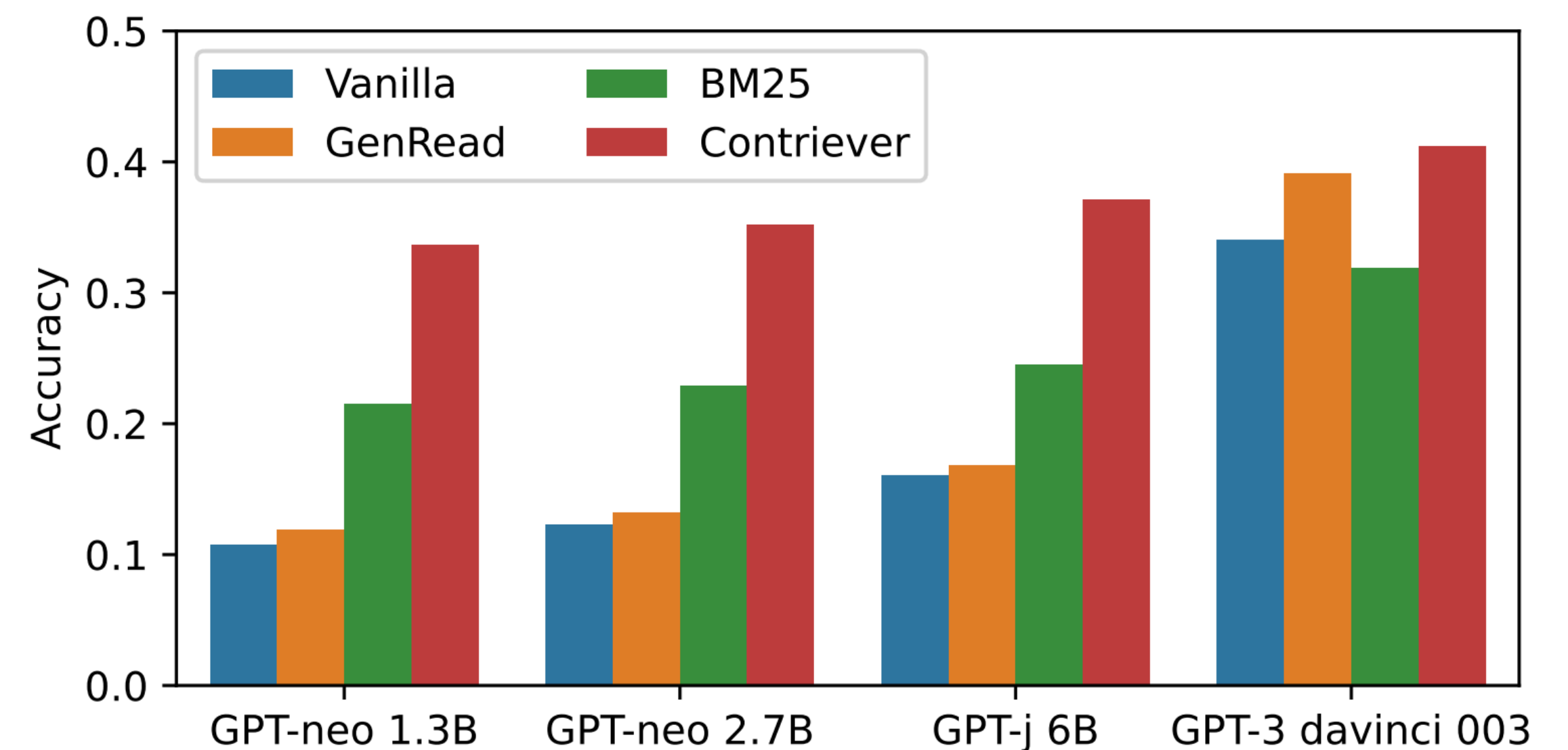
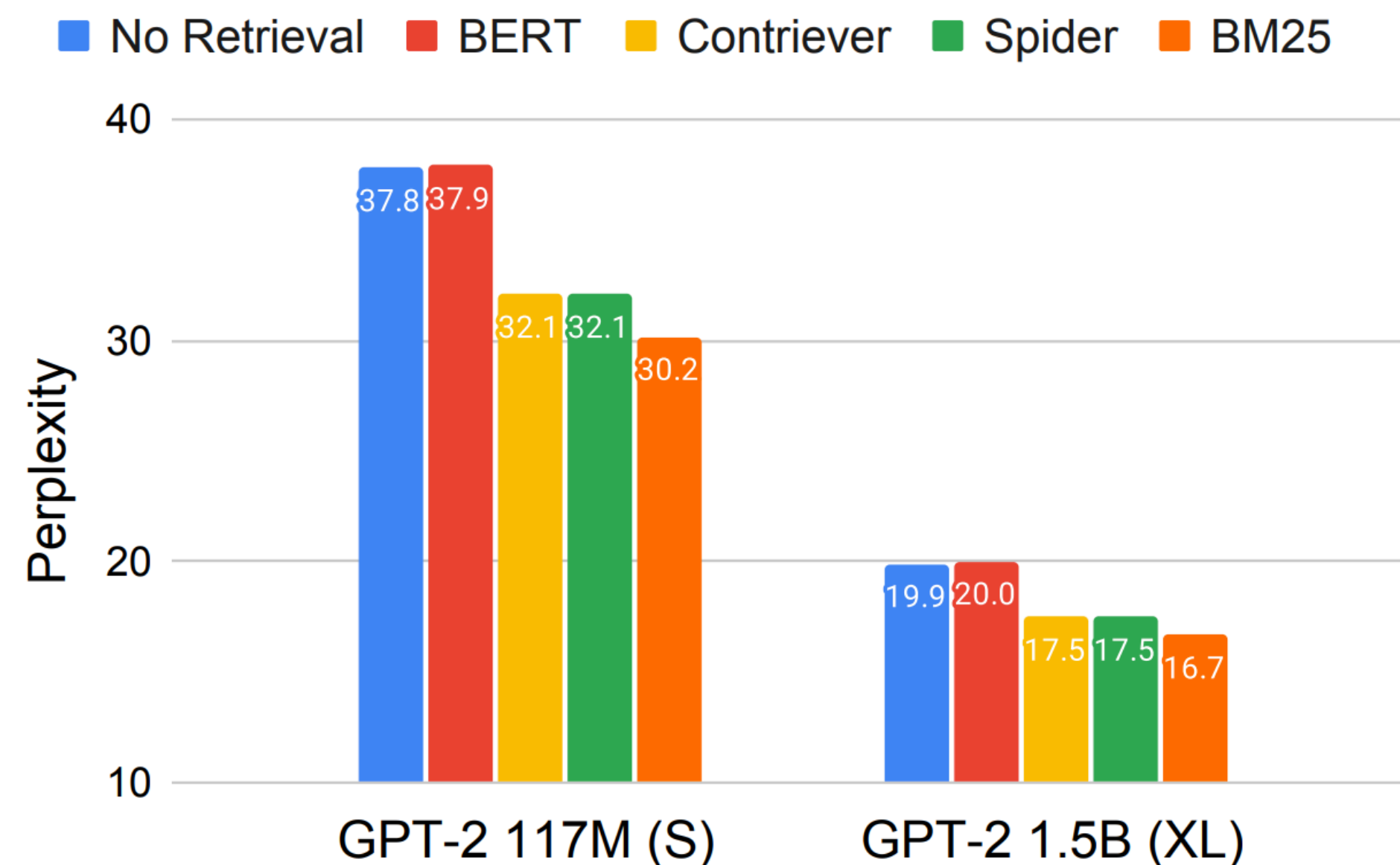




In-context RAG

Effects of retrieval systems for downstream task performance

- On language modeling, BM 25 results in best performance
- On downstream QA tasks, trained retrieval models eg Contriever results in best performance



poll: in-context cheating

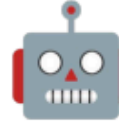


Limitations of naïve RAG

Effects of retrieval systems for downstream task performance

- In-context retrieval-augmented LMs sometimes generate content that is not fully supported by their citations

 What are the latest discoveries from the James Webb Space Telescope?



The James Webb Space Telescope is designed to peer into the dusty clouds of gas where stars and planetary systems are born. Webb has captured the first direct image of an exoplanet, and the Pillars of Creation in the Eagle Nebula[1][2]. Additionally, the telescope will be used to study the next interstellar interloper[3].

(*Some generated statements may *not be fully supported* by citations, while others are *fully supported*.)

Cited Webpages

[1]:  nasa.gov (✗ citation does not support its associated statement)
[NASA's Webb Confirms Its First Exoplanet](#)
... Researchers confirmed an exoplanet, a planet that orbits another star, using NASA's James Webb Space Telescope for the first time. ...

[2]:  cnn.com (⚠ citation partially supports its associated statement)
[Pillars of Creation: James Webb Space Telescope ...](#)
... The Pillars of Creation, in the Eagle Nebula, is a star-forming region captured in a new image (right) by the James Webb Space Telescope that reveals more detail than a 2014 image (left) by Hubble ...

[3]:  nasa.gov (✓ citation fully supports its associated statement)
[Studying the Next Interstellar Interloper with Webb](#)
...Scientists have had only limited ability to study these objects once discovered, but all of that is about to change with NASA's James Webb Space Telescope...The team will use Webb's spectroscopic capabilities in both the near-infrared and mid-infrared bands to study two different aspects of the interstellar object.



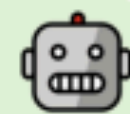
Limitations of naïve RAG

Effects of retrieval systems for downstream task performance

- In-context retrieval-augmented LMs sometimes generate content that is not fully supported by their citations
- They can easily be distracted by unhelpful context

Q: Who is the actor playing Jason on general hospital?

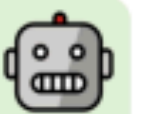
Large Language Model (no retrieval)



The answer is: Steve Burton



Retrieval Augmented Language Model



E: Jason Gerhardt (born April 21, 1974) is an American actor. He is known for playing the role of Cooper Barrett in General Hospital and Zack Kilmer in Mistresses.

The answer is: Jason Gerhardt





Limitations of naïve RAG

Effects of retrieval systems for downstream task performance

- In-context retrieval-augmented LMs sometimes generate content that is not fully supported by their citations
- They can easily be distracted by unhelpful context
- Diverse tasks require different retrieval needs e.g., content, frequency

Who is the current PM of UK?

Can be easily answered based on top documents retrieved at the beginning

Create a table listing all previous UK Prime Ministers, including their terms in office, political party, alma mater, and notable achievements.

This may require iterative retrieval, based on the current generation

The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts (from MATH)

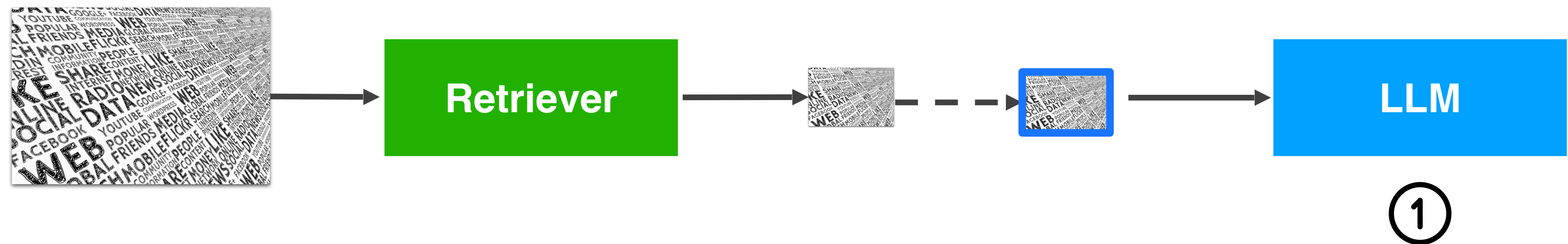
Questions with similar solutions may have limited semantic similarities in embedding space



Designing and training more reliable LLM RAG

Approaches to optimize (1) LM, (2) Retrievers, or (3) prompts for LLM RAG

1. Optimizing LLMs for RAG: training / controlling LLMs with retrieved context





Self-RAG: Teaching LLMs to learn to leverage retrieved context

Self-RAG teaches LMs to adaptively retrieve and evaluates context & own generation

Train an arbitrary LM (e.g., Llama 3) to generate special tokens for (1) triggering retrieval only when necessary and (2) evaluating the relevance of retrieved context and its own generations.

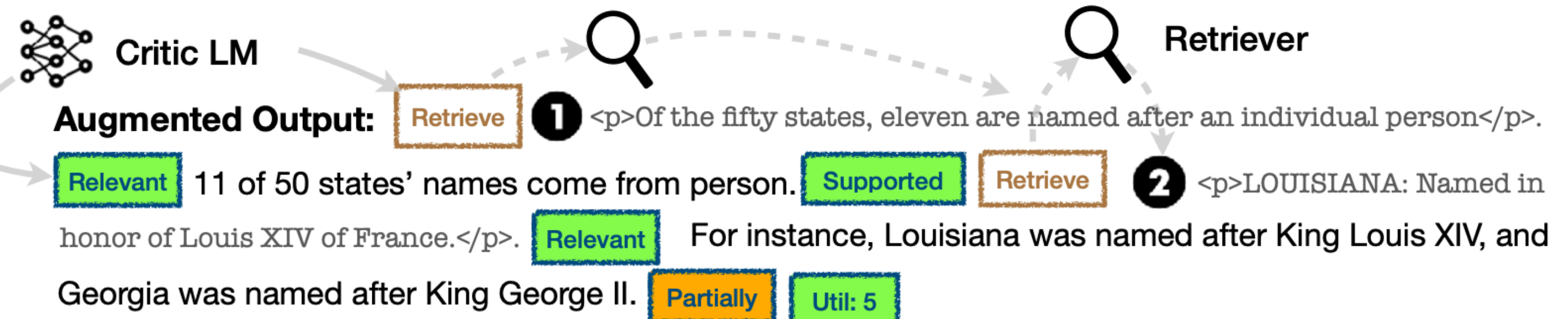
Input: Write an essay of your best summer vacation

Output: My best summer vacation was a magical escape to the coastal town of Santorini. The azure waters, charming white-washed building are unforgettable.

Augmented Output: No Retrieval My best summer vacation was a magical escape to the coastal town of Santorini. No Retrieval The azure waters, charming white-washed building are unforgettable experience. Util: 5

Input: How did US states get their names?

Output: 1 of 50 states names come from persons. For instance, Louisiana was named in honor of King Louis XIV of France and Georgia was named after King George II.

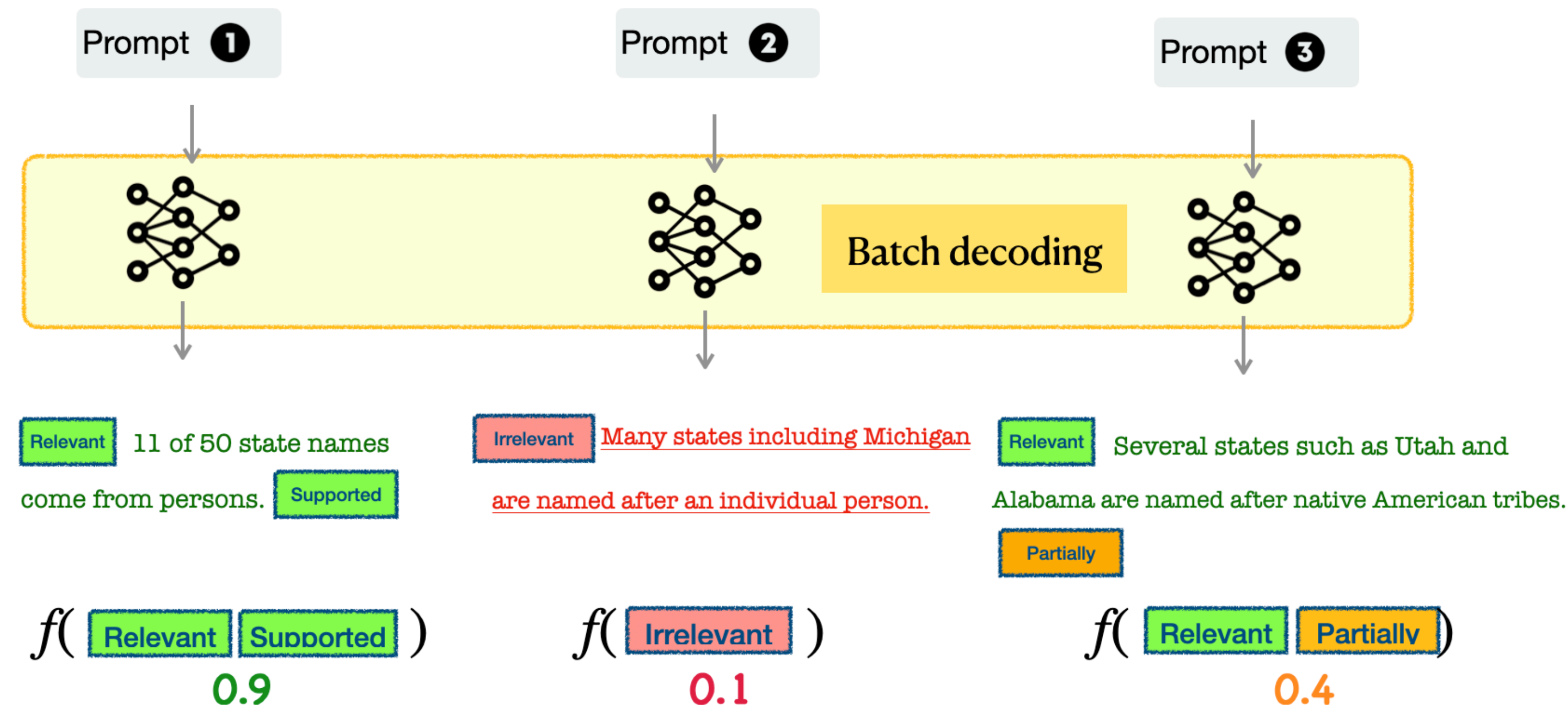




Advanced RAG inference algorithm

Advanced RAG inference algorithm to better incorporate retrieved context

Leverage model-generated tokens to improve search process at inference time

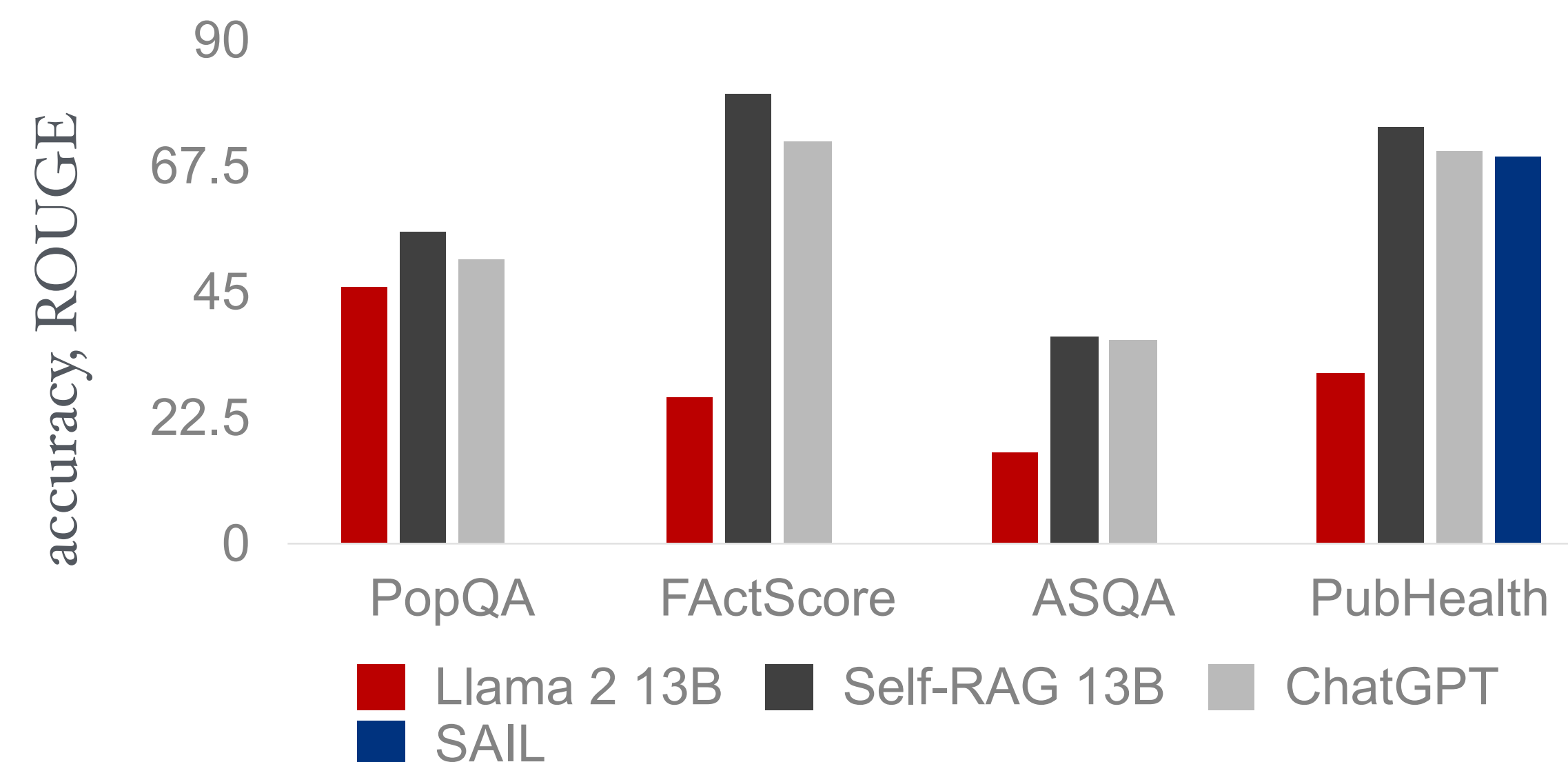




Advanced RAG inference algorithm

Advanced RAG inference algorithm to better incorporate retrieved context

- Training with 8B and 13B models significantly boosts performance compared to off-the-shelf RAG pipelines

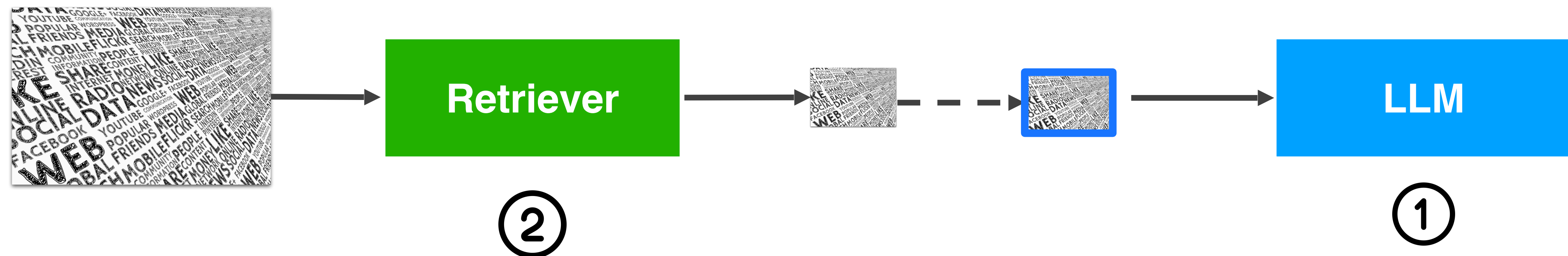




Designing and training more reliable LLM RAG

Approaches to optimize (1) LM, (2) Retrievers, or (3) prompts for LLM RAG

1. Optimizing LLMs for RAG: training / controlling LLMs with retrieved context
2. Optimizing Retriever for RAG: training retrievers for LLM RAG

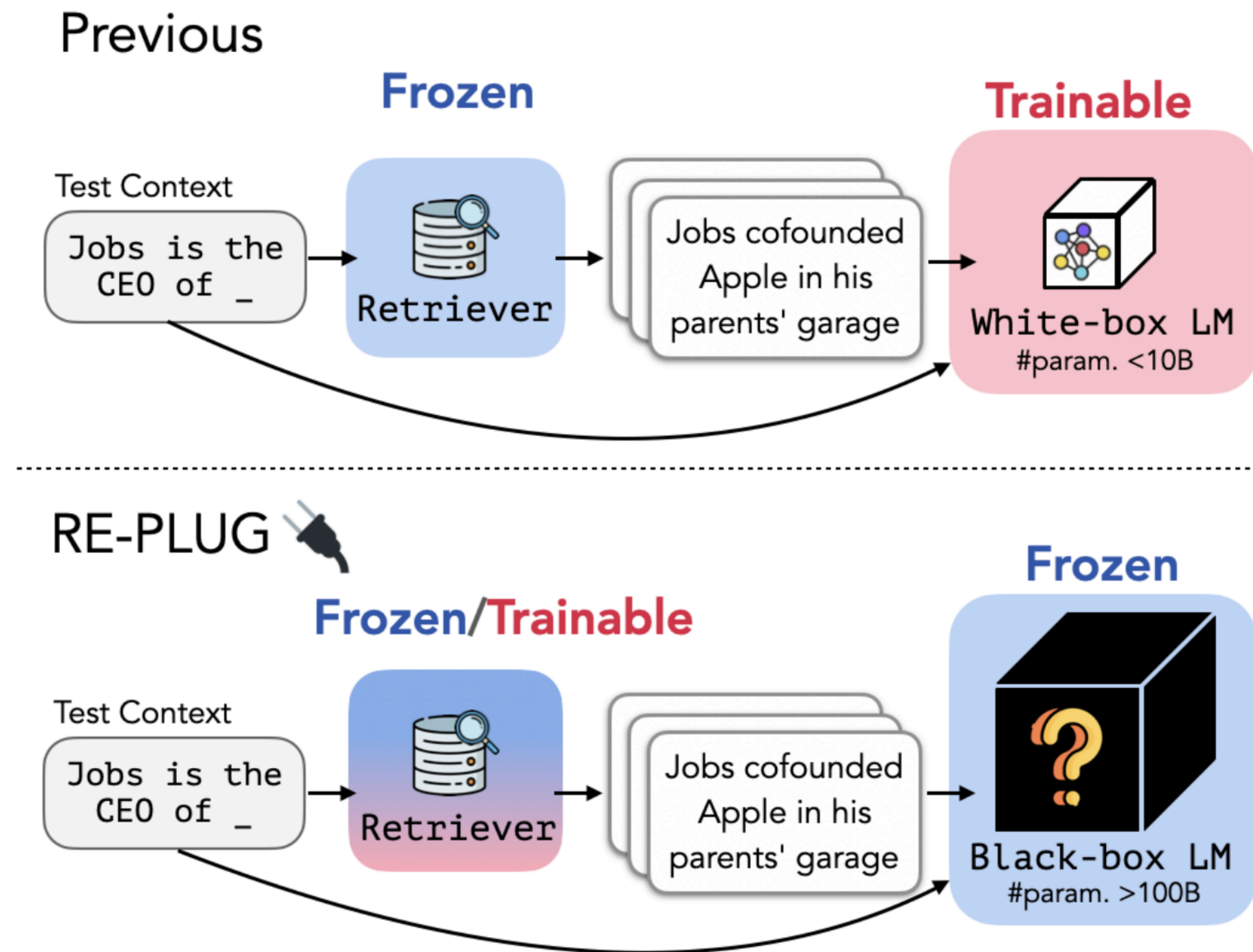




Optimizing retrievers for RAG

Training retrieval modules using LM feedback

- For RAG pipelines using blackbox LLMs e.g., GPT o1, we cannot directly train the LLMs for RAG
- Can we train retrievers instead?

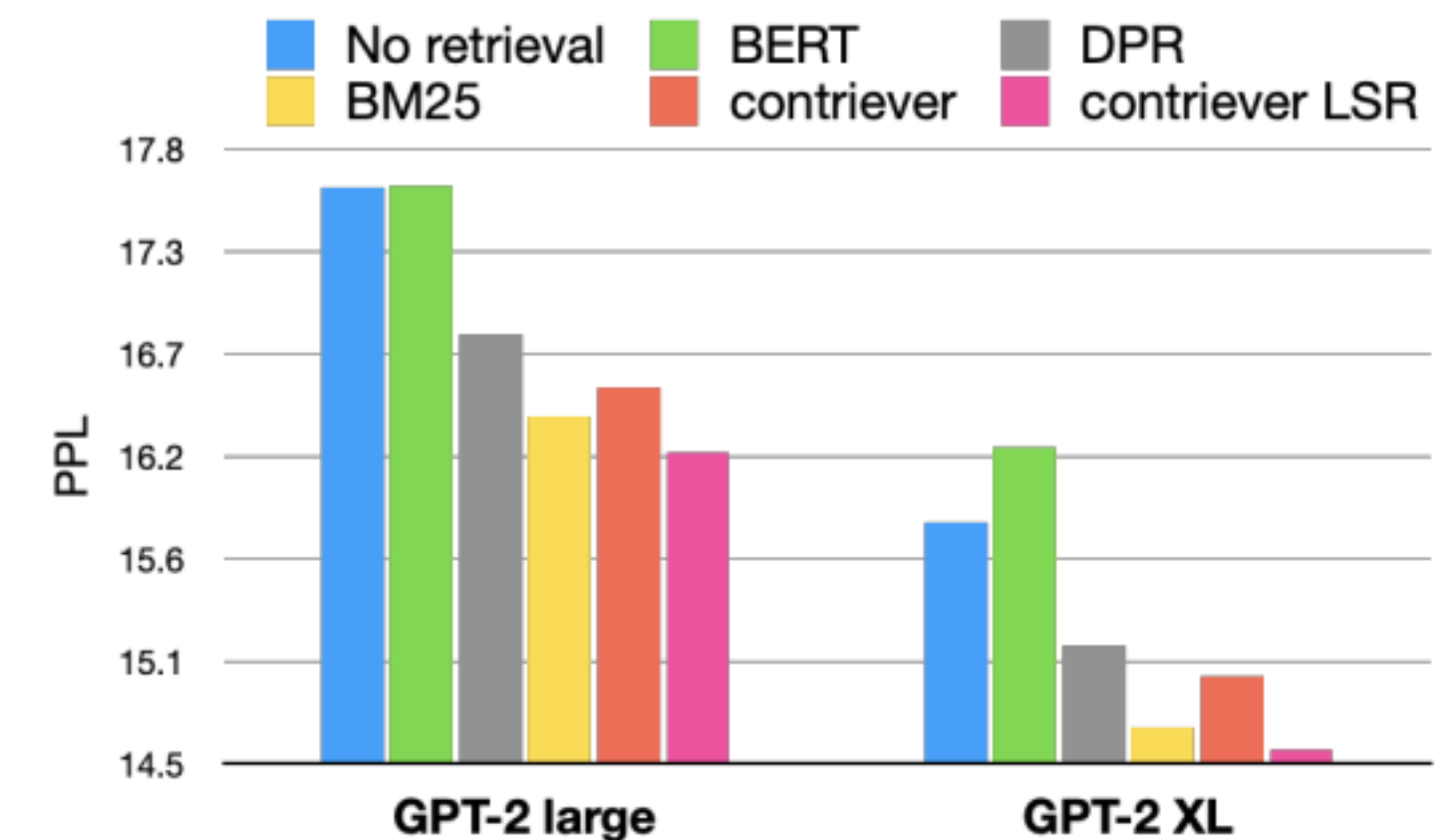
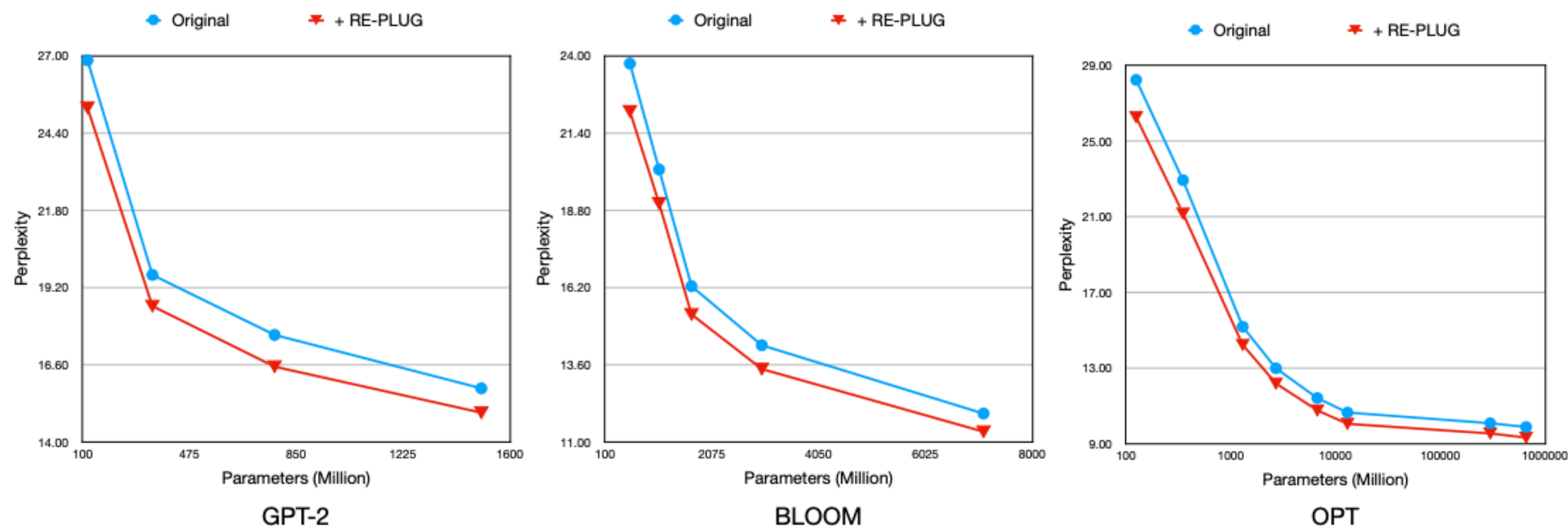




REPLUG: Training a retriever using blackbox LLM feedback

Training retrieval model using LM feedback

- models of varying sizes consistently benefit from REPLUG
- LM-supervised retriever (Contriever LSR) outperforms other off-the-shelf retrievers.





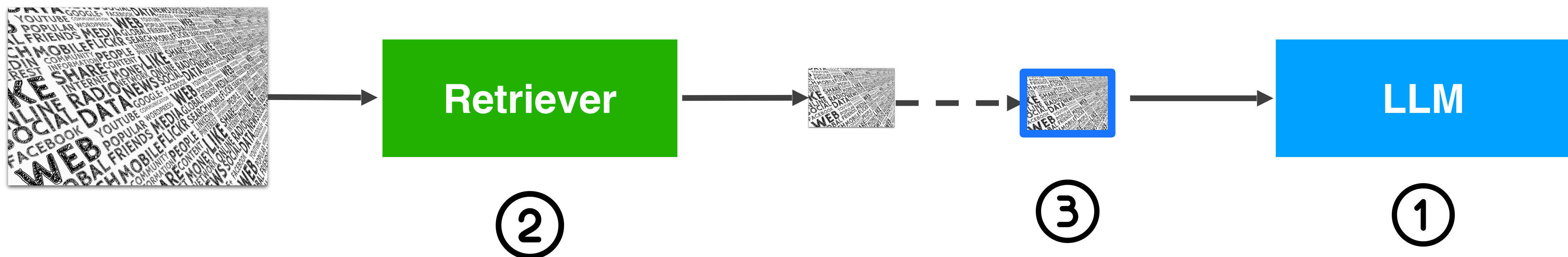
Designing and training more reliable LLM RAG

Approaches to optimize (1) LM, (2) Retrievers, or (3) prompts for LLM RAG

1. Optimizing LLMs for RAG: training / controlling LLMs with retrieved context
2. Optimizing Retriever for RAG: training retrievers for LLM RAG
3. Optimizing Prompts for RAG: advanced prompt techniques

see January 20 lecture

using methods like DSPy can improve end to end accuracy from 33% to 55%!

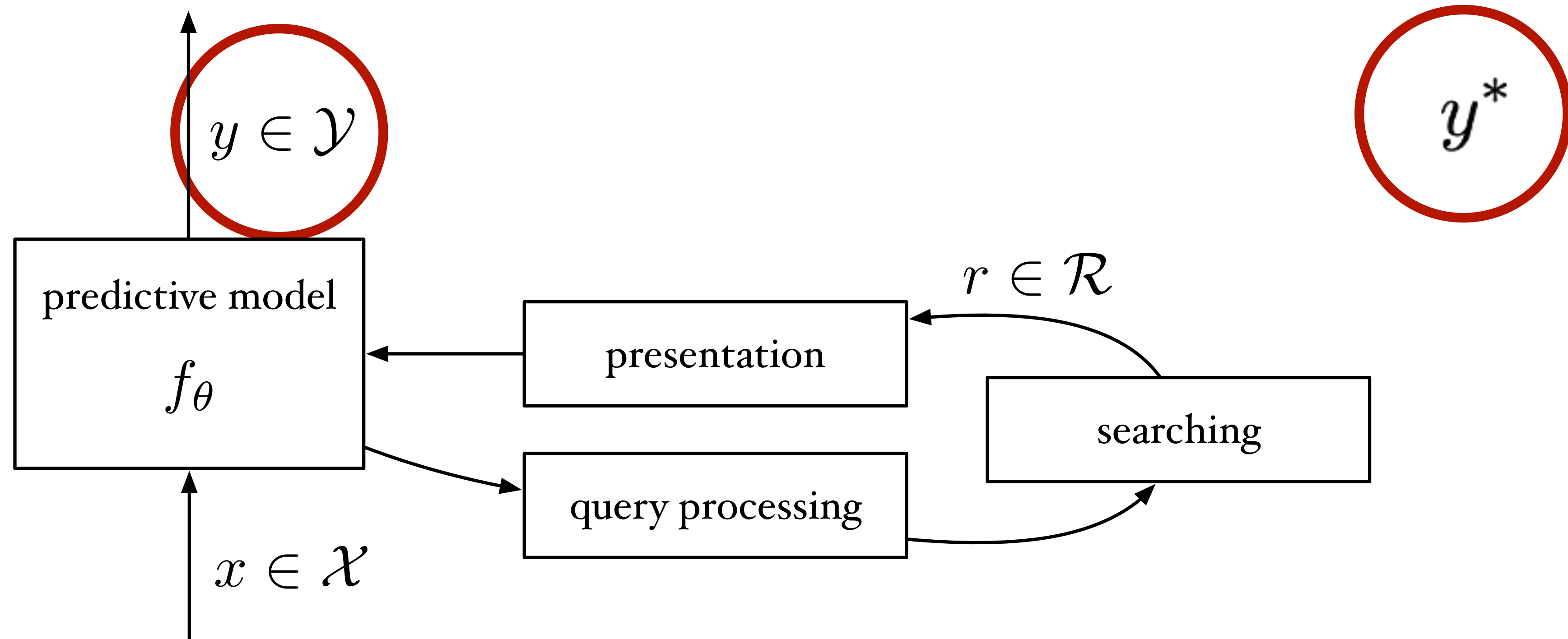


RAG evaluation

evaluation

- need to understand whether a change to the system—including a full replacement—is better than keeping the status quo
- extrinsic evaluation: final performance of the predictive model using a task-specific metric.
- intrinsic evaluation: performance of a component of the system using a local measure of quality
 - can be an efficient approximation for an extrinsic evaluation.
 - can measure some independent value such as resource consumption.

extrinsic evaluation



extrinsic evaluation

The diagram shows the formula for extrinsic evaluation with three annotations and arrows pointing to specific parts of the equation:

- predictive model output**: An arrow points from this text to the $f_\theta(x)$ term in the metric function $\mu(f_\theta(x), y)$.
- task labels**: An arrow points from this text to the y term in the metric function $\mu(f_\theta(x), y)$.
- task metric**: An arrow points from this text to the entire metric function $\mu(f_\theta(x), y)$.

$$\mathbb{E}_{\langle x, y \rangle \sim \mathcal{E}} [\mu(f_\theta(x), y)] \approx \sum_{\langle x, y \rangle \in E} \mu(f_\theta(x), y)$$

- extrinsic evaluation computes the empirical estimate of the expected value of the task metric using labeled data.
- labeled data should be sampled according the target distribution

extrinsic evaluation

- precision measures the relevant fraction of the output.
- recall measures the fraction of relevant claims in the output.
- back-translation measures the probability of an input derived from the output that are similar to the input.

$$\mu_P(\mathcal{C}_y, \mathcal{C}_{y^*}) = \frac{|\mathcal{C}_y \cap \mathcal{C}_{y^*}|}{|\mathcal{C}_y|}$$

$$\mu_R(\mathcal{C}_y, \mathcal{C}_{y^*}) = \frac{|\mathcal{C}_y \cap \mathcal{C}_{y^*}|}{|\mathcal{C}_{y^*}|}$$

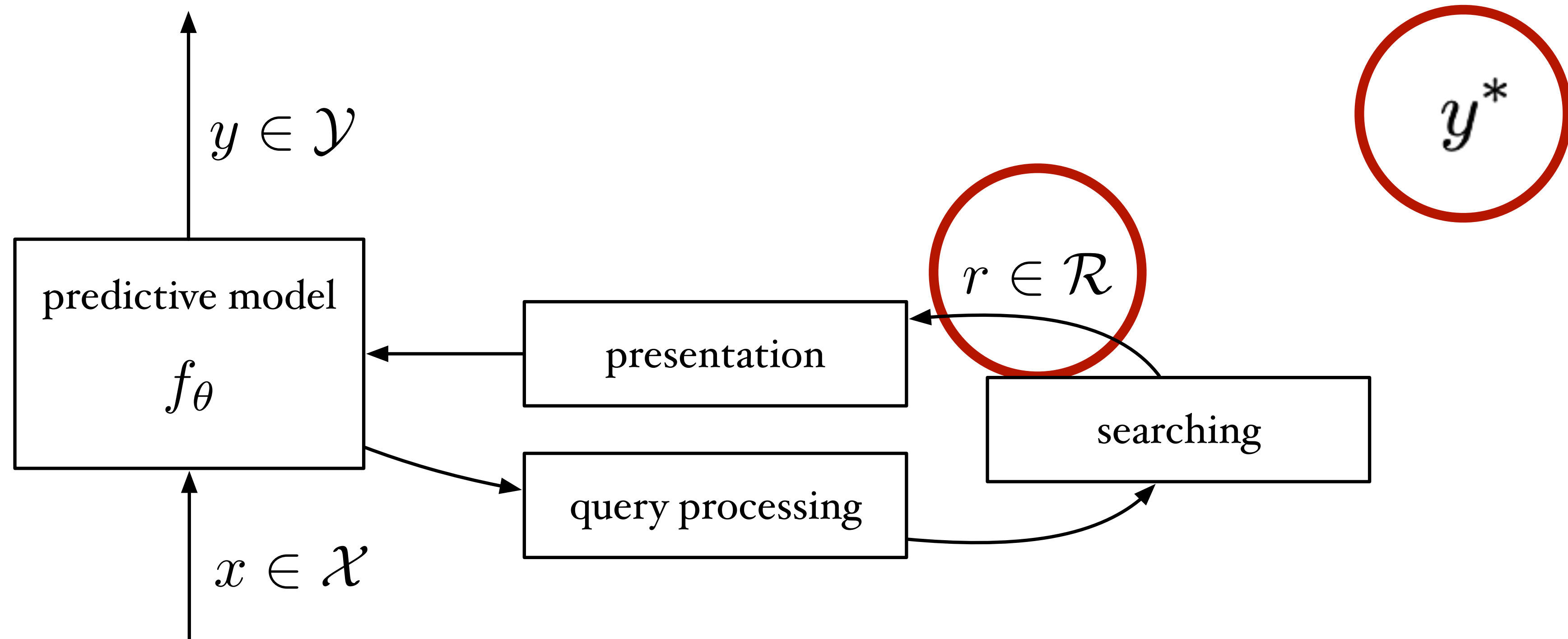
$$\mu_B(\mathcal{X}_y, x) = \frac{|\mathcal{X}_y \cap \{x\}|}{|\mathcal{X}_y|}$$

$\mathcal{C}_y$ claims in prediction y

$\mathcal{C}_{y^*}$ claims in target y^*

$\mathcal{X}_y$ input derived from
prediction y

intrinsic evaluation: retrieval



intrinsic evaluation: retrieval

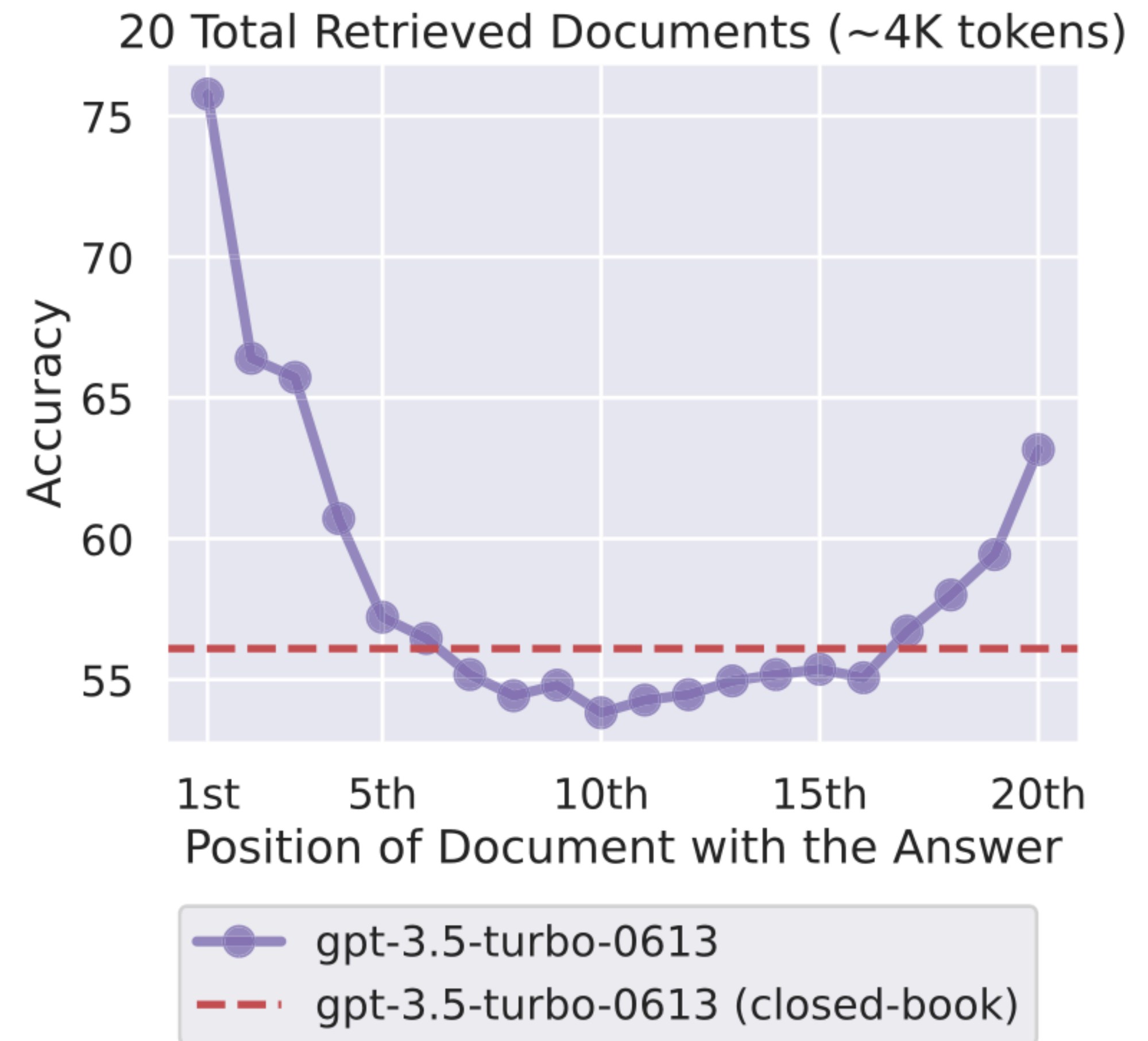
The diagram shows the formula $\mathbb{E}_{\langle x, y \rangle \sim \mathcal{E}} [\mu(f_\theta(x), y)] \propto \sum_{\langle x, \tilde{y} \rangle \in \tilde{E}} \tilde{\mu}(g_\omega(x), \tilde{y})$. Three arrows point to parts of the formula: 'retrieval output' points to $g_\omega(x)$, 'relevance labels' points to $\tilde{y}$, and 'ranking metric' points to $\tilde{\mu}$.

$$\mathbb{E}_{\langle x, y \rangle \sim \mathcal{E}} [\mu(f_\theta(x), y)] \propto \sum_{\langle x, \tilde{y} \rangle \in \tilde{E}} \tilde{\mu}(g_\omega(x), \tilde{y})$$

- classic retrieval metrics support human searchers and correlation with human task performance.
- can reuse existing metrics and new relevance judgments to measure component performance
 - relevance judgements should be task-specific

intrinsic evaluation: retrieval

- traditional retrieval metrics assume that position of relevant item is monotonically related to task performance
- REML models may not obey this!
- top and bottom of the ranking influence task performance!



intrinsic evaluation: retrieval

- for claim-based evaluation, we can inspect the claims in the retrieval.

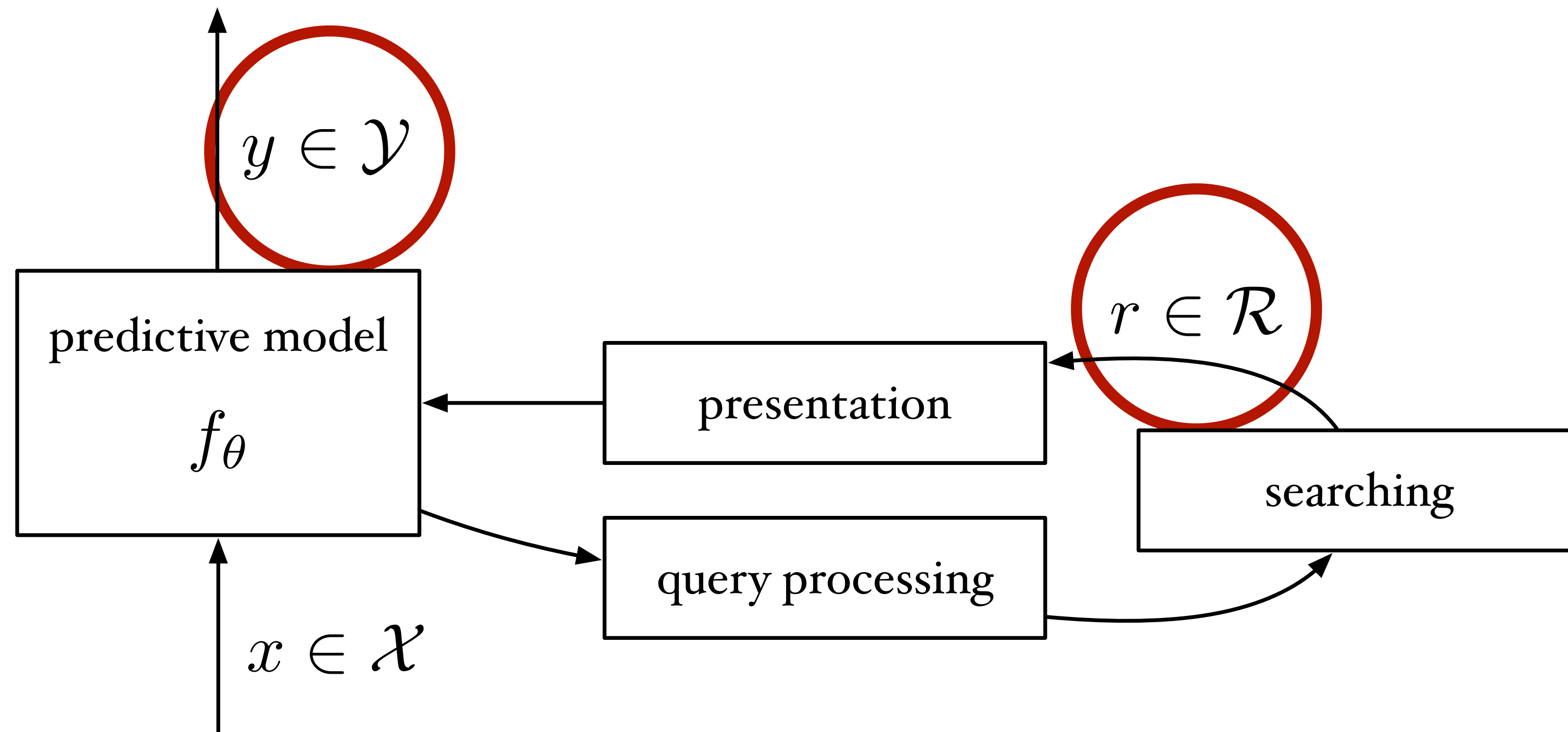
$$\mu_P(\mathcal{C}_r, \mathcal{C}_{y^*}) = \frac{|\mathcal{C}_r \cap \mathcal{C}_{y^*}|}{|\mathcal{C}_r|}$$

$$\mu_R(\mathcal{C}_r, \mathcal{C}_{y^*}) = \frac{|\mathcal{C}_r \cap \mathcal{C}_{y^*}|}{|\mathcal{C}_{y^*}|}$$

$\mathcal{C}_r$ claims in retrieval r

$\mathcal{C}_{y^*}$ claims in target y^*

intrinsic evaluation: interaction



- **faithfulness** measures the degree to which claims in output are supported by the retrieval.
- low faithfulness suggests that claims in the the output are not supported by the retrieval
- high faithfulness suggests that claims in the the output are supported by the retrieval

$$\mu_F(\mathcal{C}_y, \mathcal{C}_r) = \frac{|\mathcal{C}_y \cap \mathcal{C}_r|}{|\mathcal{C}_y|}$$

$\mathcal{C}_y$ claims in prediction y
 $\mathcal{C}_r$ claims in retrieval r

- **utilization** measures the degree to which *relevant* claims in retrieval are present in the output.
- low utilization suggests that claims in the the retrieval are not present in the output
- high utilization suggests that claims in the the retrieval are present in the output

$$\mu_U(\mathcal{C}_y, \mathcal{C}_r, \mathcal{C}_{y^*}) = \frac{|\mathcal{C}_y^* \cap \mathcal{C}_r^*|}{|\mathcal{C}_r^*|}$$

$$\mathcal{C}_y^* = \mathcal{C}_y \cap \mathcal{C}_{y^*}$$

$$\mathcal{C}_r^* = \mathcal{C}_r \cap \mathcal{C}_{y^*}$$

- **sensitivity** measures the degree to which *nonrelevant* claims in output are present in the retrieval.
- low sensitivity suggests that nonrelevant claims in the the output might come from the retrieval.
- high sensitivity suggests that nonrelevant claims in the the output might not come from the retrieval.

$$\mu_S(\mathcal{C}_y, \mathcal{C}_r, \mathcal{C}_{y^*}) = \frac{|\mathcal{C}_y^- \cap \mathcal{C}_r^-|}{|\mathcal{C}_y|}$$

$$\mathcal{C}_y^- = \mathcal{C}_y \setminus \mathcal{C}_{y^*}$$

$$\mathcal{C}_r^- = \mathcal{C}_r \setminus \mathcal{C}_{y^*}$$

- **hallucination** measures the degree to which *nonrelevant* claims in output are not present in the retrieval.
- low hallucination suggests that nonrelevant claims in the the output might come from the retrieval.
- high hallucination suggests that nonrelevant claims in the the output might not come from the retrieval.

$$\mu_H(\mathcal{C}_y, \mathcal{C}_r, \mathcal{C}_{y^*}) = \frac{|\mathcal{C}_y^- \setminus \mathcal{C}_r^-|}{|\mathcal{C}_y|}$$

$$\mathcal{C}_y^- = \mathcal{C}_y \setminus \mathcal{C}_{y^*}$$

$$\mathcal{C}_r^- = \mathcal{C}_r \setminus \mathcal{C}_{y^*}$$

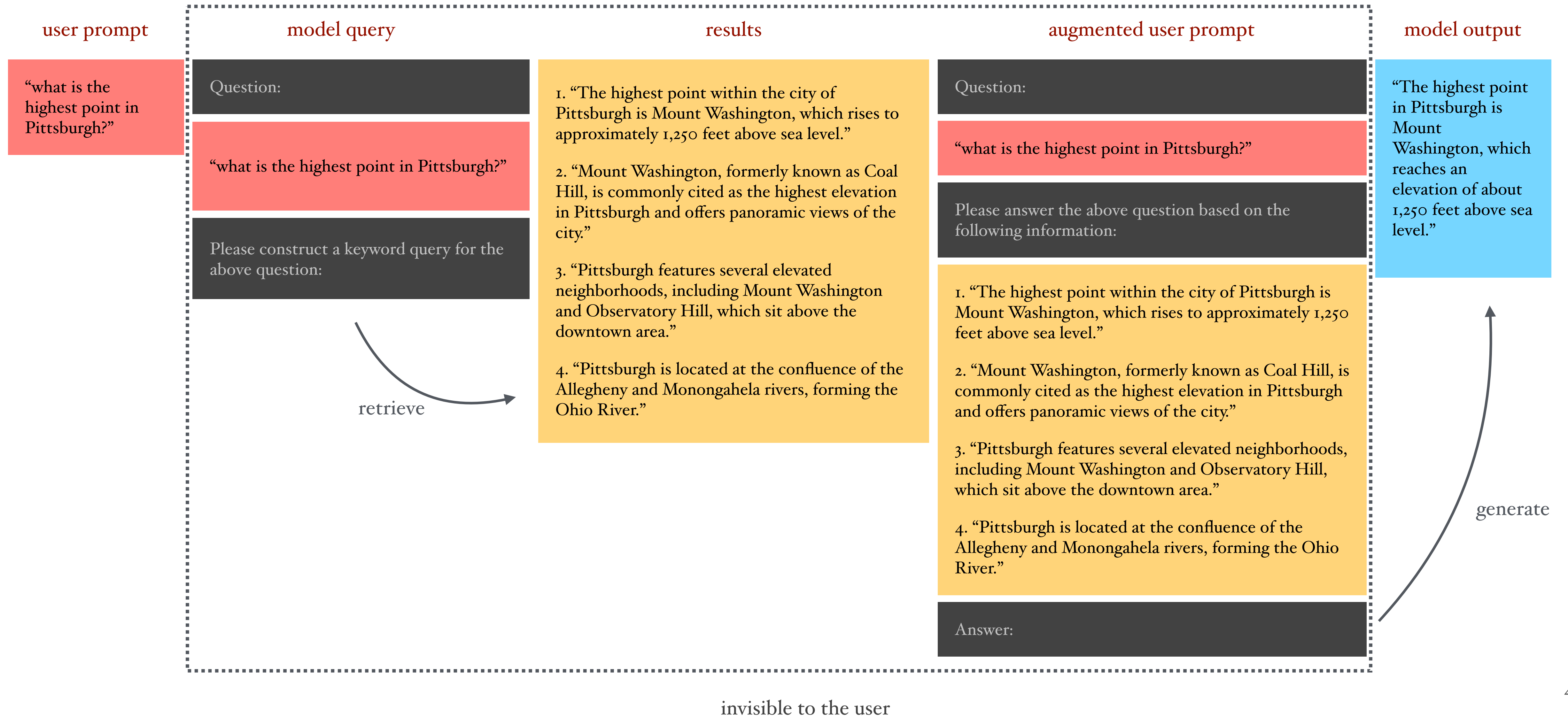
- **knowledge** measures the degree to which *relevant* claims in output are not present in the retrieval.
- low knowledge suggests that relevant claims in the the output might come from the retrieval.
- high knowledge suggests that relevant claims in the the output might not come from the retrieval.

$$\mu_K(\mathcal{C}_y, \mathcal{C}_r, \mathcal{C}_{y^*}) = \frac{|\mathcal{C}_y^* \setminus \mathcal{C}_r|}{|\mathcal{C}_y|}$$

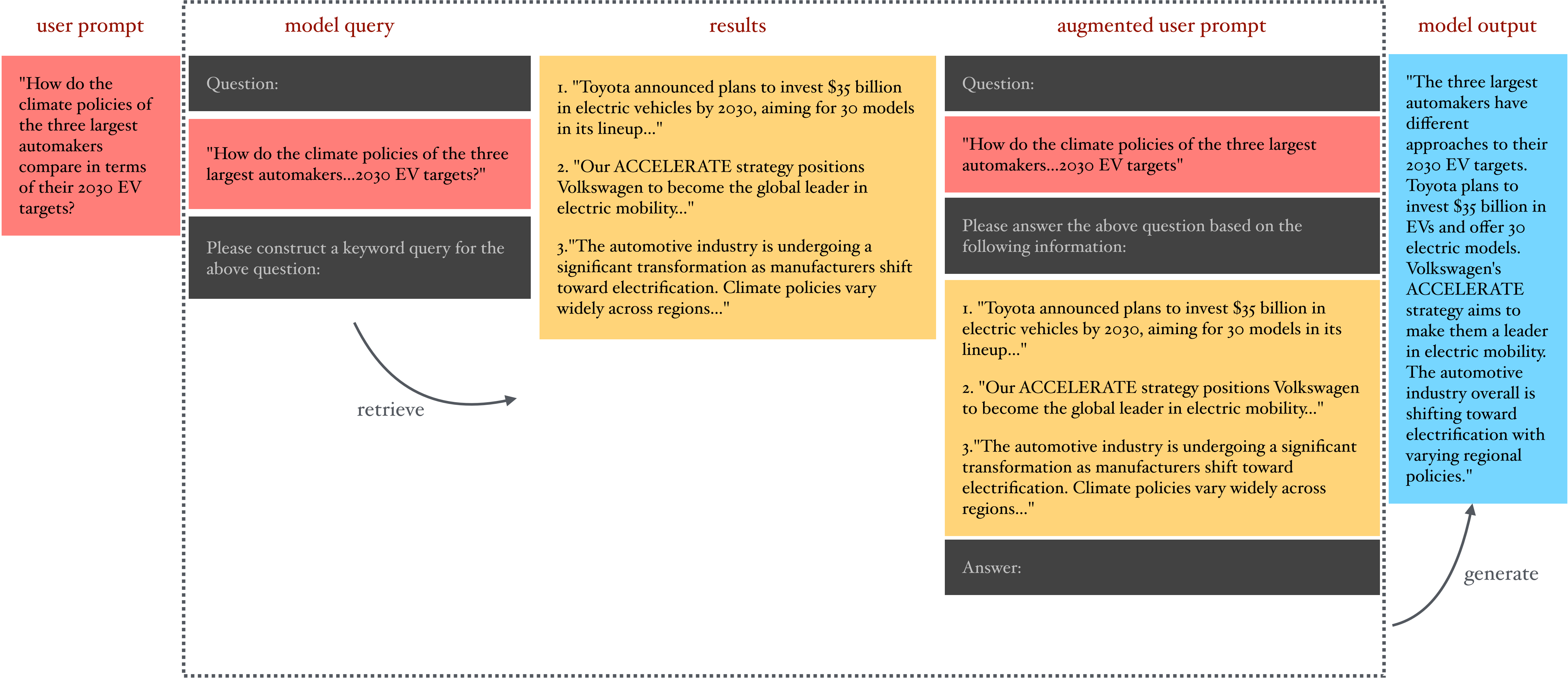
$$\mathcal{C}_y^* = \mathcal{C}_y \cap \mathcal{C}_{y^*}$$

Deep research

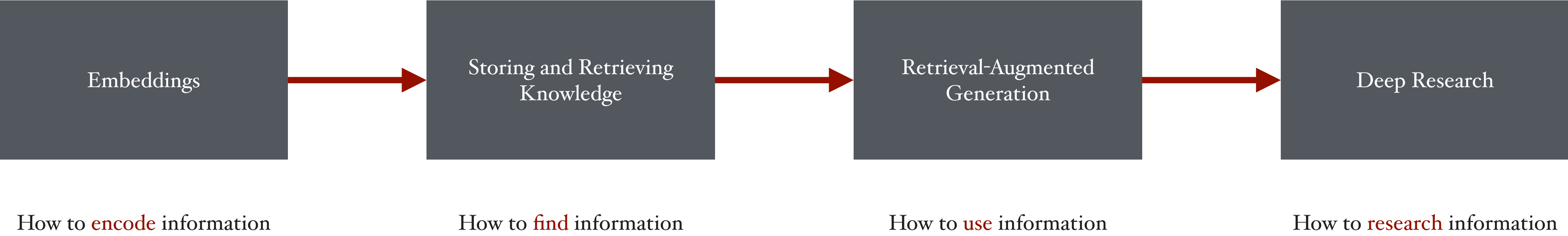
RAG tends to be good when the desired model output is a function of one (or a few) retrievals



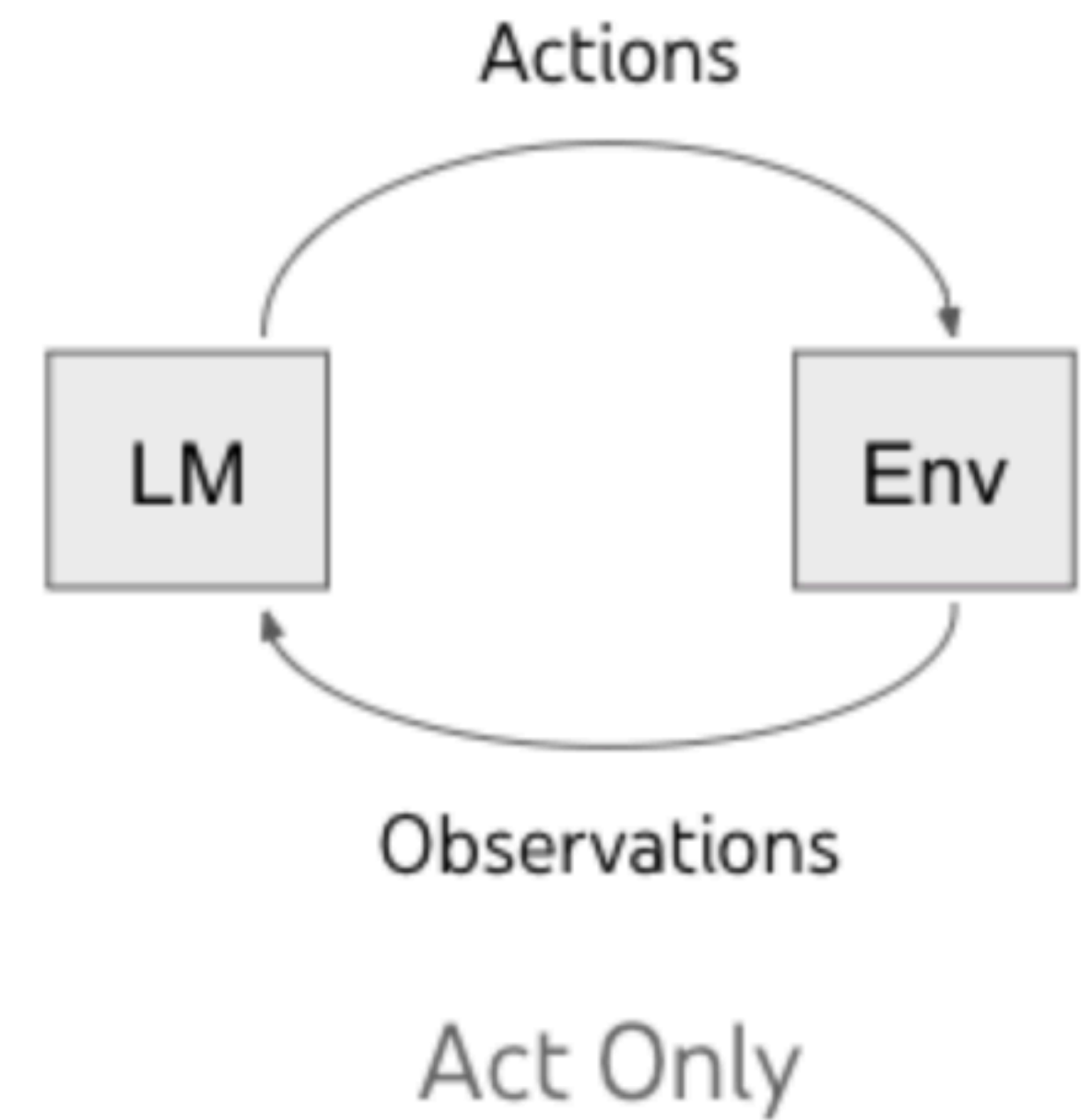
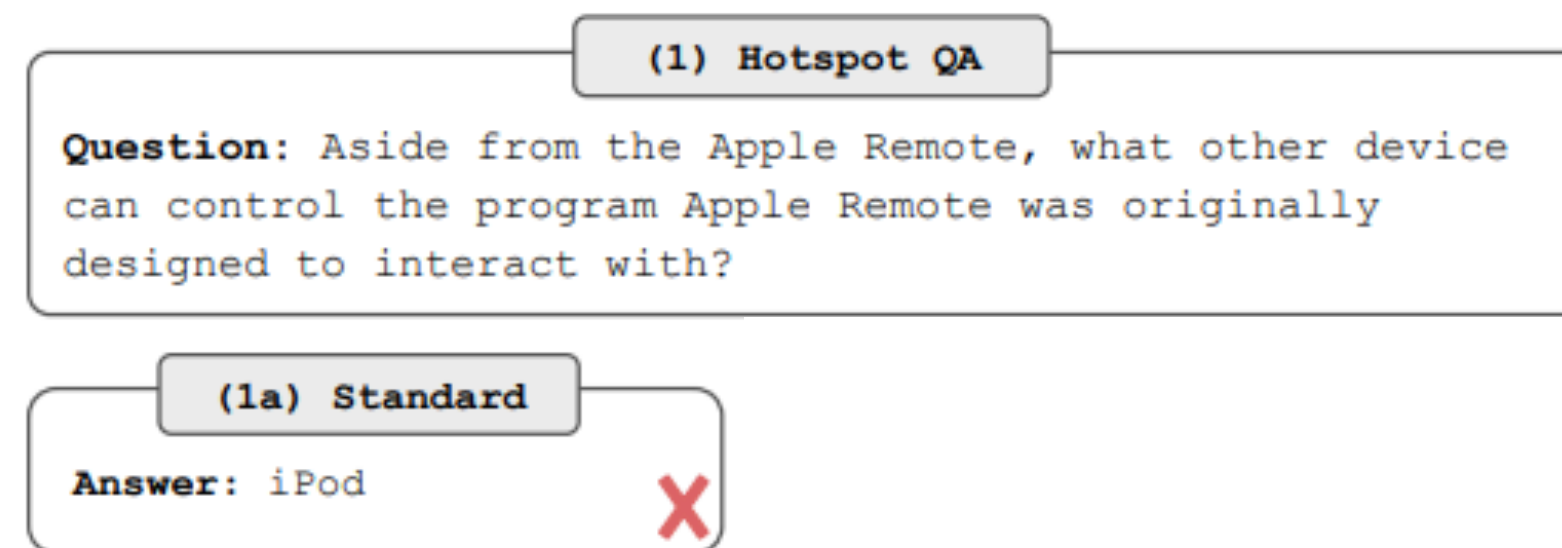
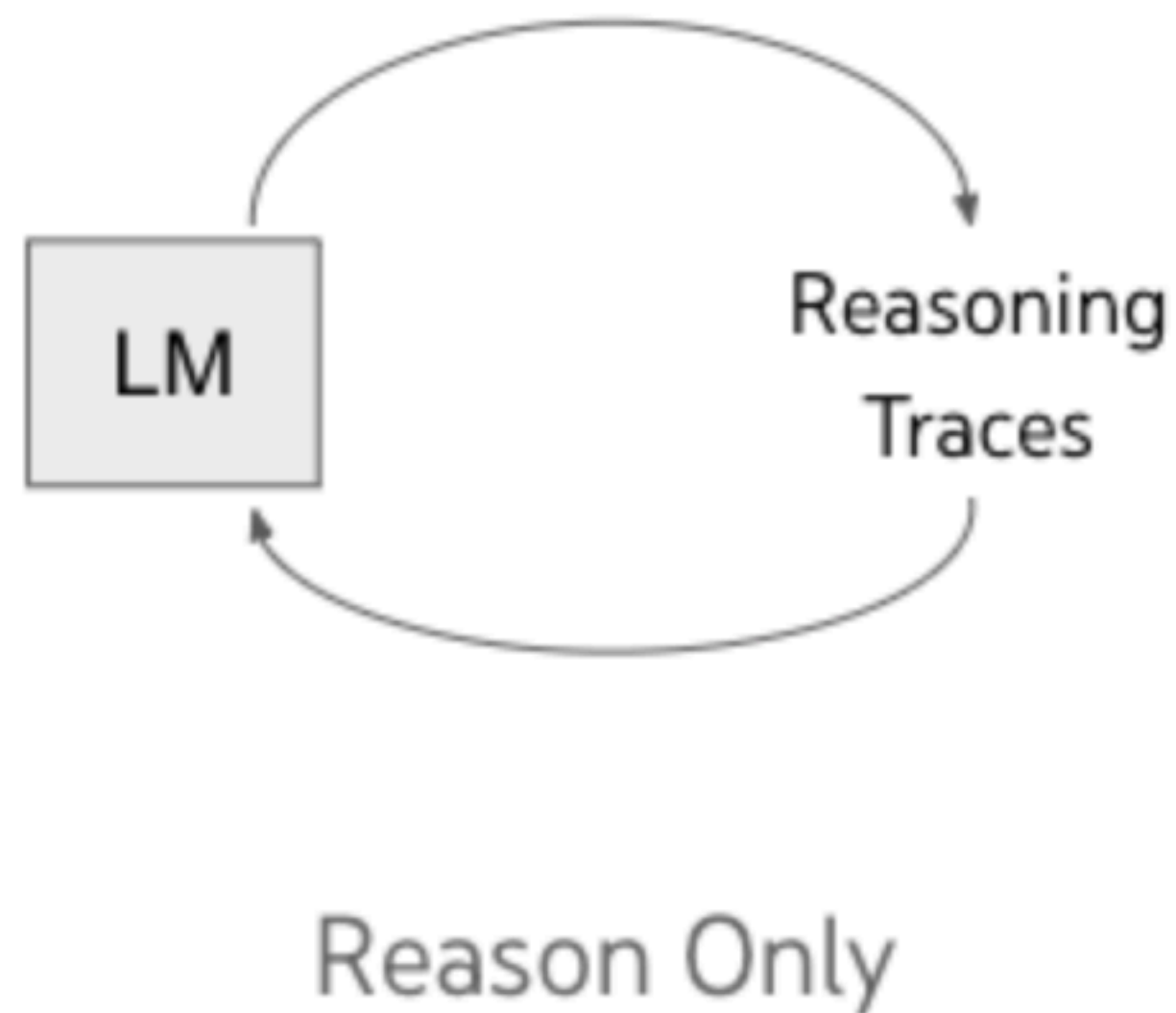
RAG tends to be good when the desired model output is a function of one (or a few) retrievals



poll: RAG failures...



QA: reasoning versus acting



poll: acting vs reasoning...

ReAct

(1) Hotspot QA

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

(1a) Standard

Answer: iPod

X

(1b) CoT (Reason Only)

Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch

X

(1c) Act-Only

Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control ...

Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ...

Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software ...

Act 4: Finish[yes]

X

(1d) ReAct (Reason + Act)

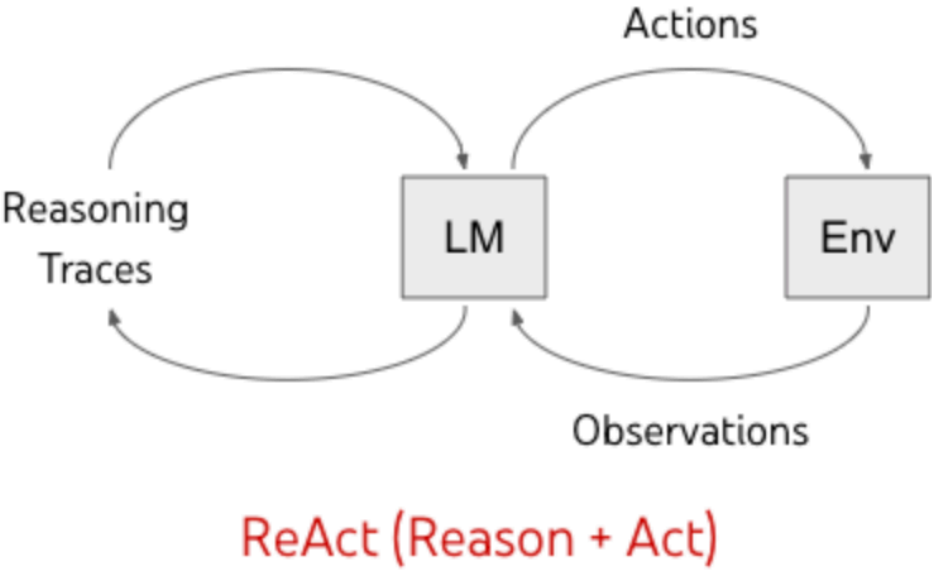
Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.
Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the Front Row media center program ...

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.
Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

Thought 3: Front Row is not found. I need to search Front Row (software) .
Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software ...

Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.
Act 4: Finish[keyboard function keys]

✓



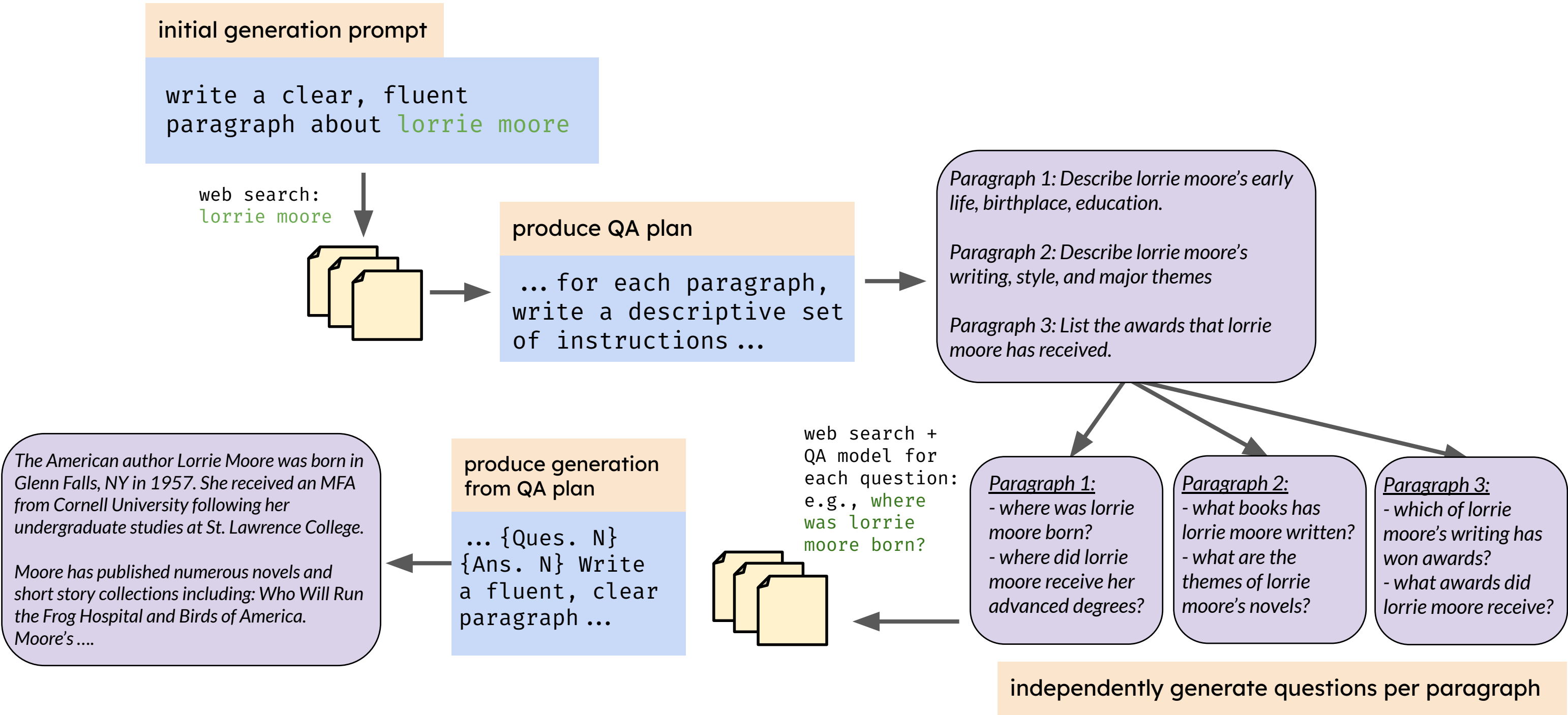
ReAct

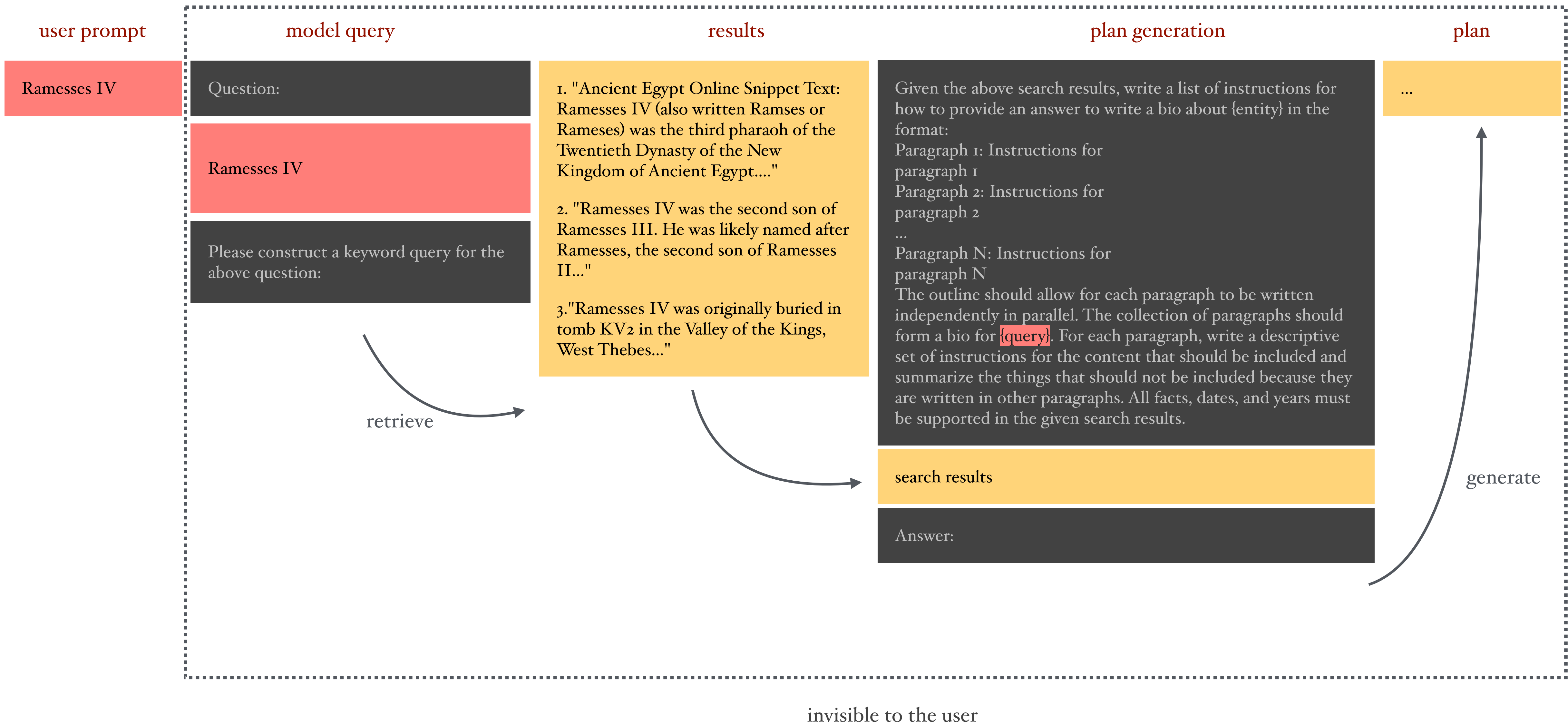
	HotpotQA (exact match, 6-shot)	FEVER (accuracy, 3-shot)
Standard	28.7	57.1
Reason-only (CoT)	29.4	56.3
Act-only	25.7	58.9
ReAct	27.4	60.9
Best ReAct + CoT Method	35.1	64.6
Supervised SoTA	67.5 (using ~140k samples)	89.5 (using ~90k samples)

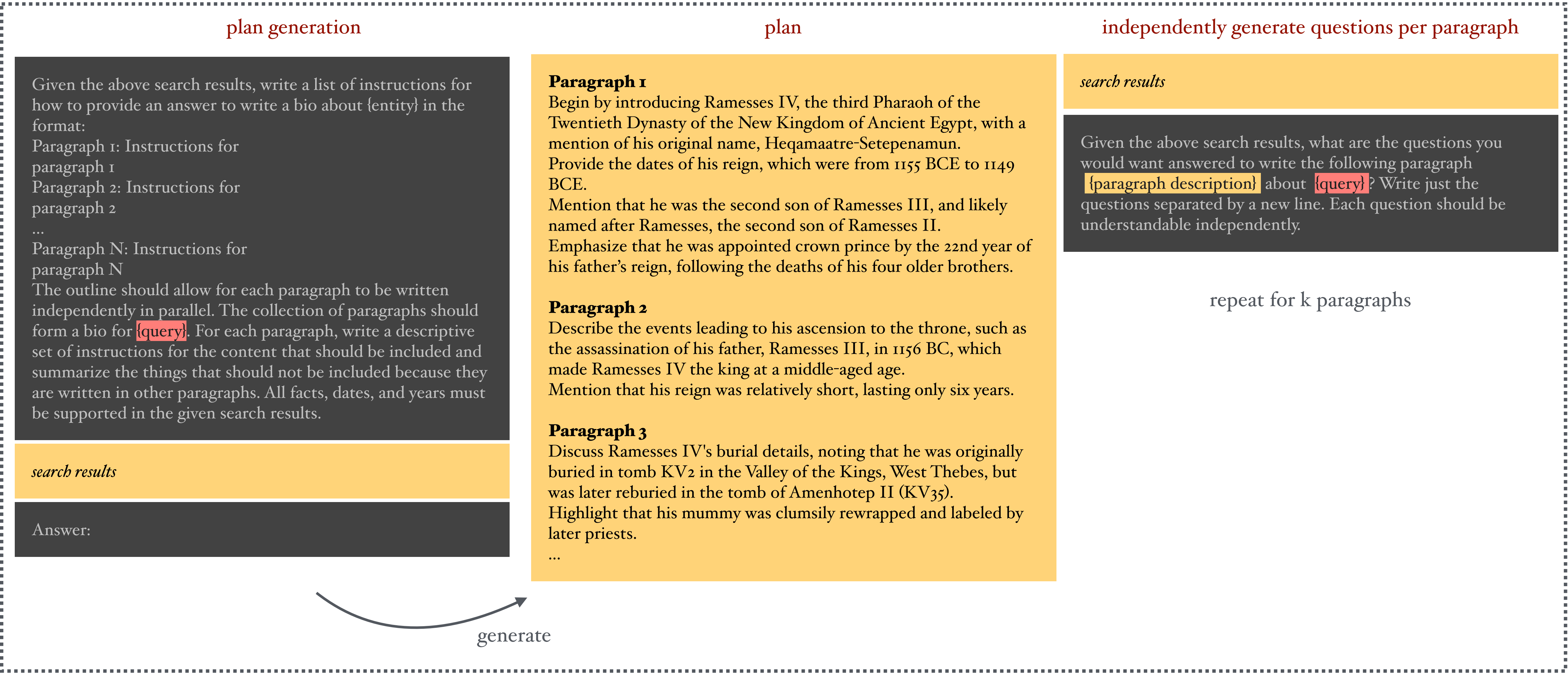
Planning

- need to plan search process because agent information needs may be distributed across a corpus
 - situations where there is no "optimal query"
- subqueries: a single question can often be decomposed into multiple subqueries
 - when there are no dependencies, you can query in parallel
 - some subqueries will be dependent on each other (e.g., multi-hop reasoning)
- requires generating a plan for searching

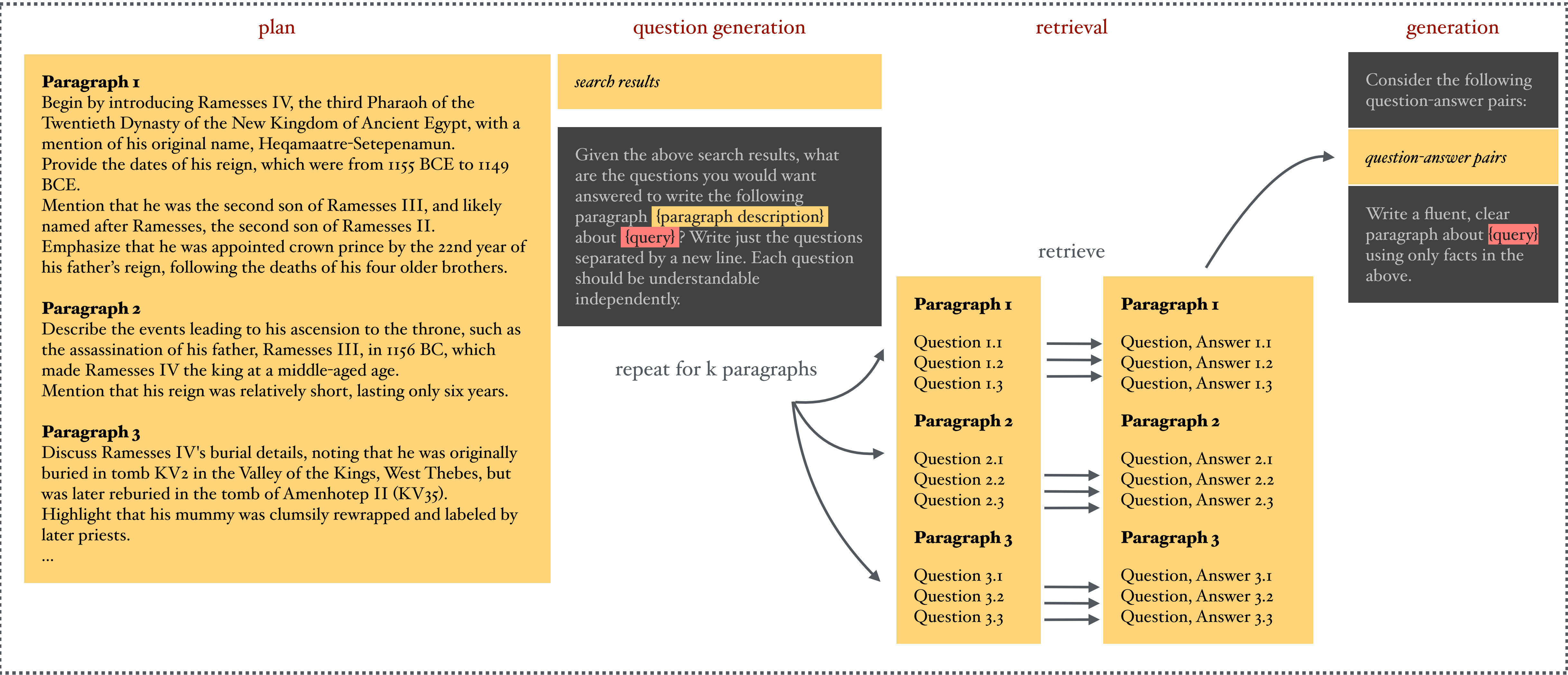
Planning



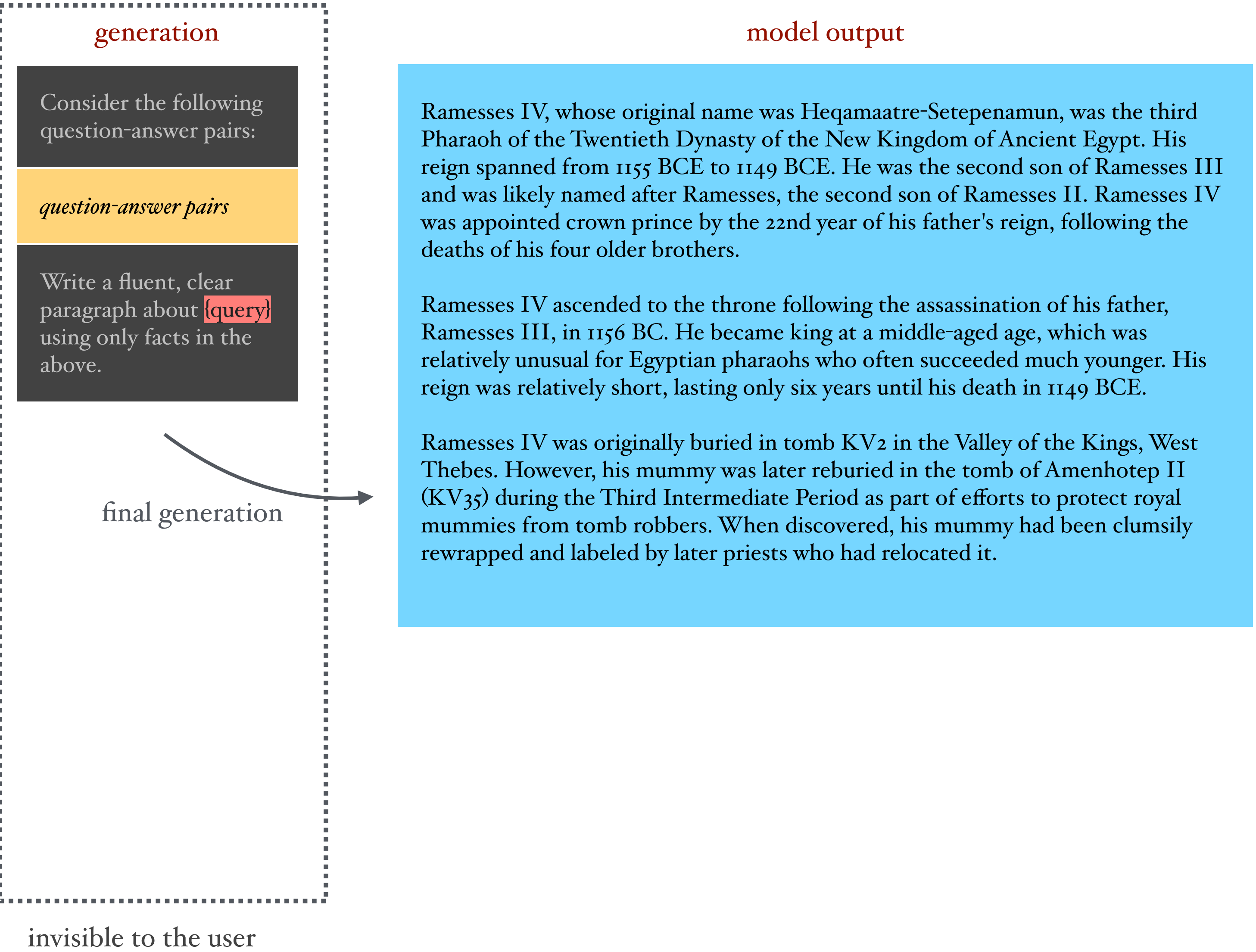




invisible to the user



invisible to the user



Planning

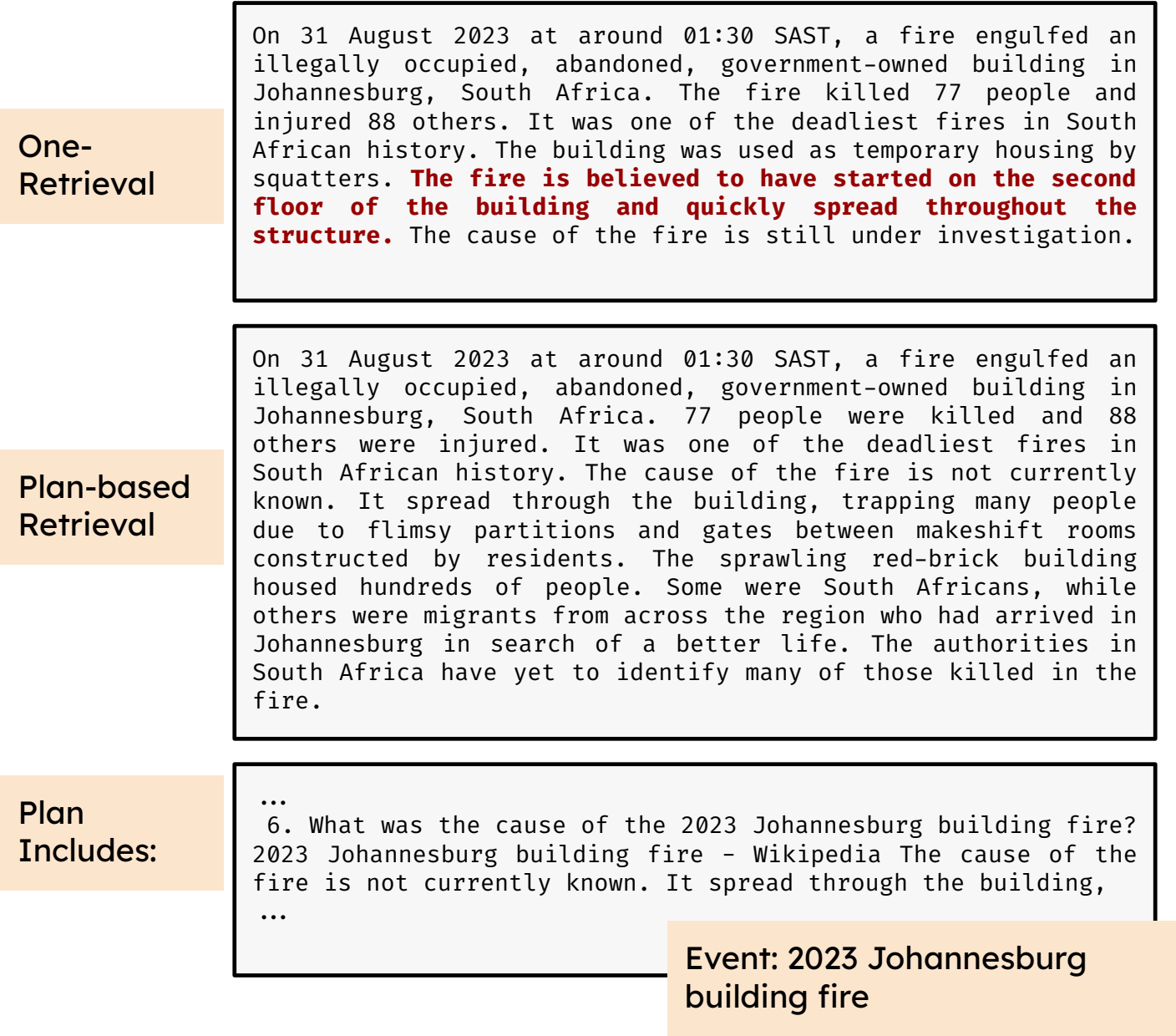


Figure 2: **Example Generation.** One of the hallucinations in the One-Retrieval model is the focus of one of the questions provided in the question-based plan.

	Approach	AIS			Rouge Prec.			Length # Tokens
		Strict	Macro	Micro	R1	R2	RL	
Wiki-Event	No Retrieval	0.00	15.30	15.74	69.45	30.16	66.63	88.63
	One-Retrieval	74.54	90.46	90.35	96.23	84.94	95.09	79.26
	One-Retrieval (2x snippets)	75.39	90.99	91.10	97.65	87.01	97.04	78.63
	Plan-based Retrieval (Var.A)	89.99	96.18	95.14	99.59	94.52	99.44	82.76
	Plan-based Retrieval (Var.B)	72.96	92.02	91.85	99.43	90.30	99.19	113.92
Wiki-Ent	No Retrieval	0.00	13.80	16.27	55.67	16.95	53.14	97.29
	One-Retrieval	60.91	83.09	83.11	91.52	74.17	90.12	87.62
	One-Retrieval (2x snippets)	67.03	87.44	86.67	95.79	80.45	94.92	86.86
	Plan-based Retrieval (Var.A)	84.77	95.16	94.23	99.59	92.99	99.44	93.03
	Plan-based Retrieval (Var.B)	63.37	90.40	90.70	99.28	89.45	99.04	134.47
Researcher	No Retrieval	0.00	6.74	6.73	57.10	17.85	54.84	80.58
	One-Retrieval	62.89	83.68	84.70	91.55	73.90	90.53	74.24
	One-Retrieval (2x snippets)	69.81	86.34	87.57	95.90	80.03	95.27	75.01
	Plan-based Retrieval (Var.A)	79.56	91.63	92.16	97.87	83.99	97.48	77.71
	Plan-based Retrieval (Var.B)	64.78	88.31	88.93	97.72	83.40	97.23	98.00
News Events	No Retrieval	0.00	5.46	5.14	68.81	24.41	66.00	89.44
	One-Retrieval	64.10	84.76	87.77	95.95	80.93	94.61	83.51
	One-Retrieval (2x snippets)	61.54	86.72	87.62	97.63	80.70	96.50	79.83
	Plan-based Retrieval (Var.A)	80.77	94.47	93.72	99.76	93.47	99.60	92.15
	Plan-based Retrieval (Var.B)	66.67	90.88	91.49	99.61	92.53	99.46	124.67

Table 1: **Comparison of Generation Approaches using text-bison-001 model.** The plan-based retrieval models that are at the center of our analysis yield more attributable text than One- and No- Retrieval methods. Of the two variants, Var.A produces more attributable sentences and Var.B produces longer texts.