

University of Amsterdam
Informatics Institute



(NeurIPS 2026)

LatentRAG

Latent Reasoning and Retrieval for Efficient Agentic RAG

Yijia Zheng, Marcel Worring
{y.zheng, m.worring}@uva.nl

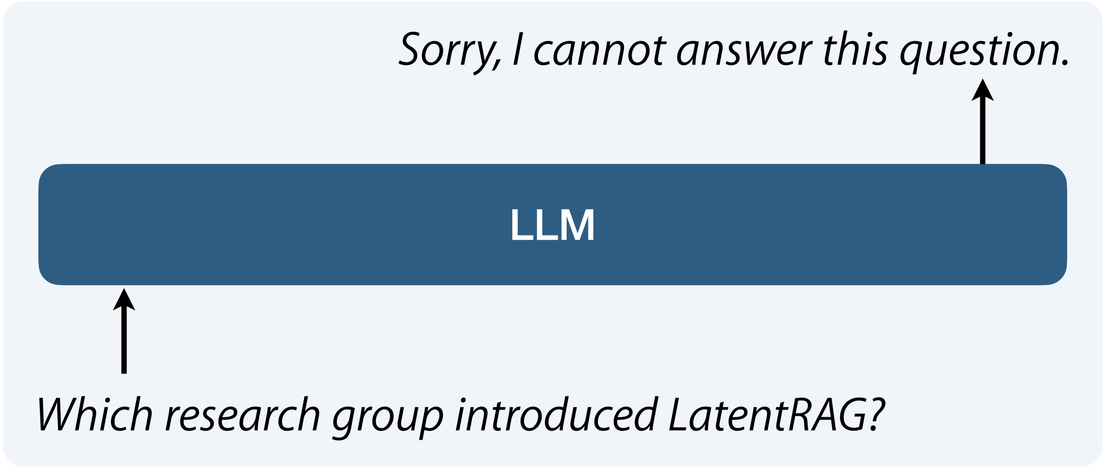
MultiX Group
Informatics Institute, University of Amsterdam

What is Retrieval-Augmented Generation (RAG)?

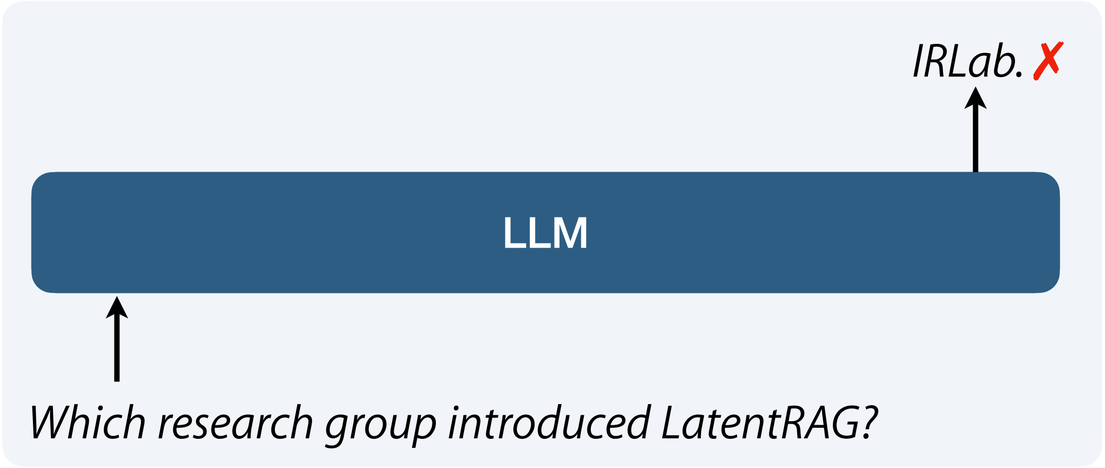
What is Retrieval-Augmented Generation (RAG)?

LLMs are powerful for text modeling and generation,
but they do not have access to **external knowledge**.

- **Lack of knowledge**

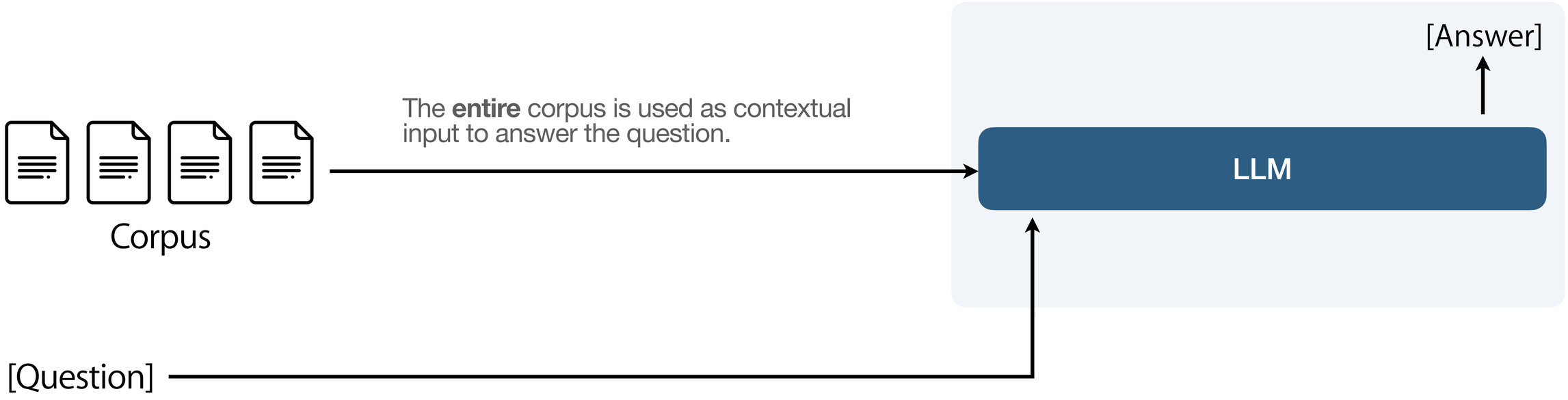


- **Hallucination**



What is Retrieval-Augmented Generation (RAG)?

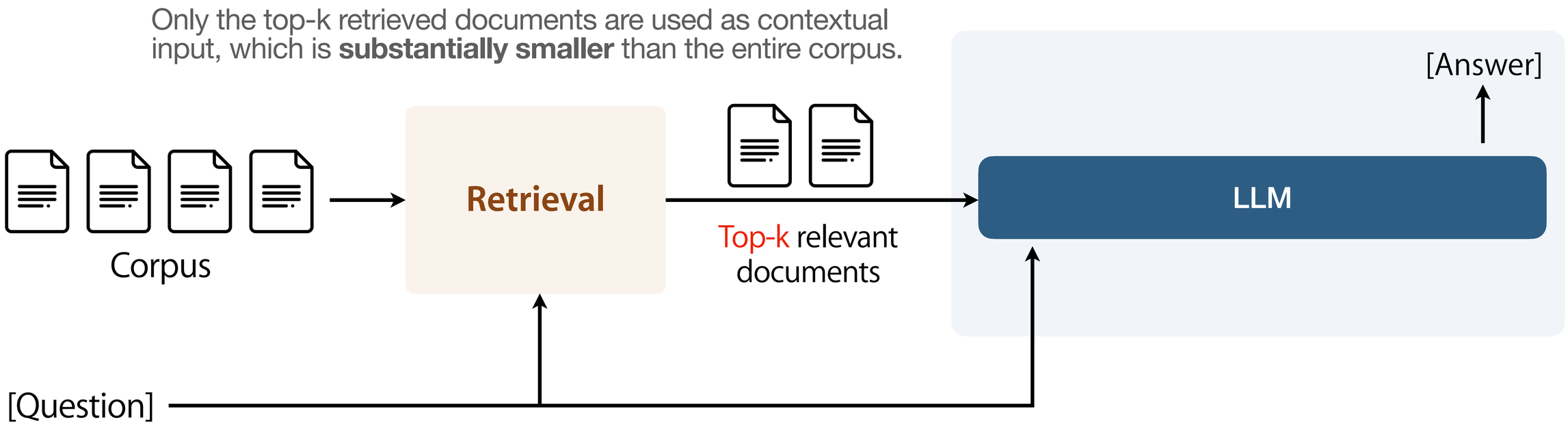
Long-Context LLMs



- High computational costs and latency.
- Limited context window size.

What is Retrieval-Augmented Generation (RAG)?

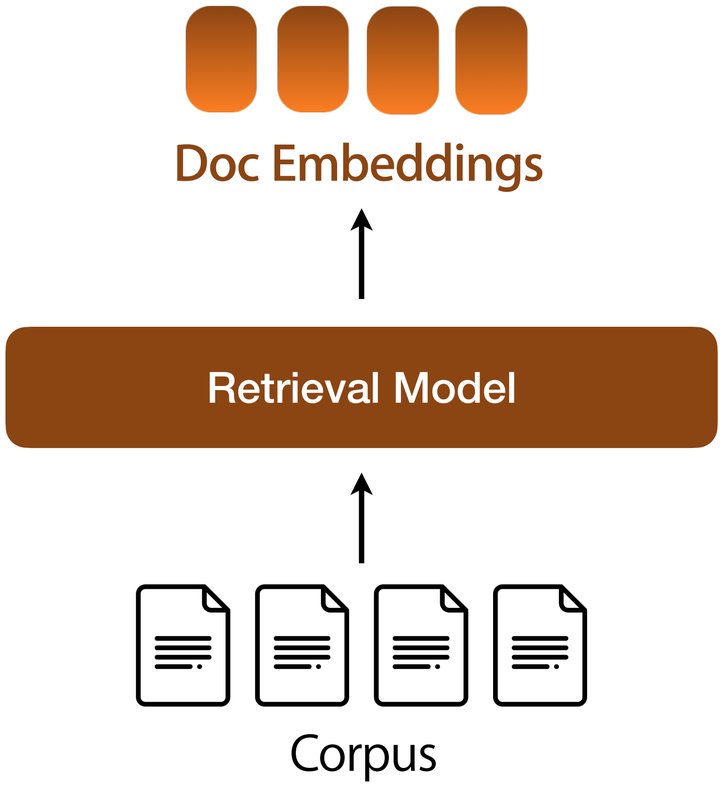
Retrieval-Augmented Generation (RAG)



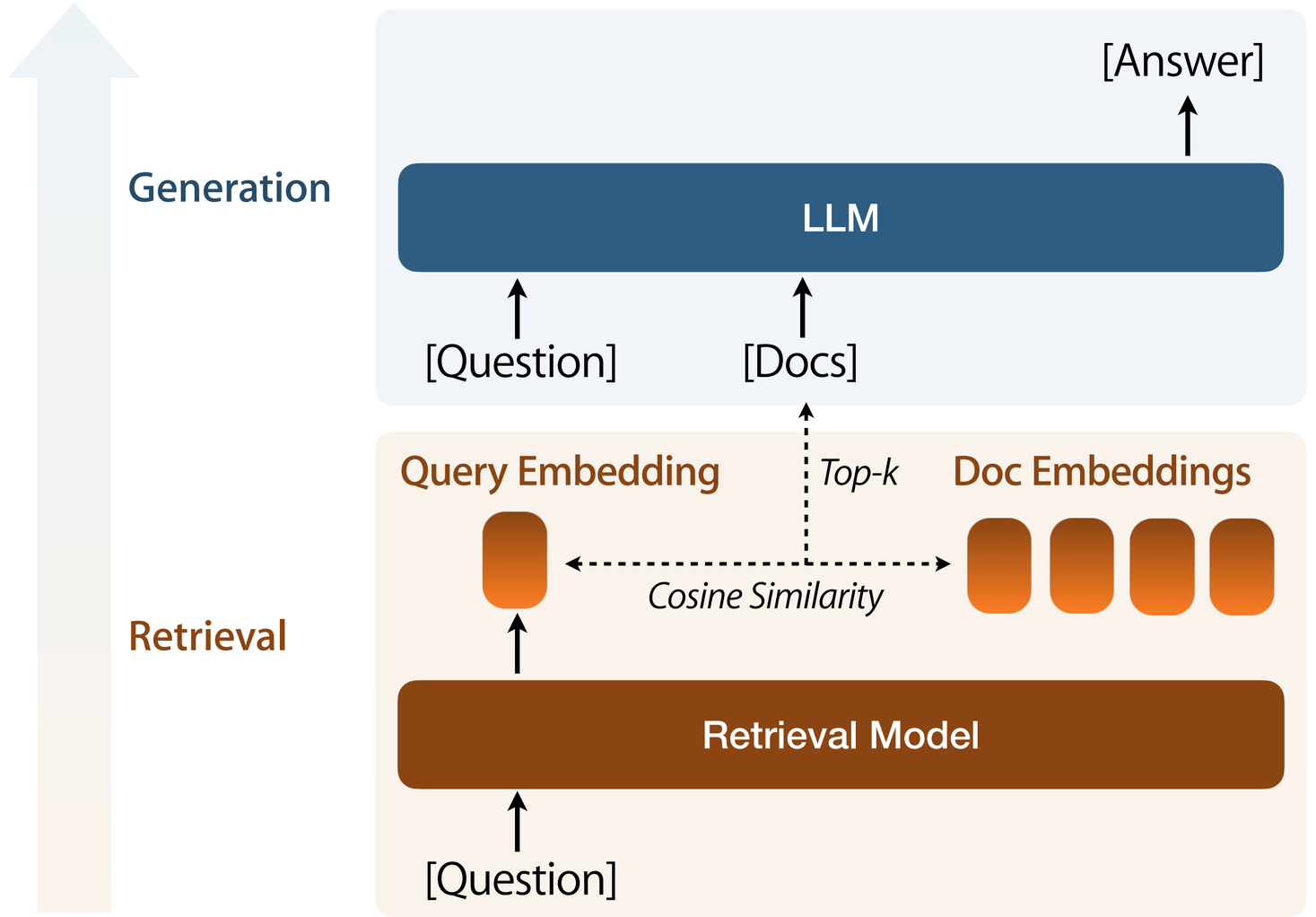
- RAG provides a reliable (tractable) and efficient solution

Traditional RAG

1. Offline indexing



2. Online retrieval and generation



Traditional RAG is not effective enough

- **The query may require multi-hop retrieval.**

Which research group introduced LatentRAG?

Direct retrieval using the original query:

Retrieved documents about LatentRAG, but none mention the research group. ✗

Which research group introduced LatentRAG?

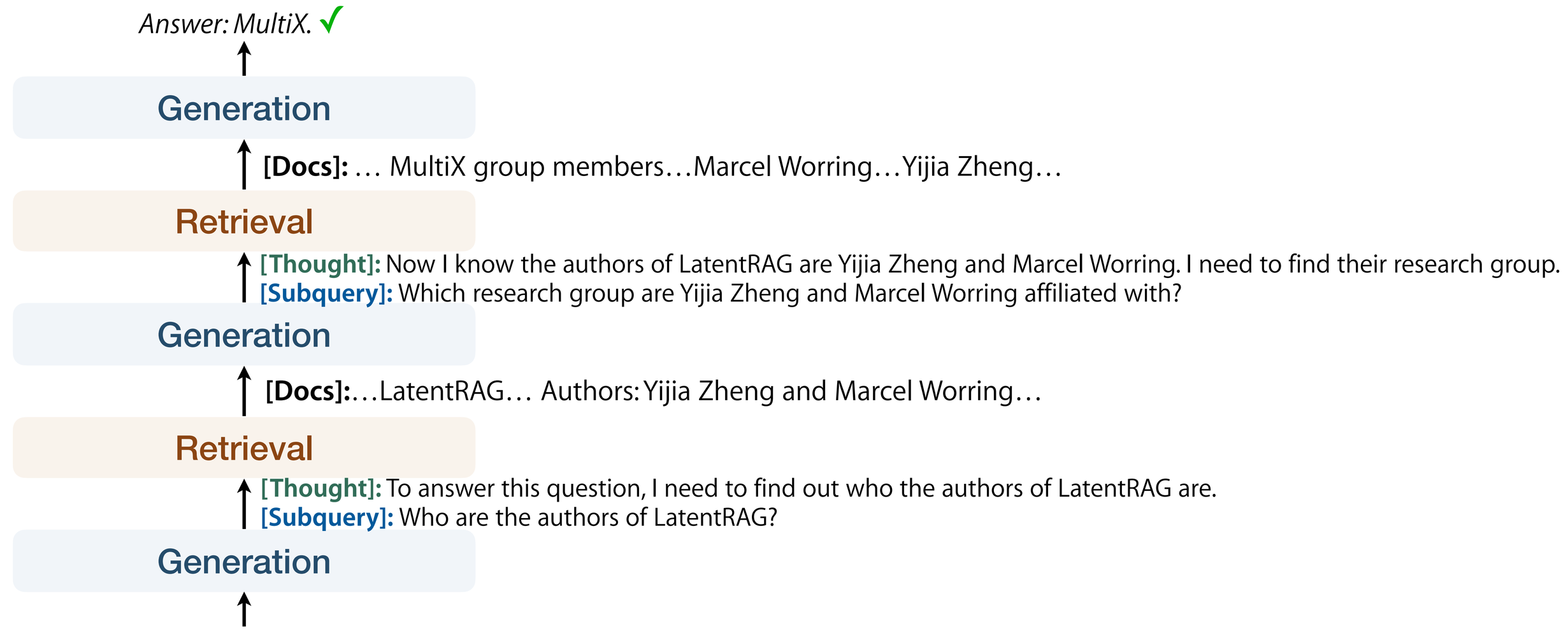
Multi-hop decomposition and retrieval are required:

1. *Who are the authors of LatentRAG?*
   (Answer: Yijia Zheng and Marcel Worring)

2. *Which research group are Yijia Zheng and Marcel Worring affiliated with?*
   (Answer: MultiX) ✓

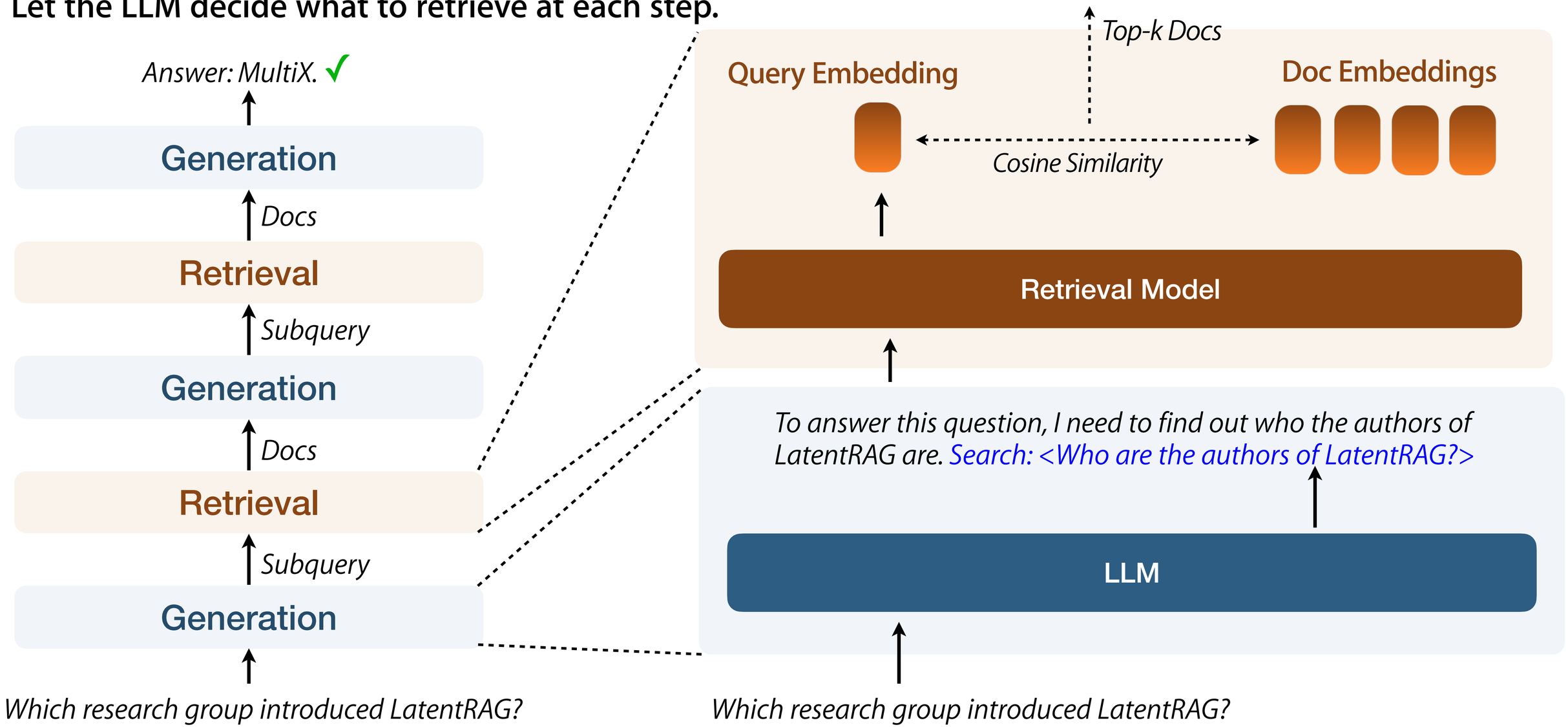
Agentic RAG

Let the LLM decide what to retrieve at each step.



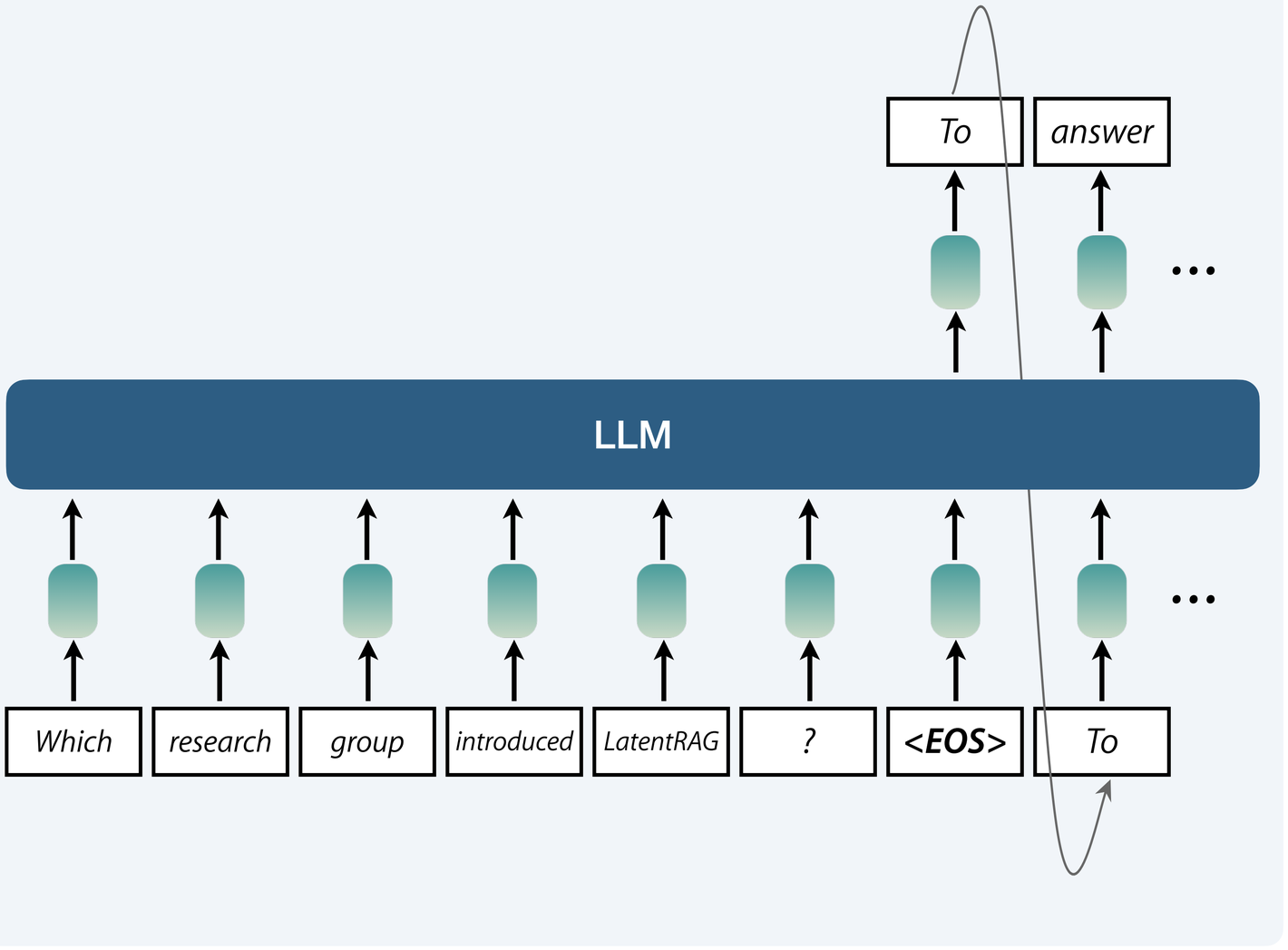
Agentic RAG

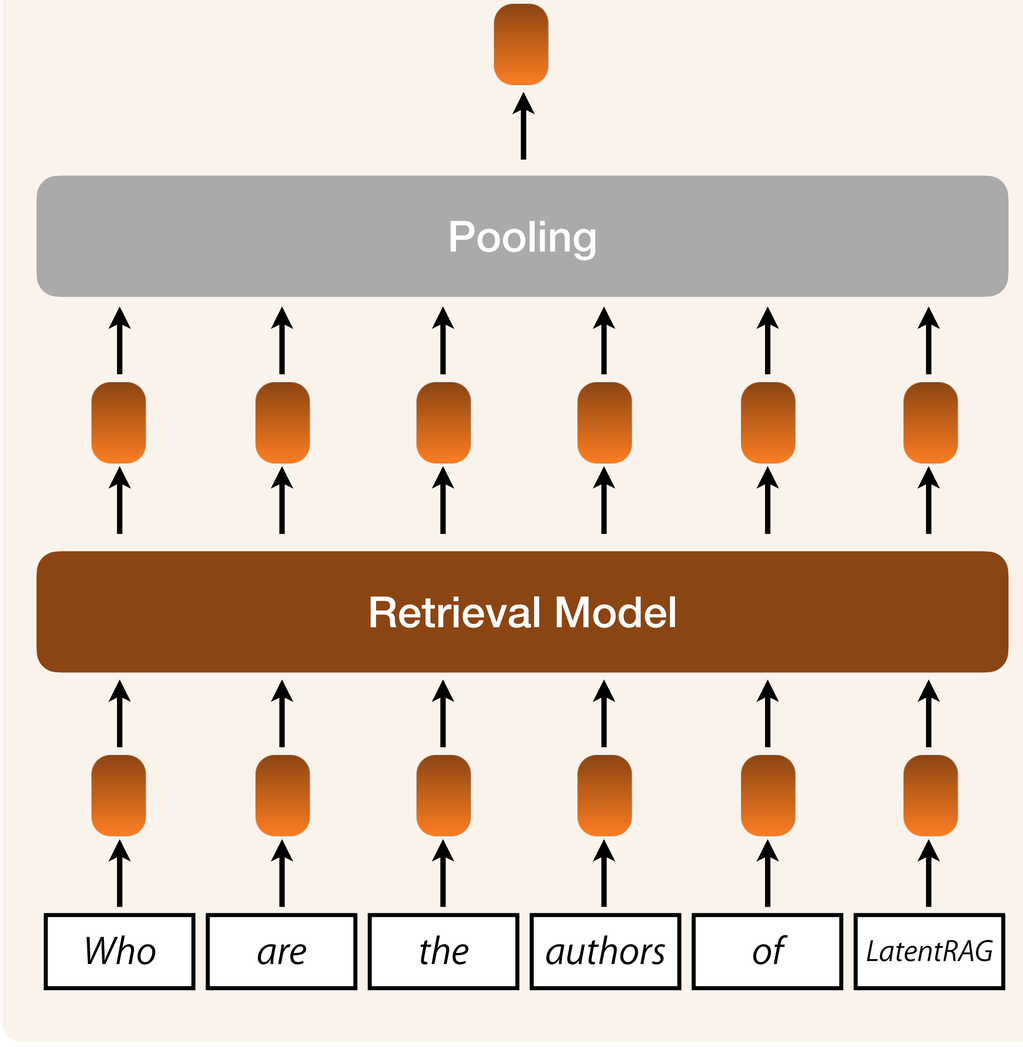
Let the LLM decide what to retrieve at each step.



LLM and Retrieval Model Architectures

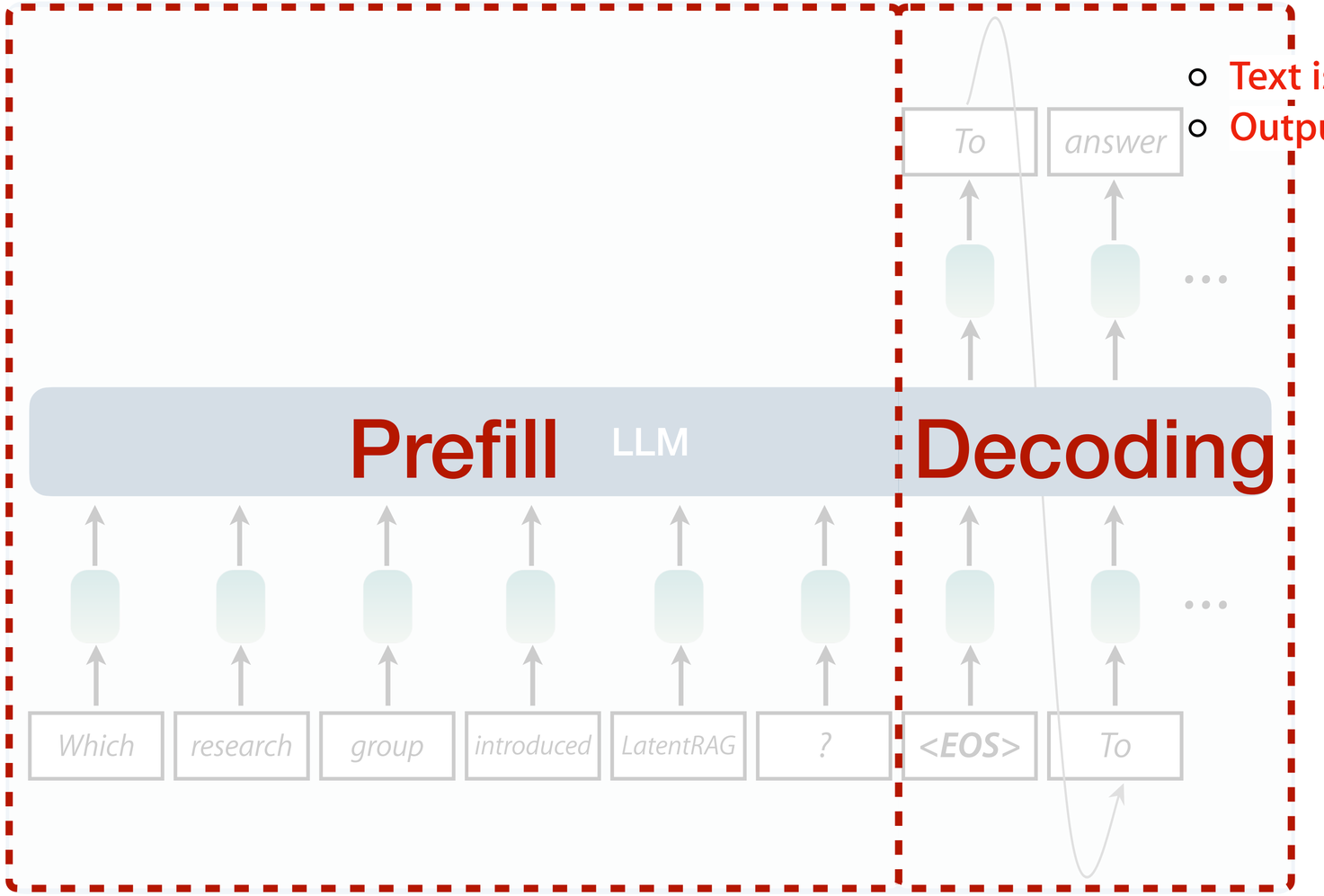
The LLM decodes latent representations into text, which is then fed into the retrieval model.
The retrieval model re-encodes the text into latent representations.
Do we really need these intermediate text representations?





Limitations of text-based agentic RAG

- Autoregressive text decoding is extremely slow.



  - Text is a verbose low-information-density medium
  - Output tokens must be decoded one by one.

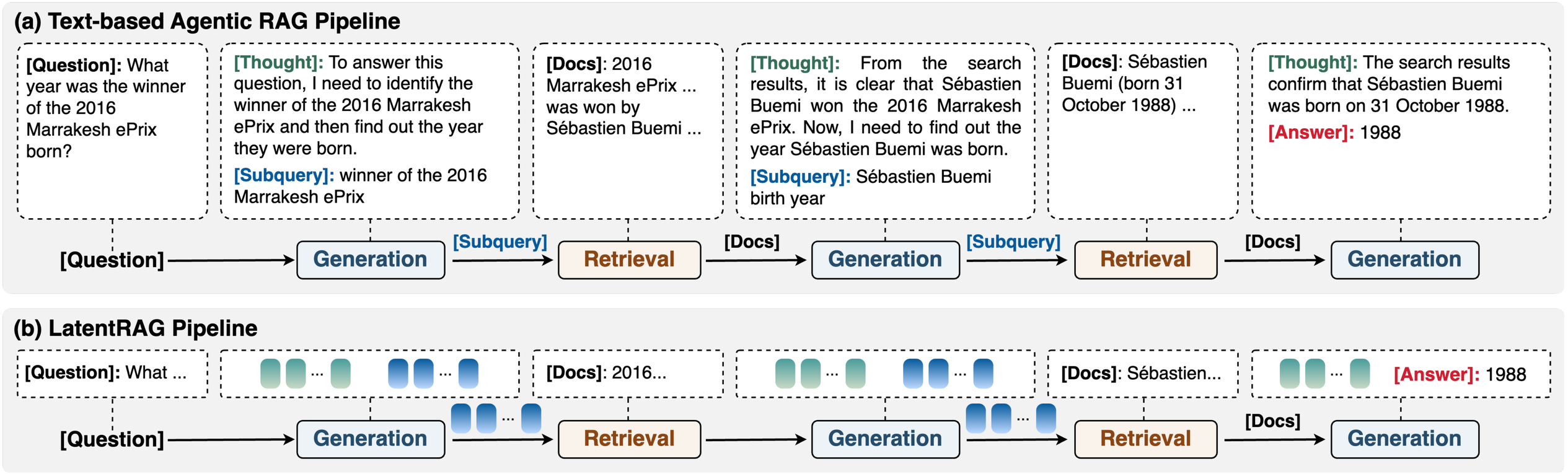
- Discrete textural tokens limit end-to-end optimization between the retrieval model and the LLM.

From text-based agentic RAG to LatentRAG

People naturally use words to search for information, but does an AI agent need to do the same?

LatentRAG

Both the LLM and the retrieval model internally work with latent representations other than words, making it possible for the LLM agent to search without first converting its internal representations into words.

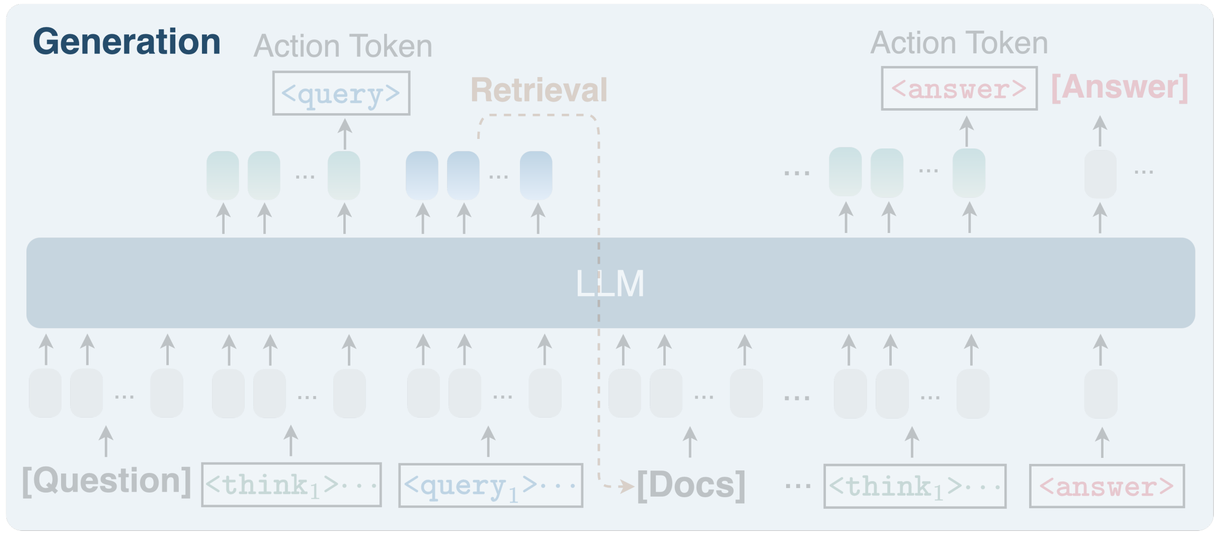


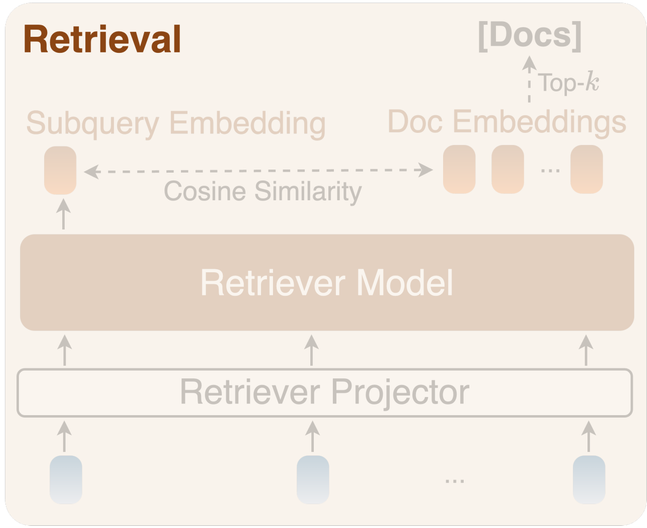
··· Latent thought tokens

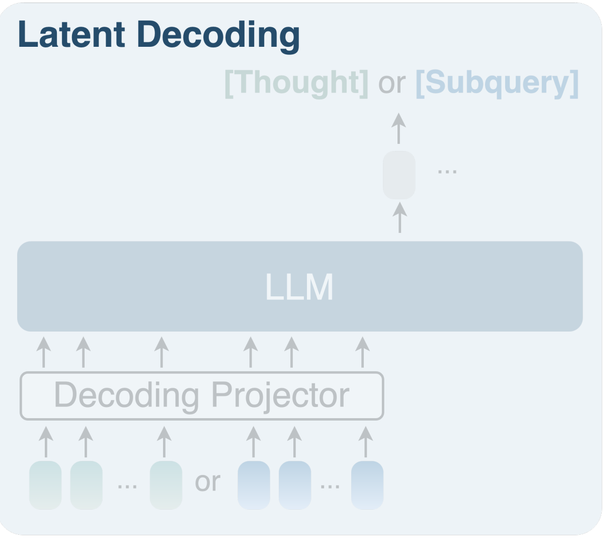
··· Latent subquery tokens

··· Hidden states of text tokens

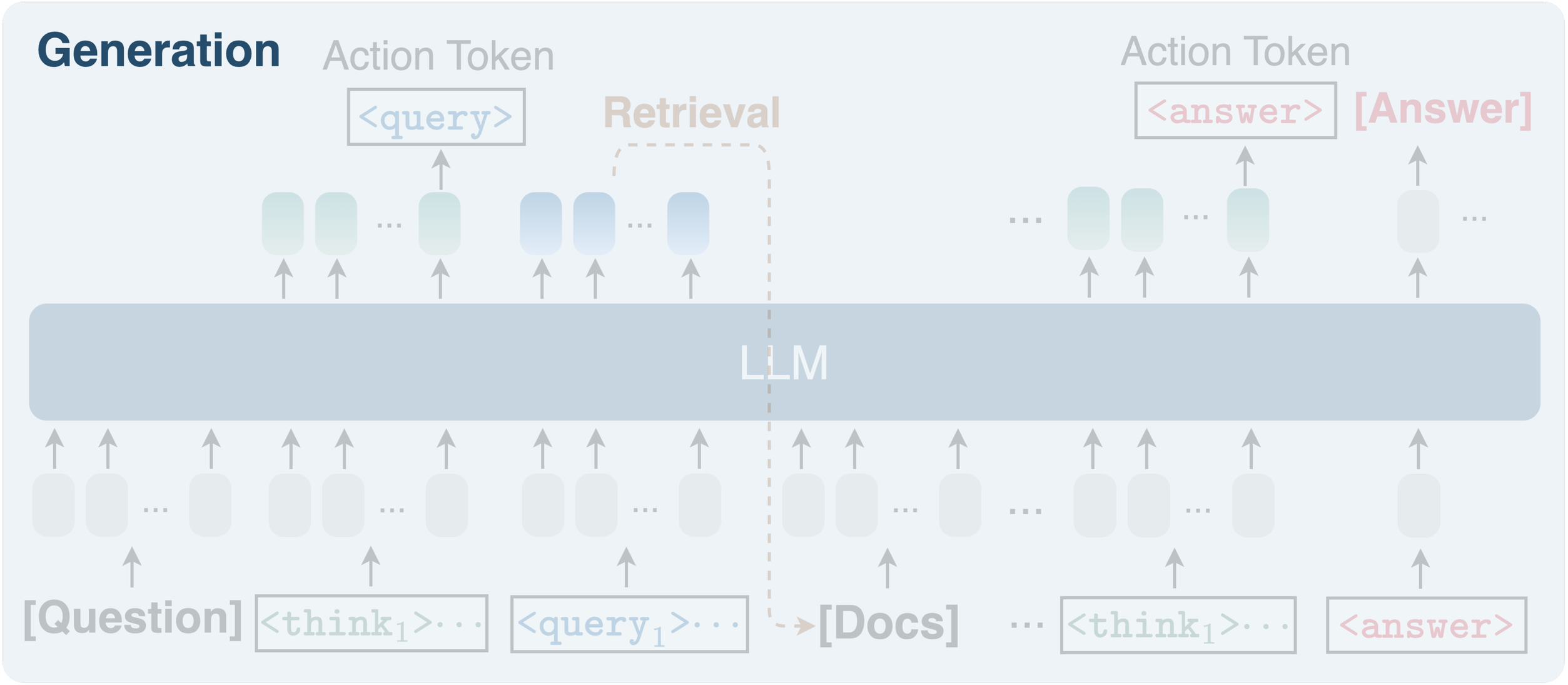
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
<think1>···
<answer>

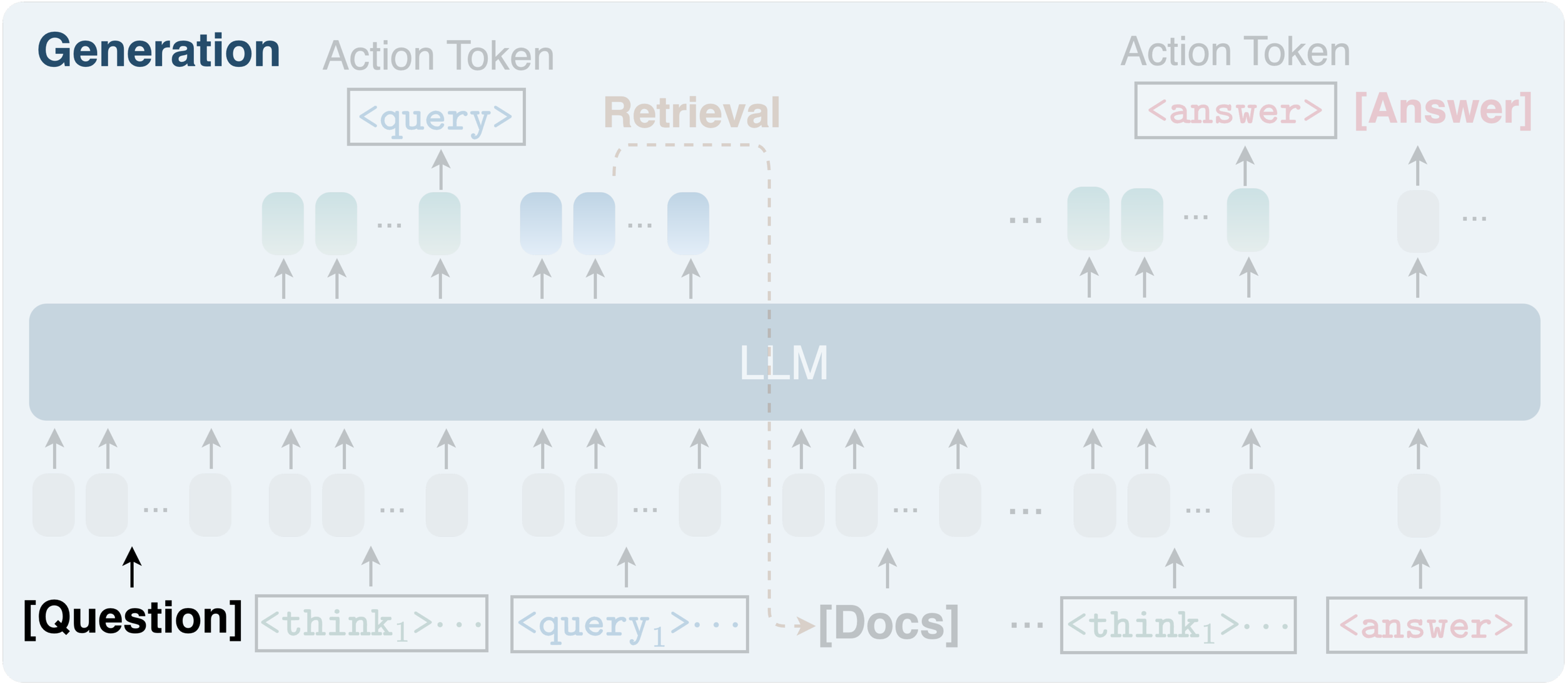

Retrieval
[Docs]
Top-k
Subquery Embedding
Doc Embeddings
Cosine Similarity
Retriever Model
Retriever Projector


Latent Decoding
[Thought] or [Subquery]
LLM
Decoding Projector
or

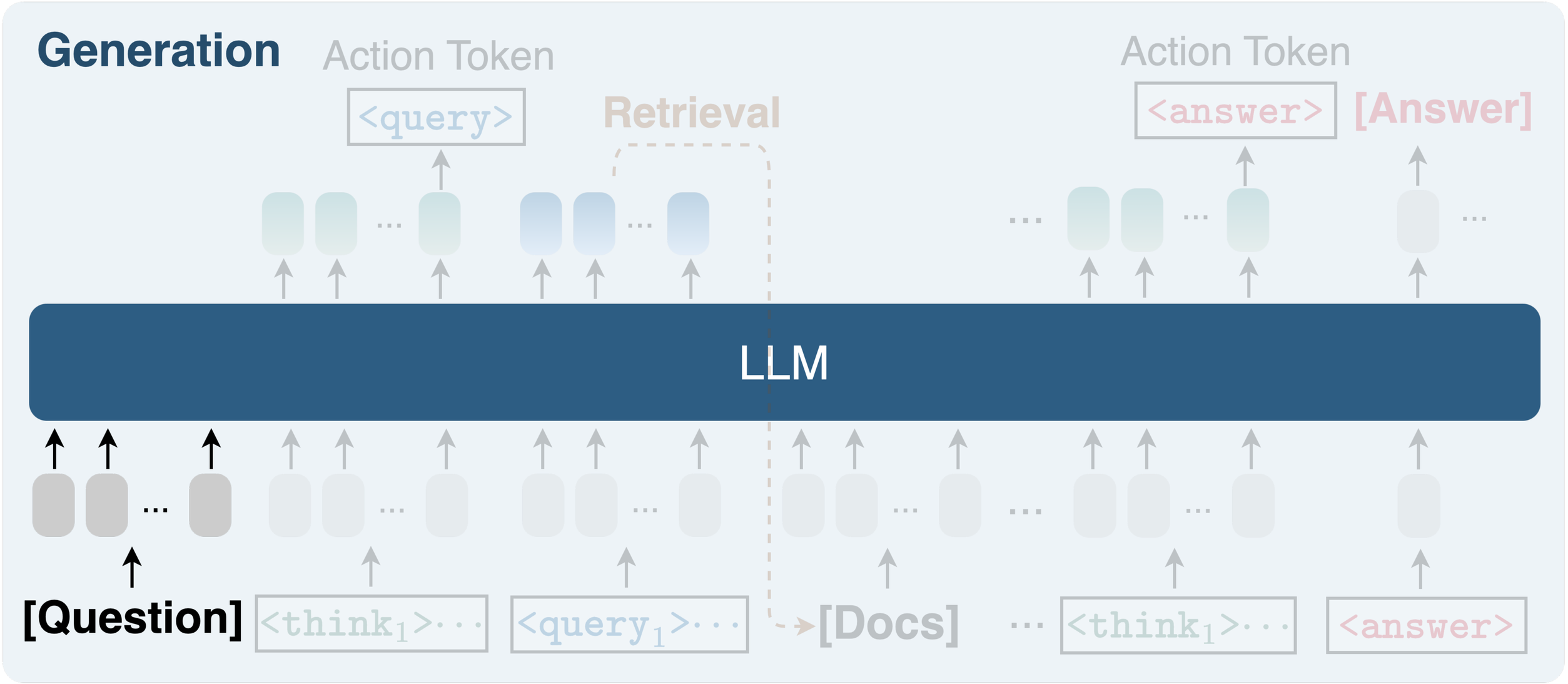
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
···
<think1>···
<answer>

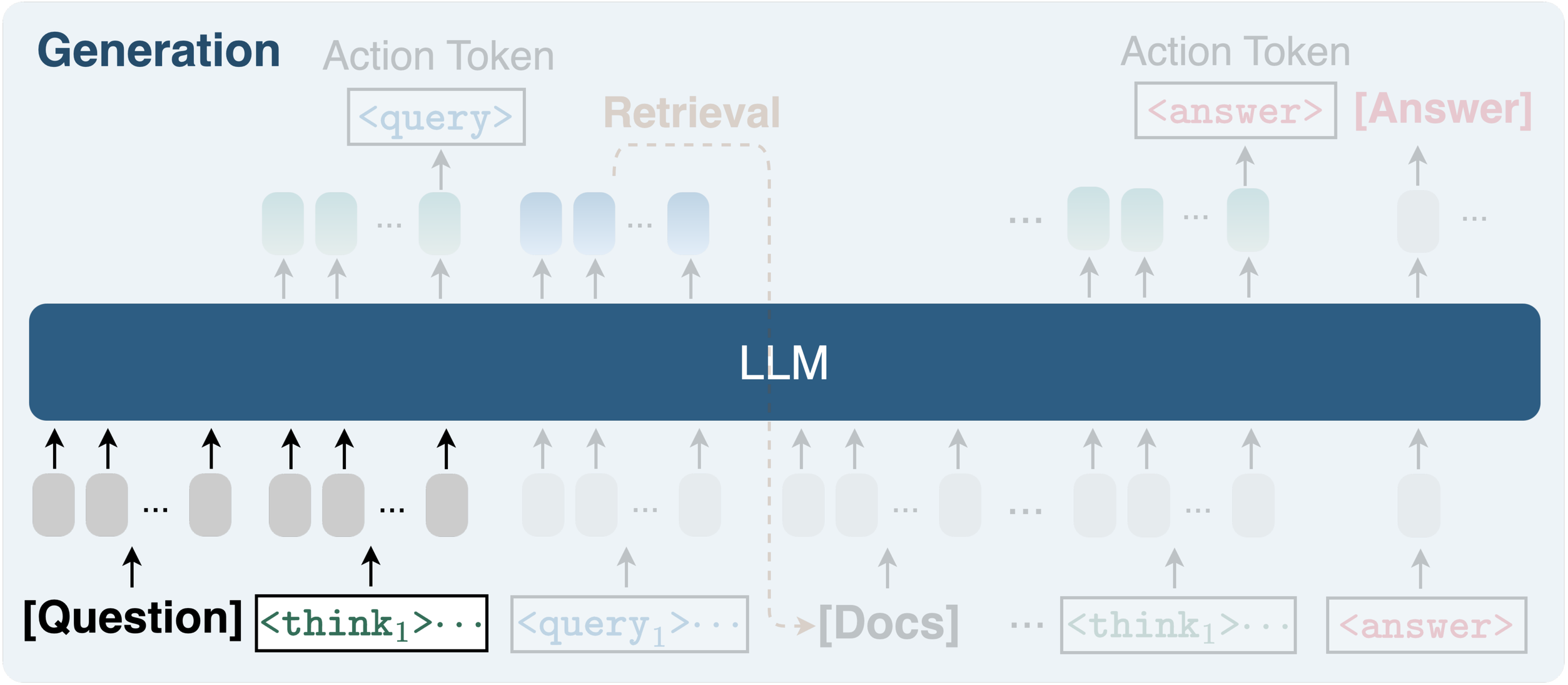
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
···
<think1>···
<answer>

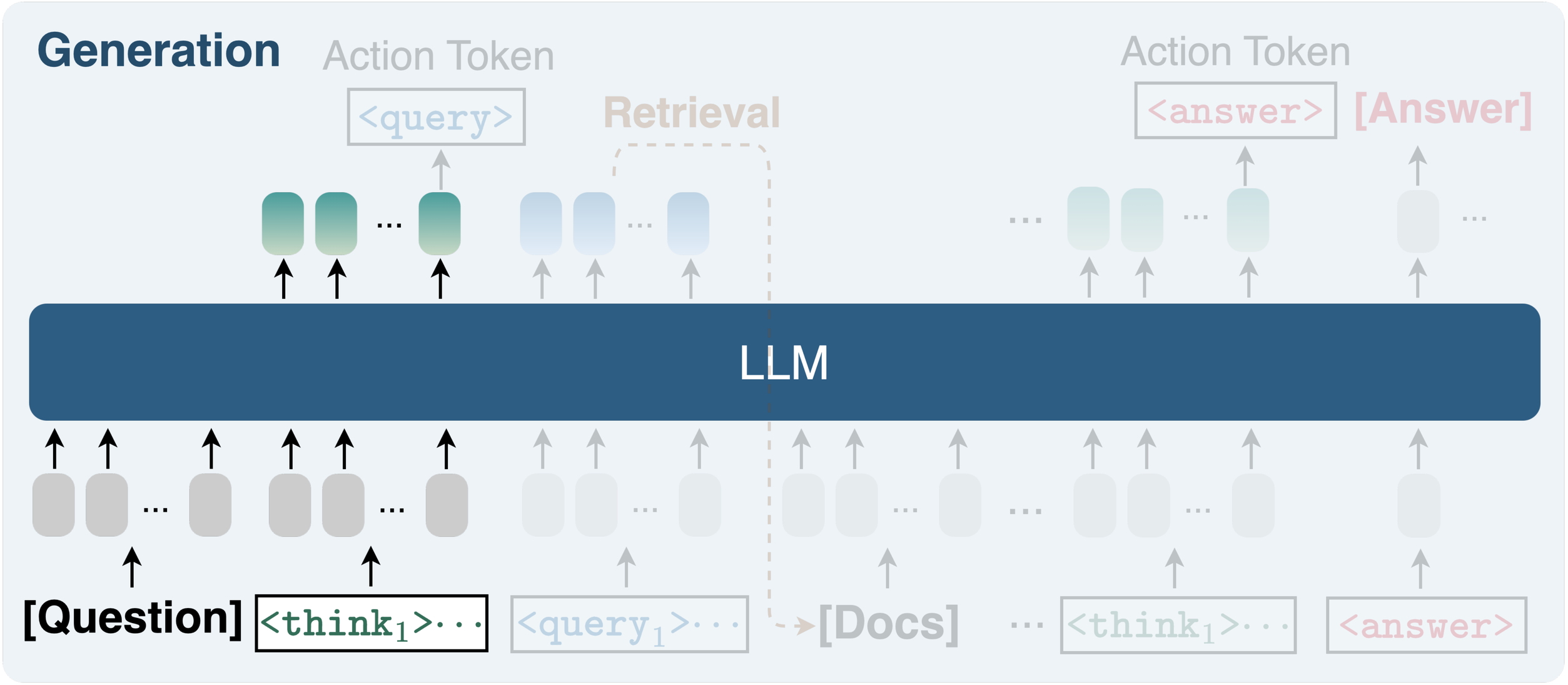
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
<think1>···
<answer>

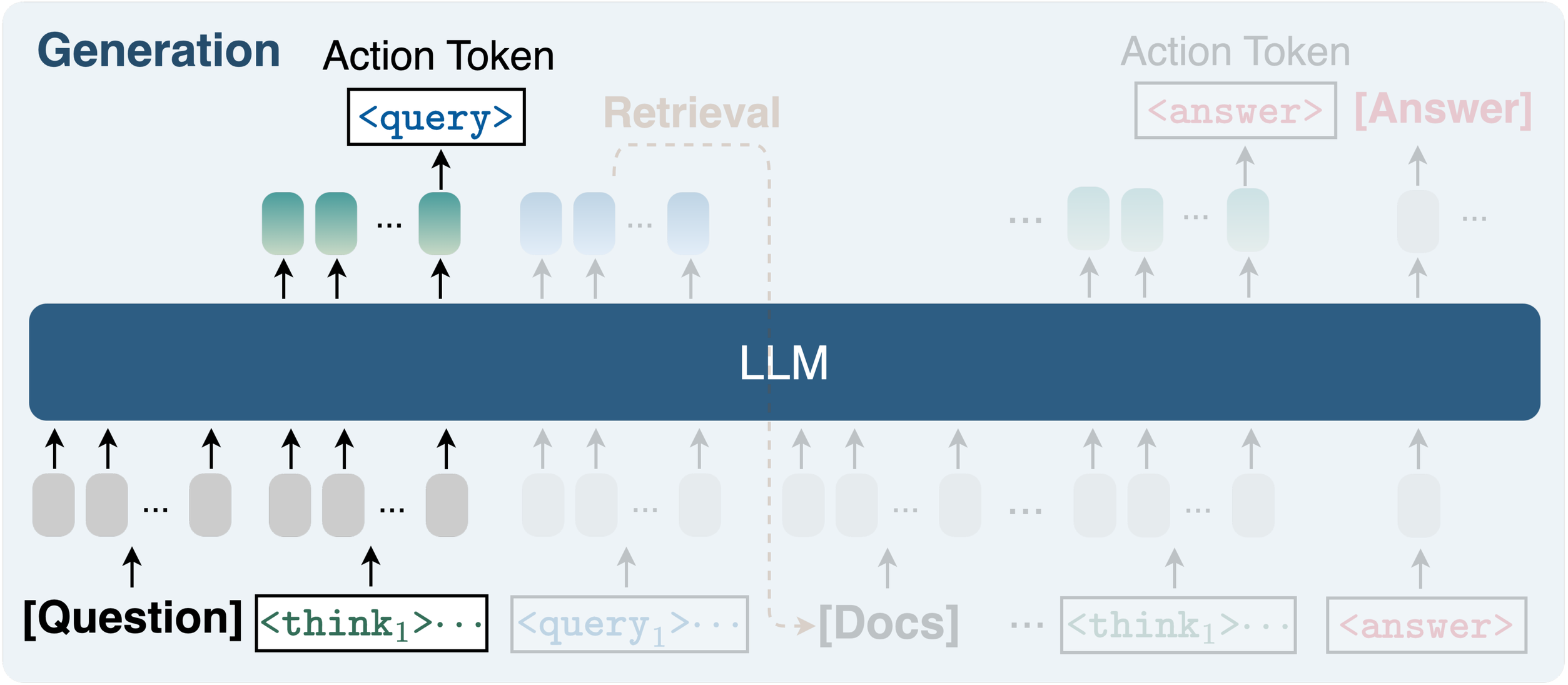
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
<think1>···
<answer>

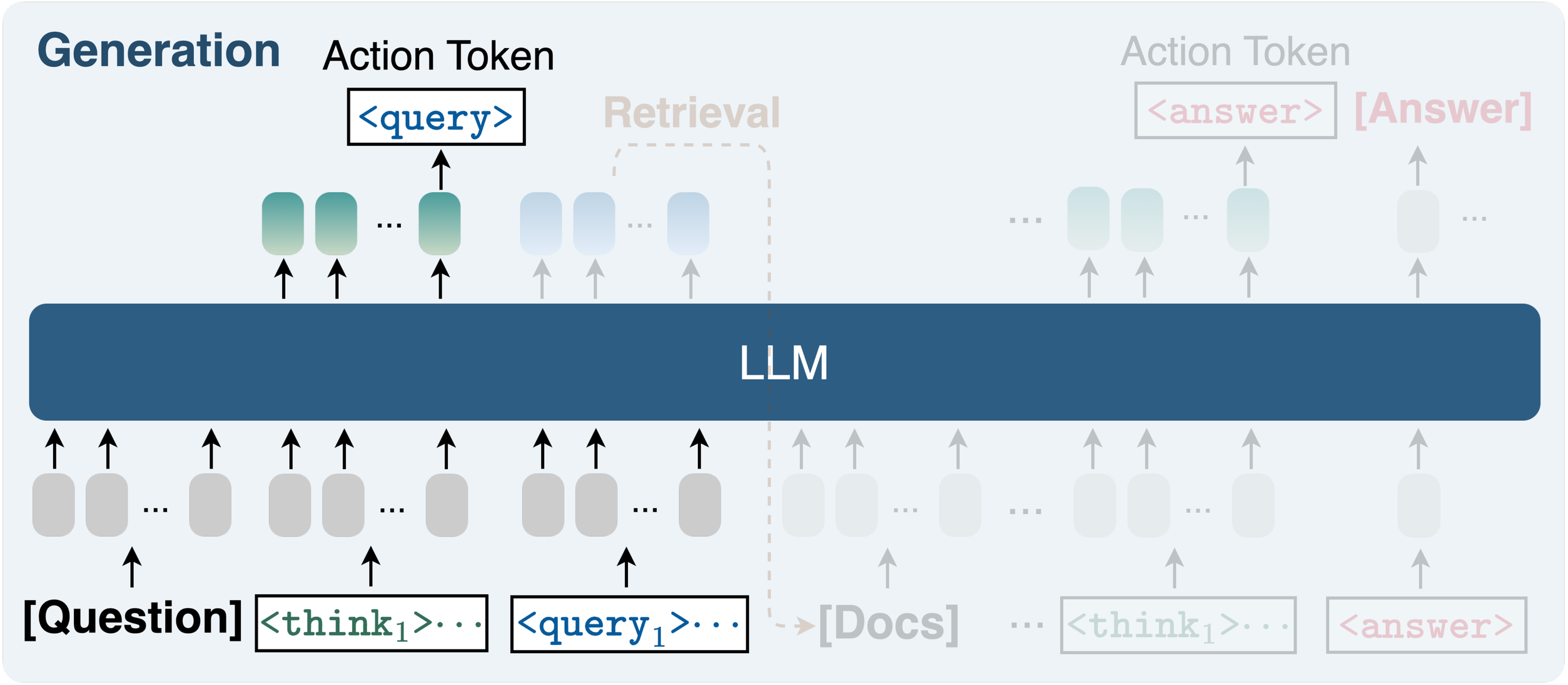
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
<think1>···
<answer>

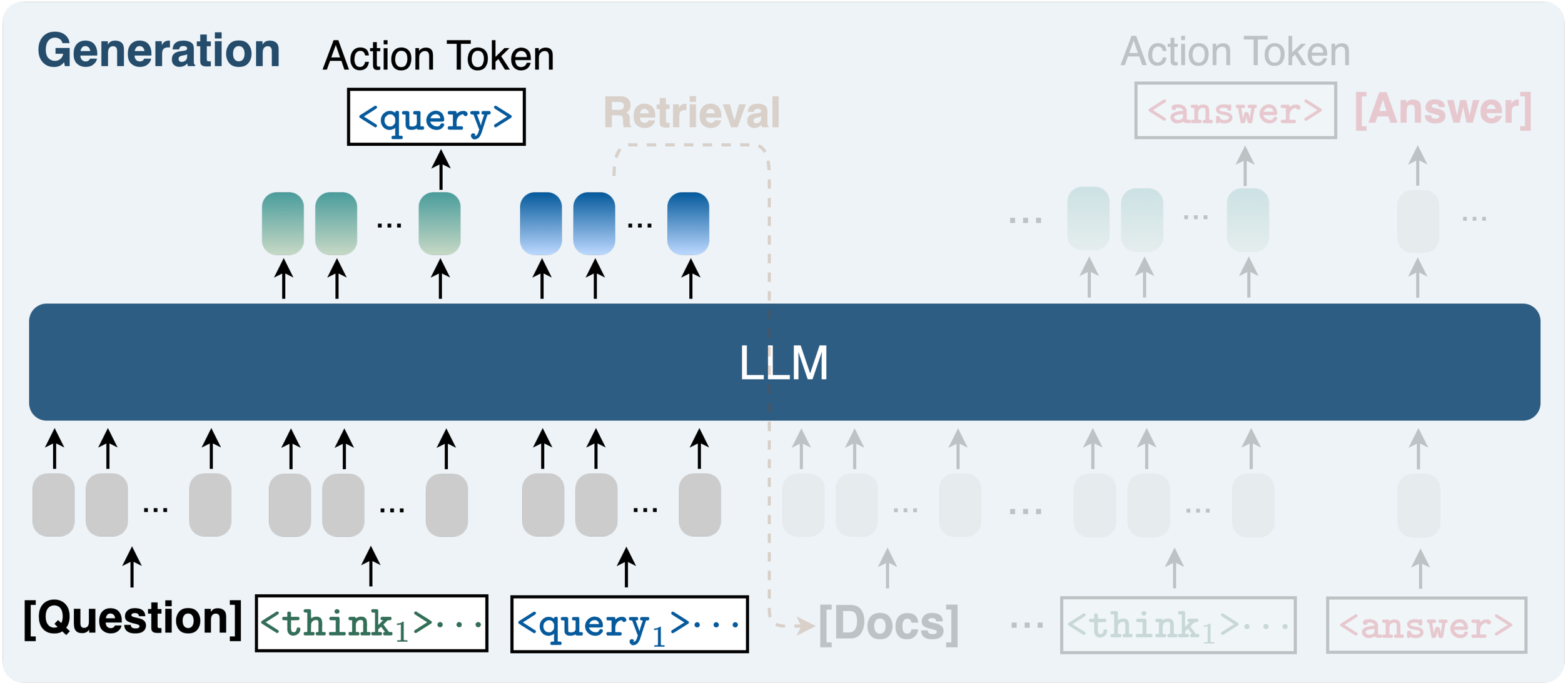
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
<think1>···
<answer>

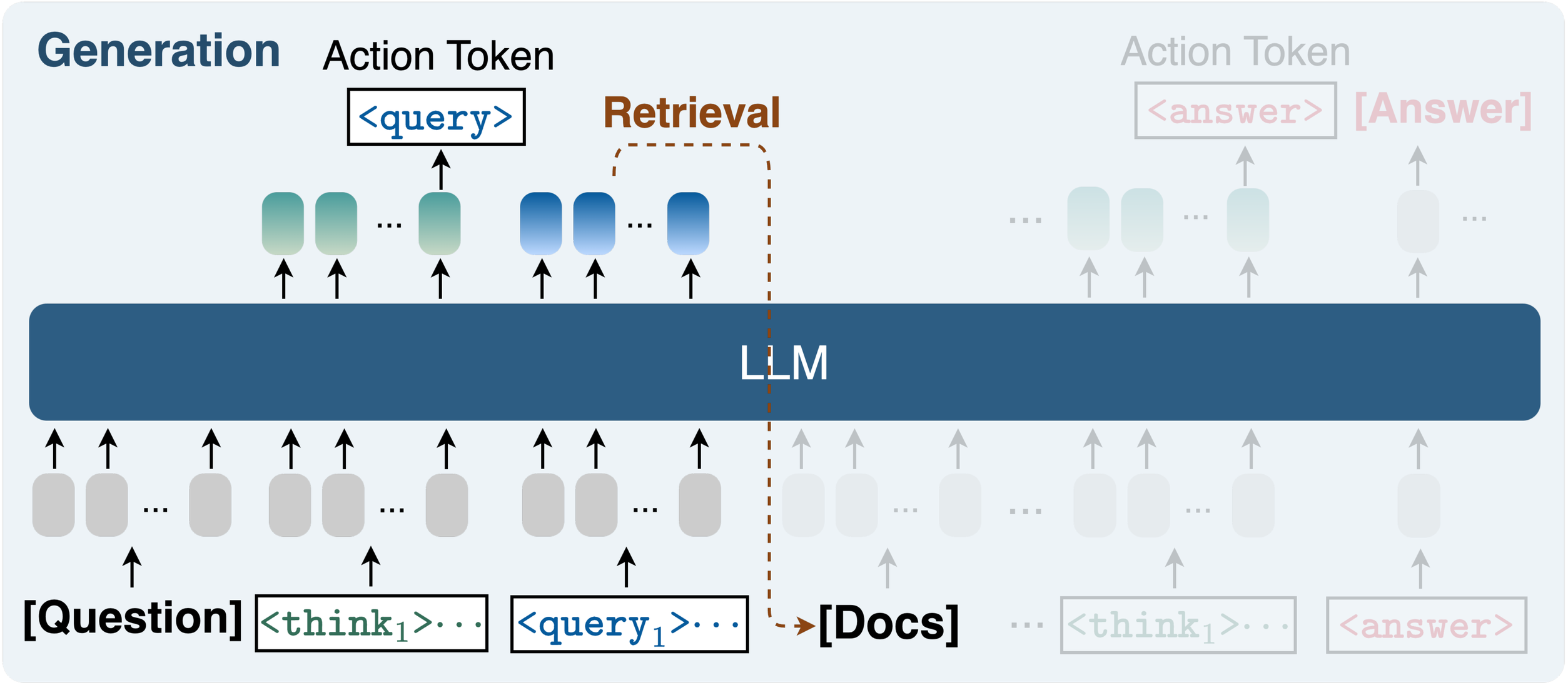
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think$_1$>···
<query$_1$>···
[Docs]
<think$_1$>···
<answer>

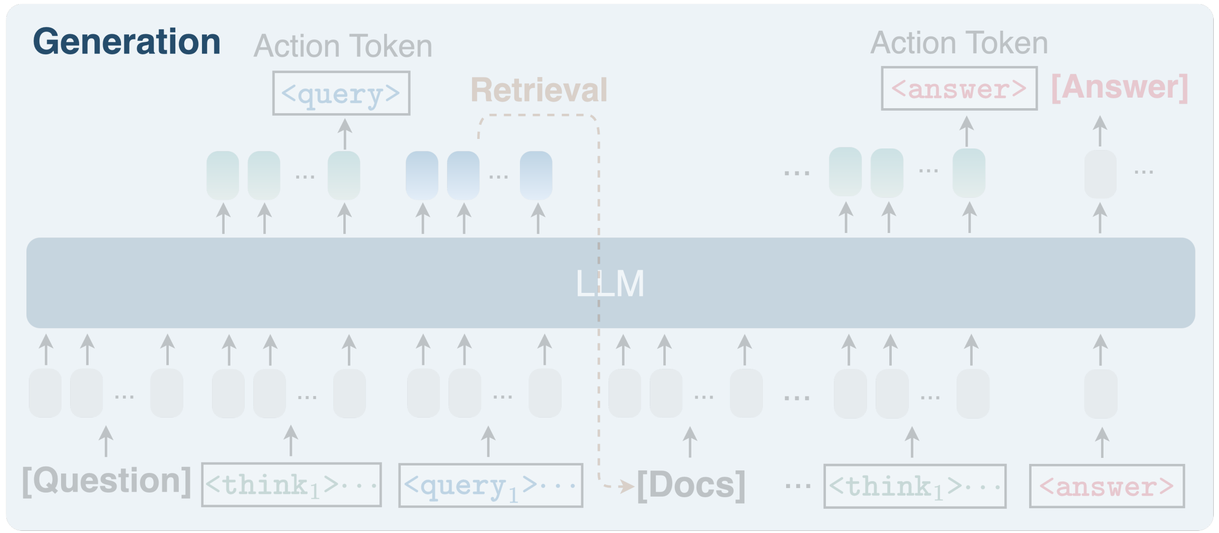
LatentRAG

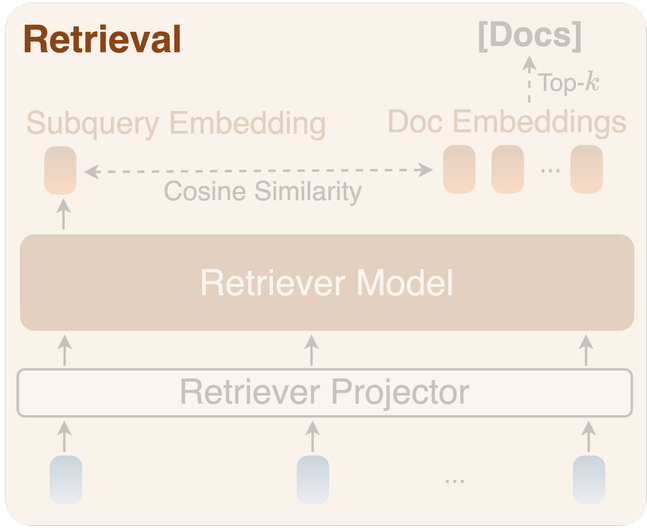

Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think$_1$>···
<query$_1$>···
[Docs]
<think$_1$>···
<answer>

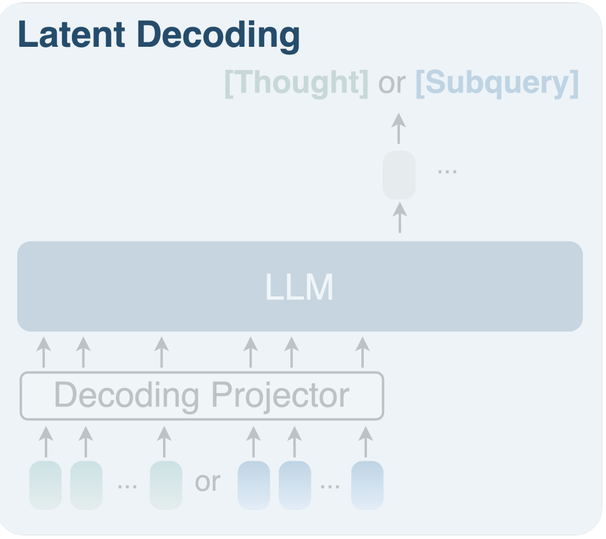
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
<think1>···
<answer>

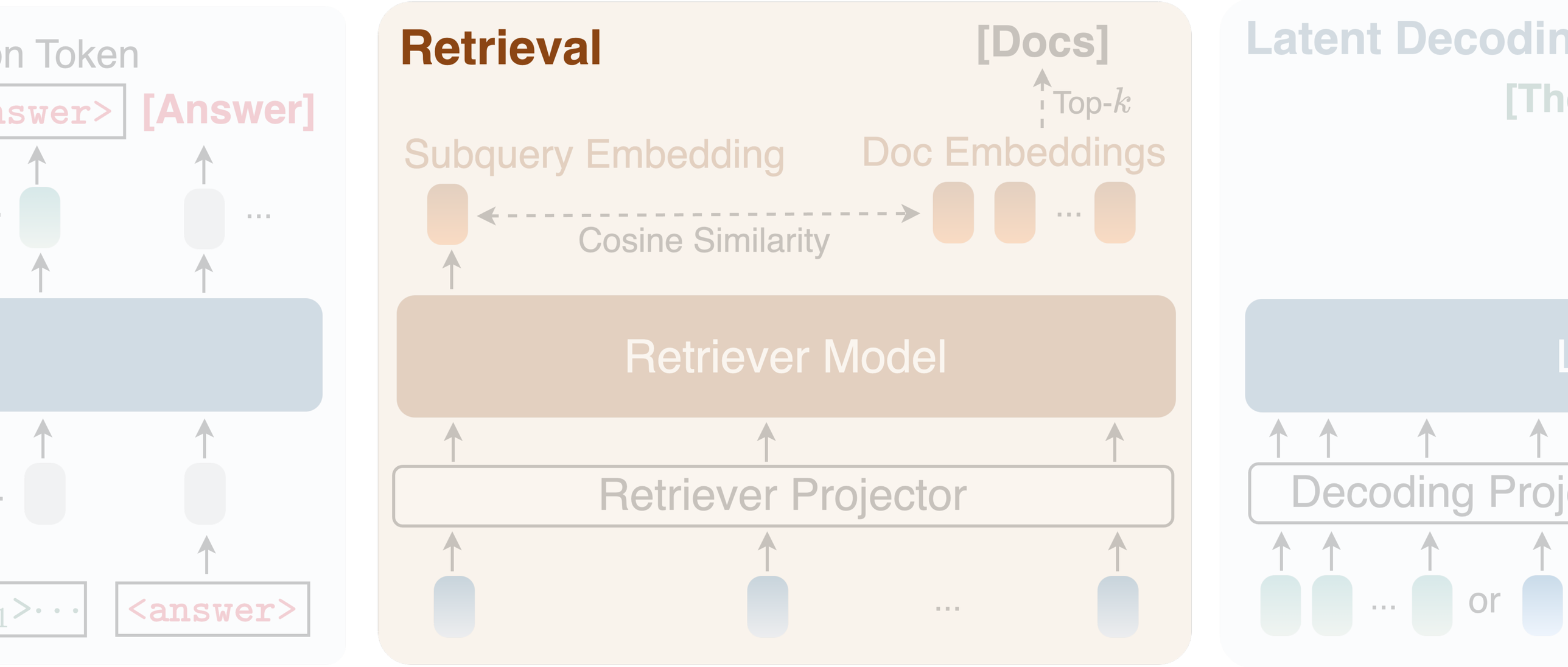
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
<think1>···
<answer>

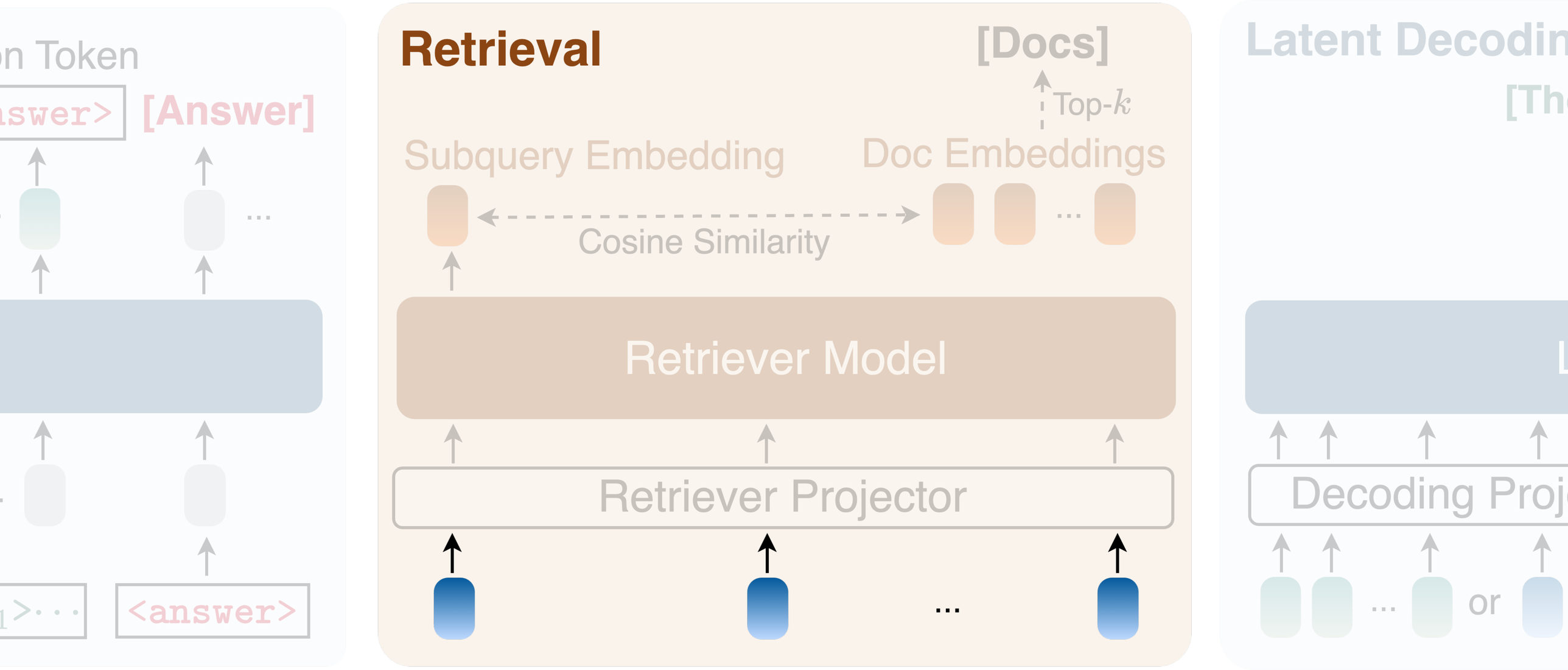

Retrieval
[Docs]
Top-k
Subquery Embedding
Doc Embeddings
Cosine Similarity
Retriever Model
Retriever Projector


Latent Decoding
[Thought] or [Subquery]
LLM
Decoding Projector
or

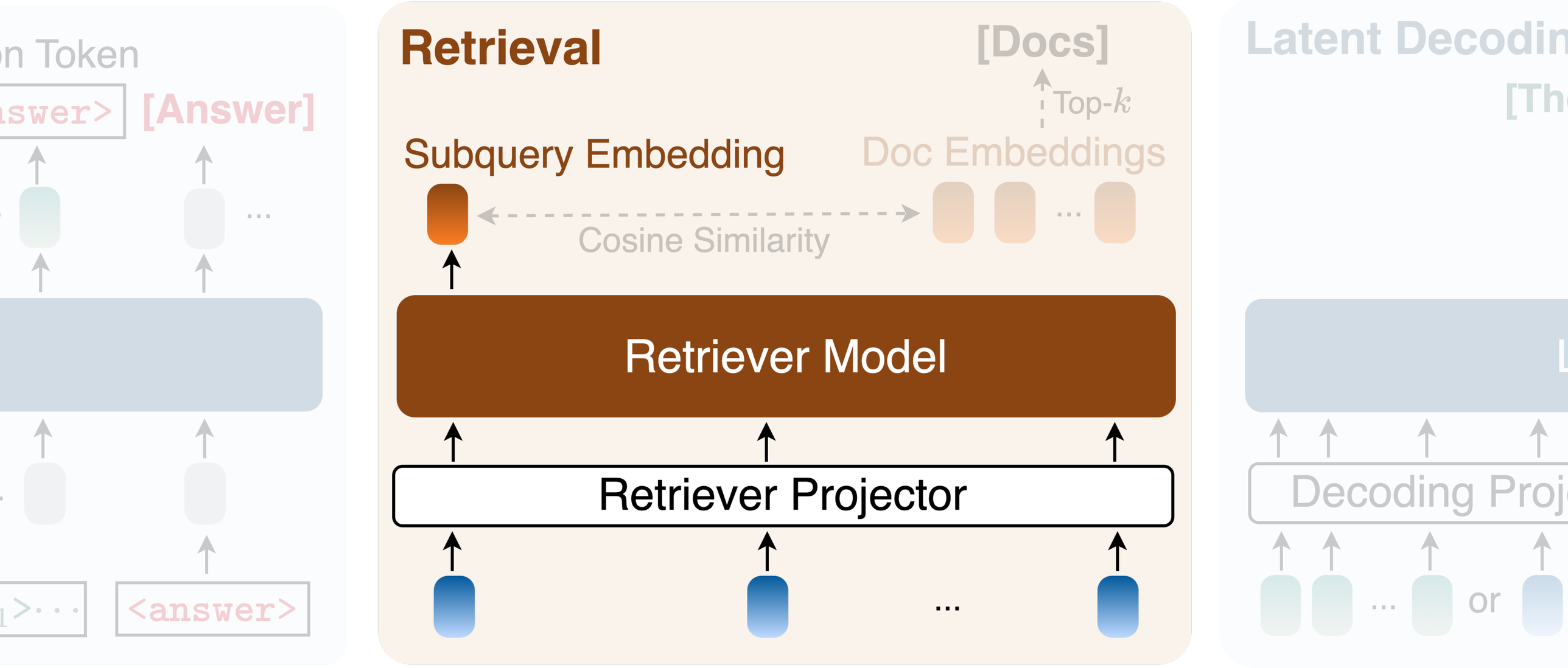
LatentRAG


n Token
swer>
[Answer]
...
<answer>
Retrieval
[Docs]
Top-k
Subquery Embedding
Doc Embeddings
Cosine Similarity
Retriever Model
Retriever Projector
...
Latent Decodin
[Th
Decoding Proj
...
or

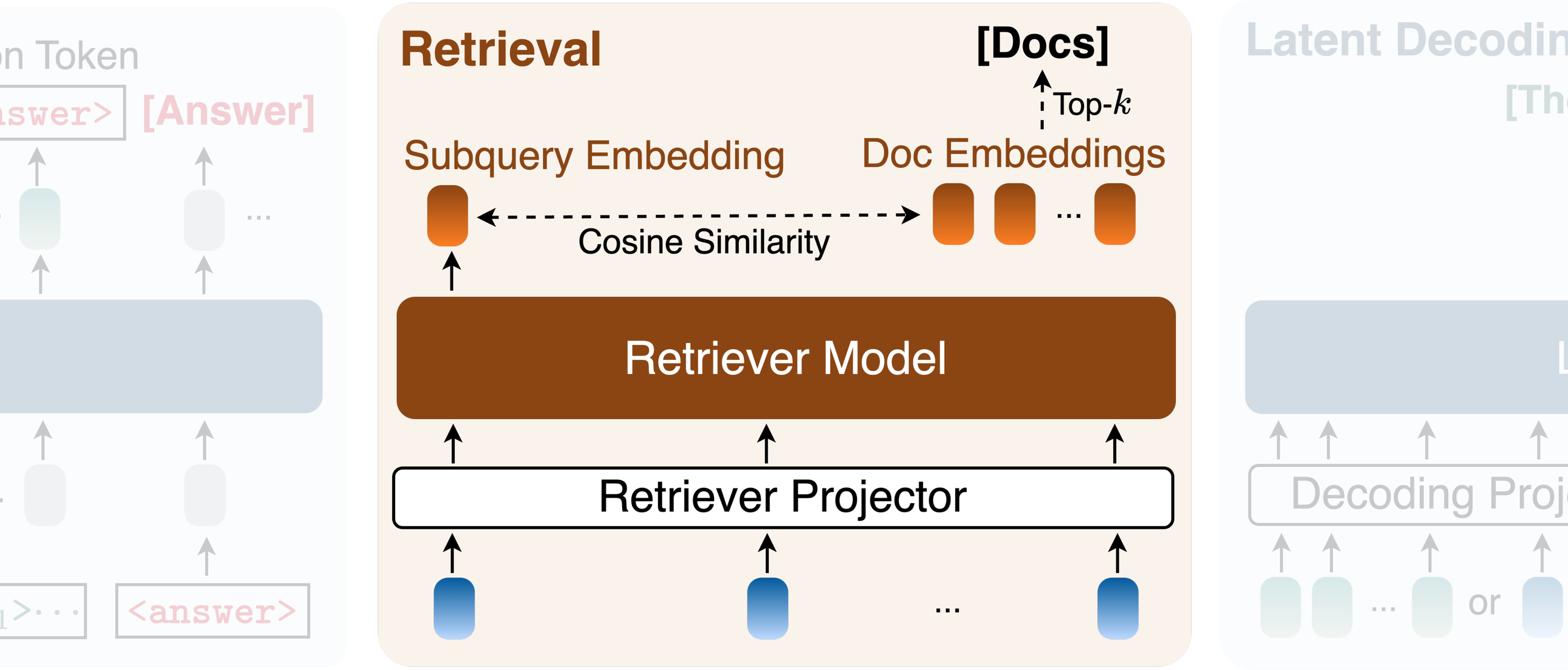
LatentRAG


n Token
swer>
[Answer]
...
<answer>
Retrieval
[Docs]
Top-k
Subquery Embedding
Doc Embeddings
Cosine Similarity
Retriever Model
Retriever Projector
...
Latent Decodin
[Th
Decoding Proj
...
or

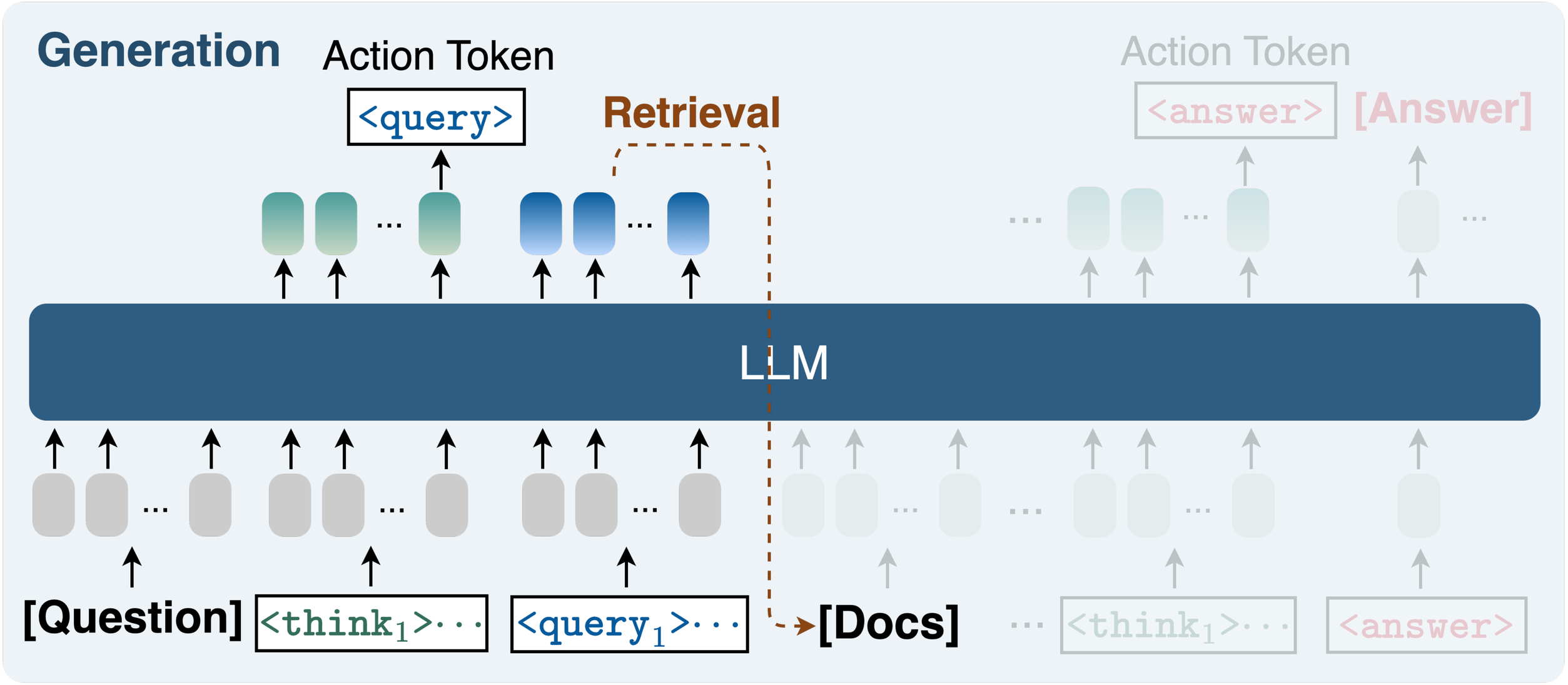
LatentRAG


Retrieval
[Docs]
Top-k
Subquery Embedding
Doc Embeddings
Cosine Similarity
Retriever Model
Retriever Projector
...
Latent Decodin
[Th
Decoding Proj
or
[Answer]
<answer>

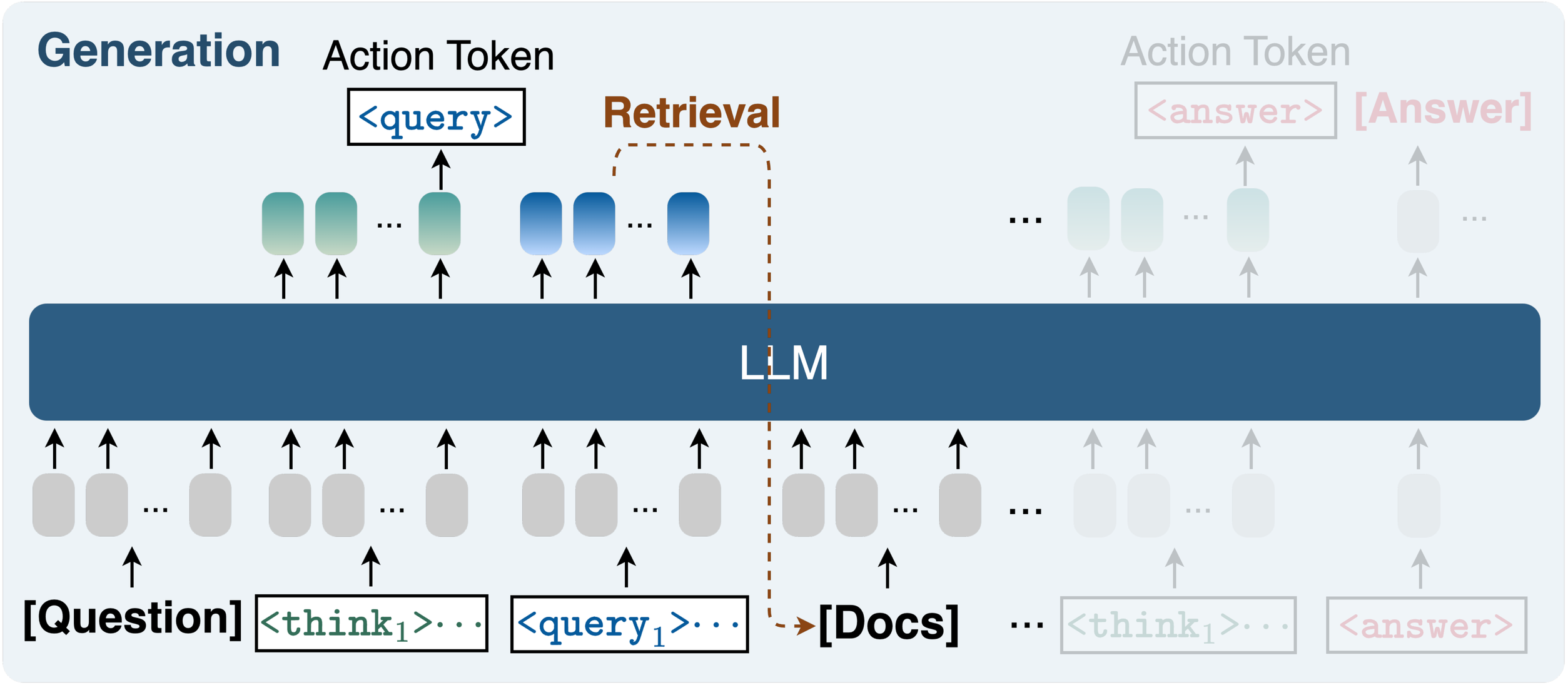
LatentRAG


Retrieval
[Docs]
Top-k
Subquery Embedding
Doc Embeddings
Cosine Similarity
Retriever Model
Retriever Projector
...
[Answer]
<answer>
Decoding Proj
or

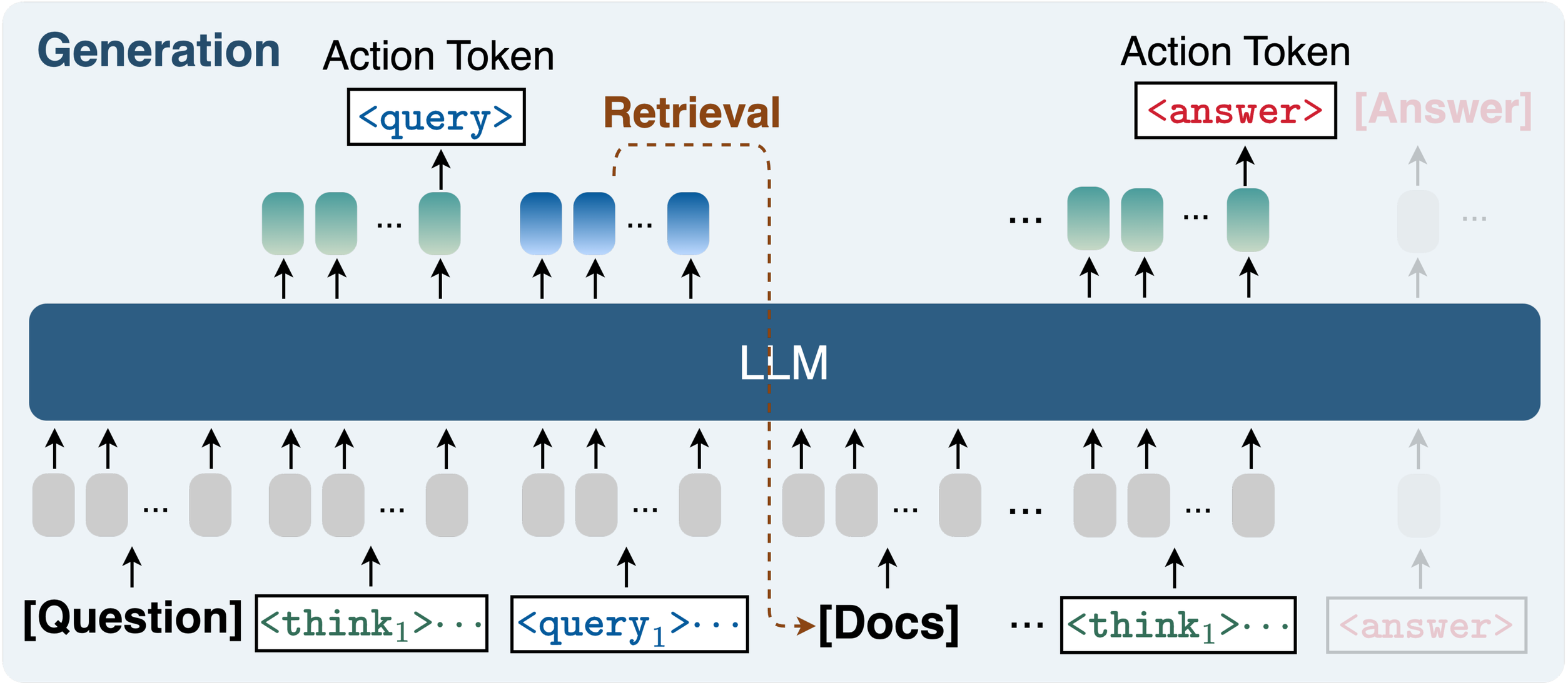
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
<think1>···
<answer>

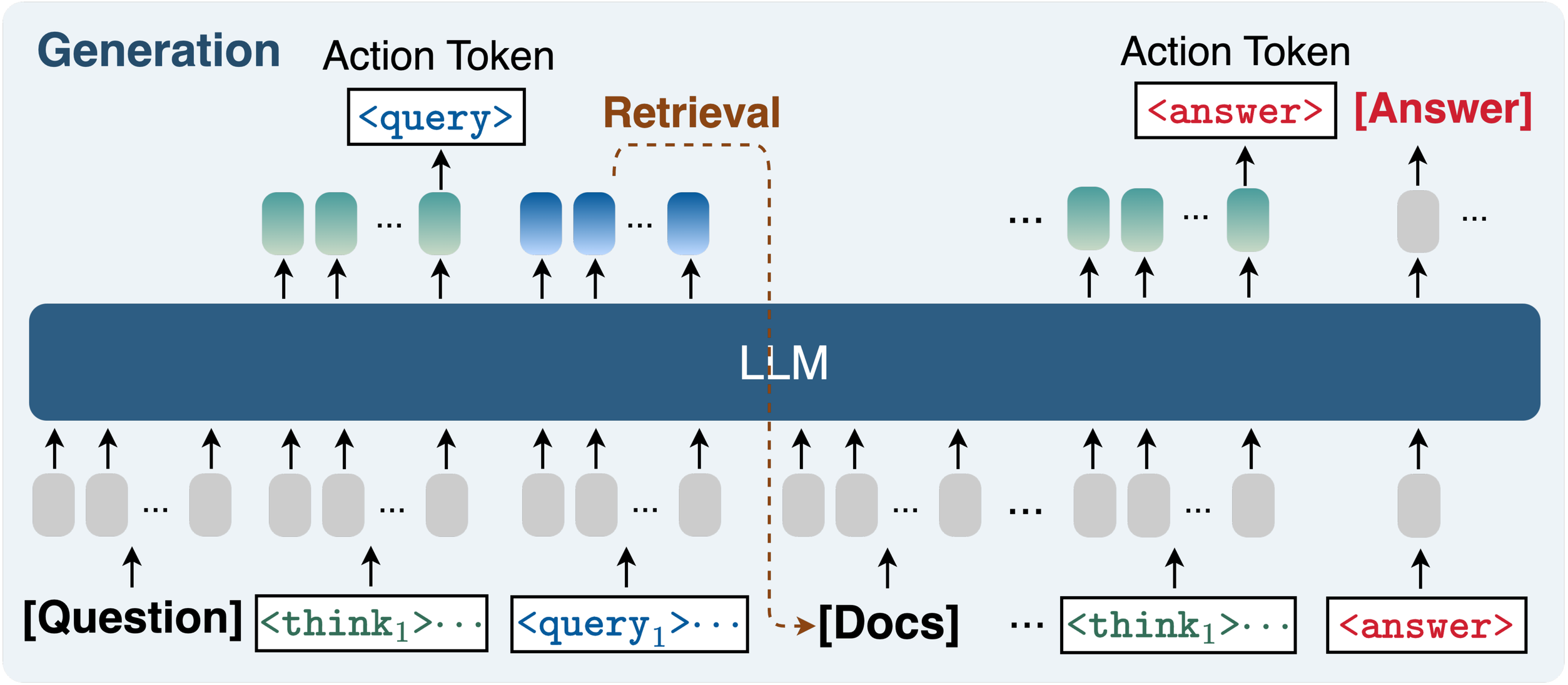
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
<think1>···
<answer>

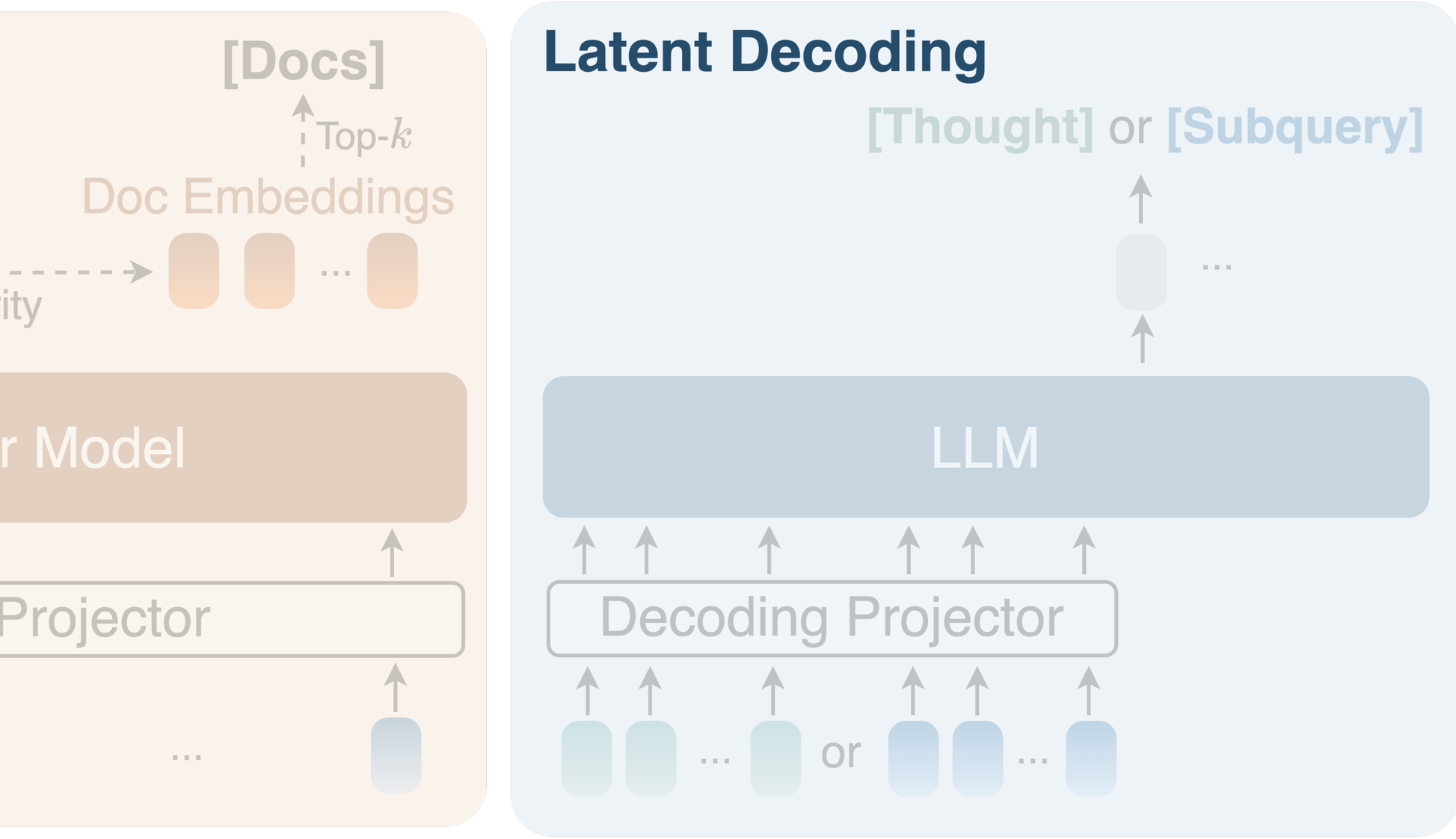
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
<think1>···
<answer>

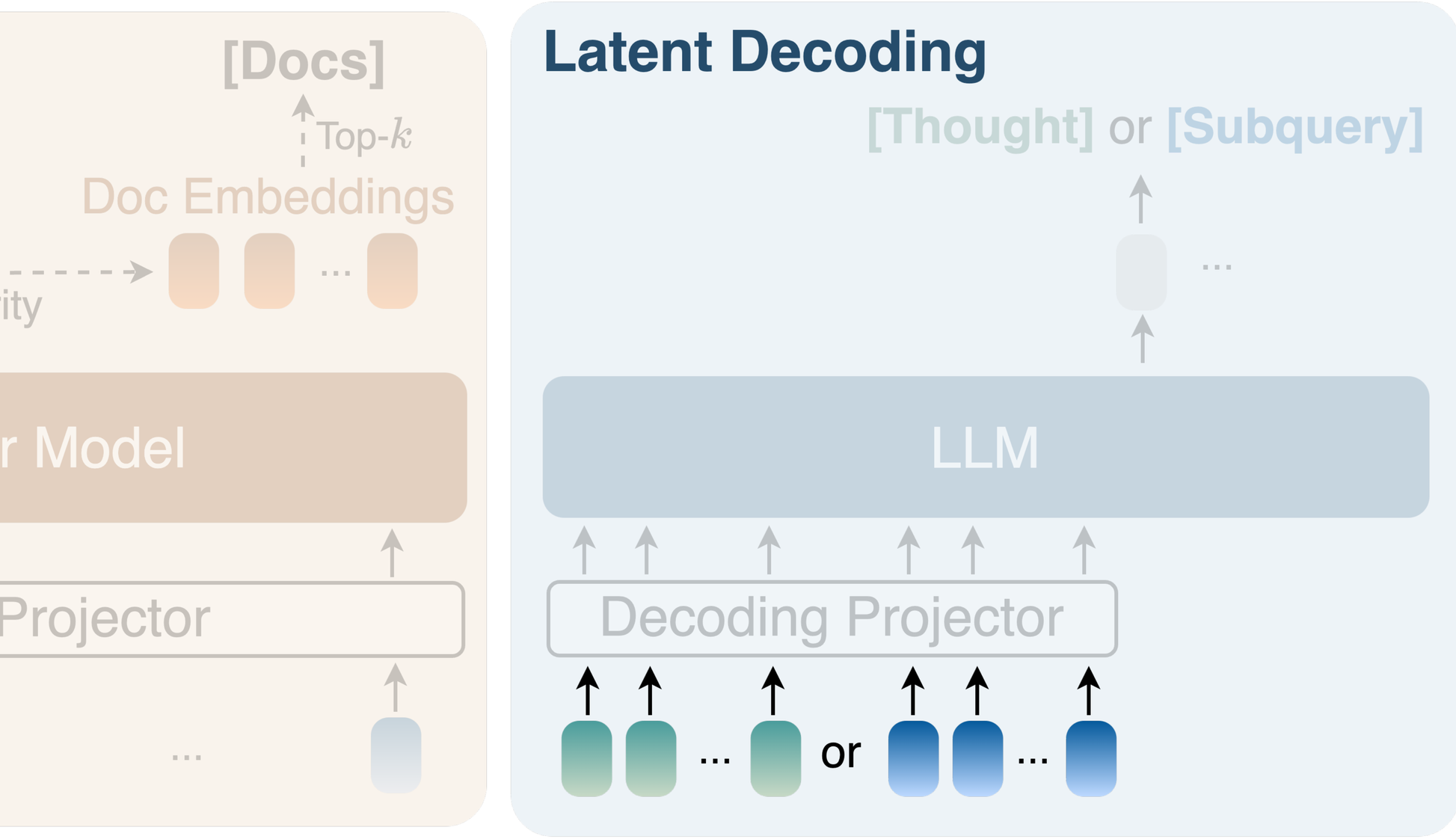
LatentRAG


Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
<think1>···
<answer>

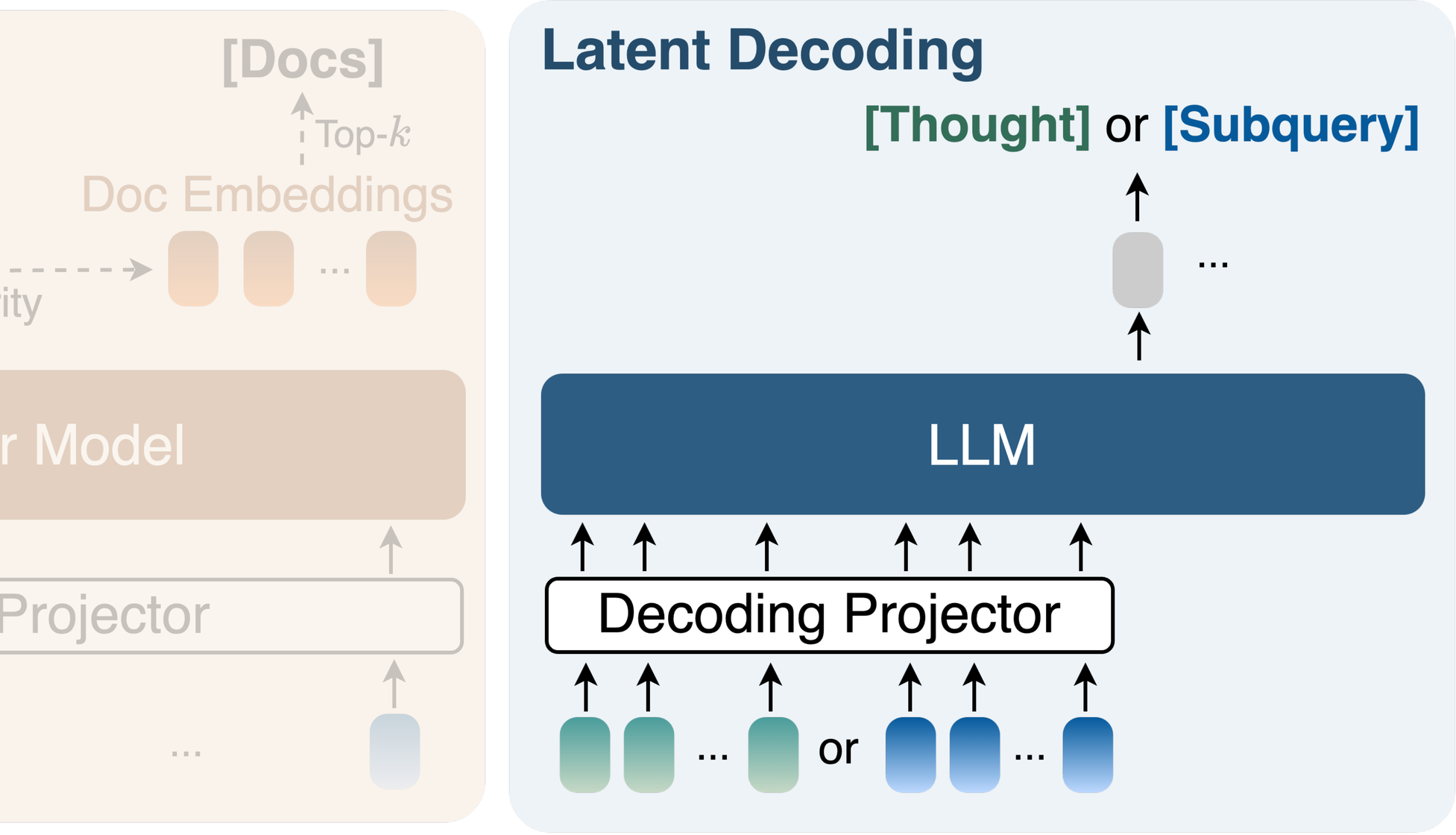
LatentRAG


[Docs]
Top-k
Doc Embeddings
...
ity
r Model
Projector
...
Latent Decoding
[Thought] or [Subquery]
...
LLM
Decoding Projector
...
or
...

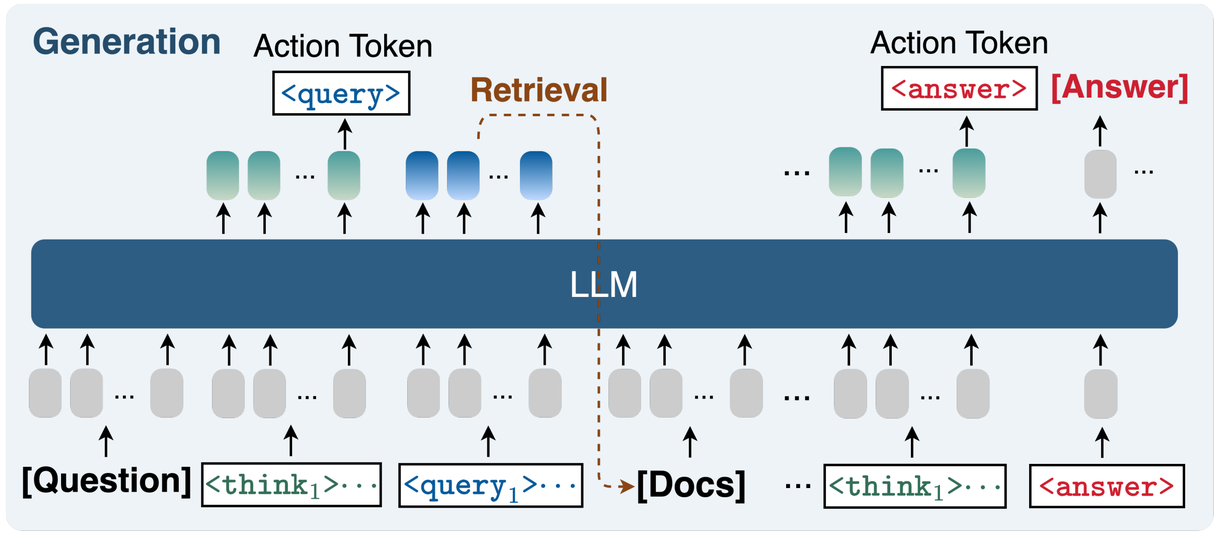
LatentRAG

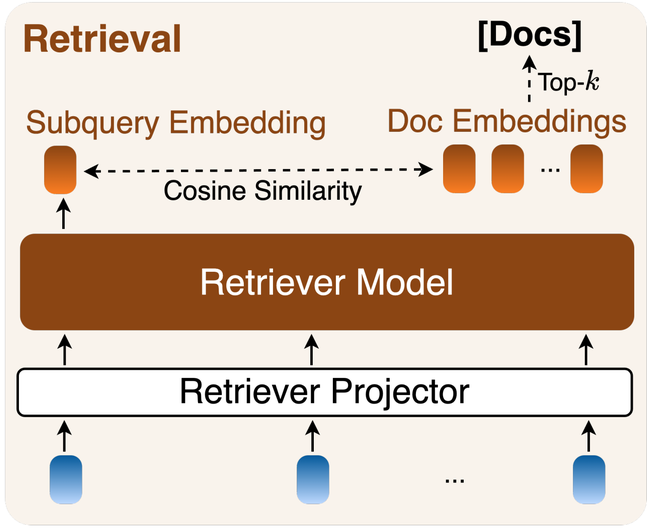

[Docs]
Top-k
Doc Embeddings
...
ity
r Model
Projector
...
Latent Decoding
[Thought] or [Subquery]
...
LLM
Decoding Projector
...
or
...

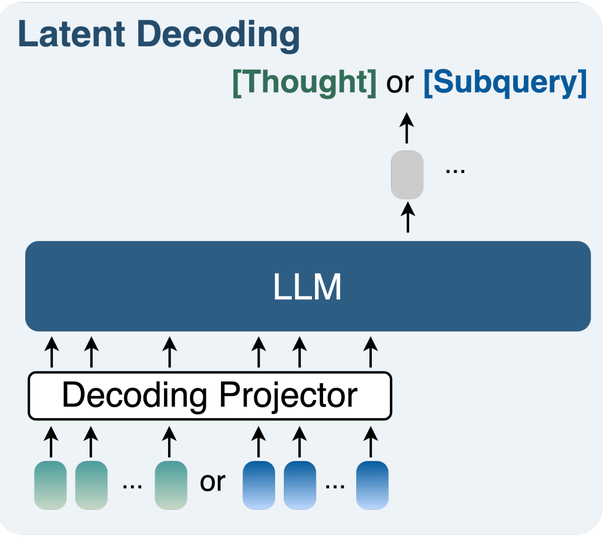
LatentRAG


[Docs]
Top-k
Doc Embeddings
...
ity
r Model
Projector
...
Latent Decoding
[Thought] or [Subquery]
...
LLM
Decoding Projector
...
or
...

LatentRAG

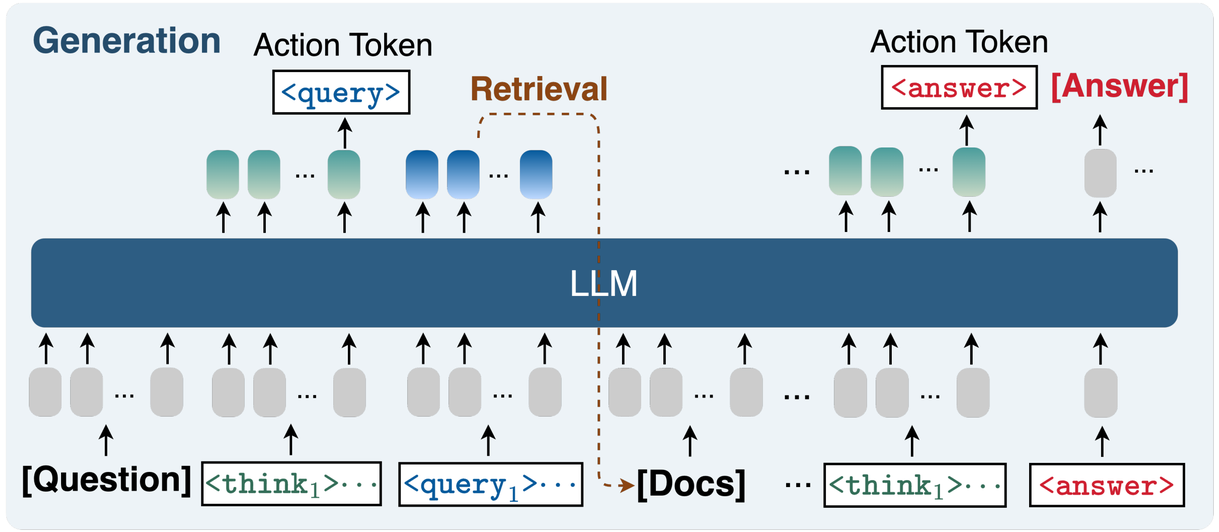

Generation
Action Token
<query>
Retrieval
Action Token
<answer>
[Answer]
LLM
[Question]
<think1>···
<query1>···
[Docs]
<think1>···
<answer>

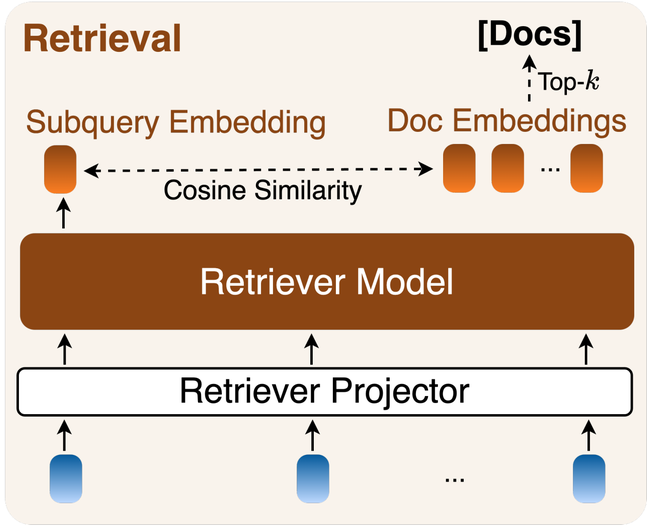

Retrieval
[Docs]
Top-k
Subquery Embedding
Doc Embeddings
Cosine Similarity
Retriever Model
Retriever Projector

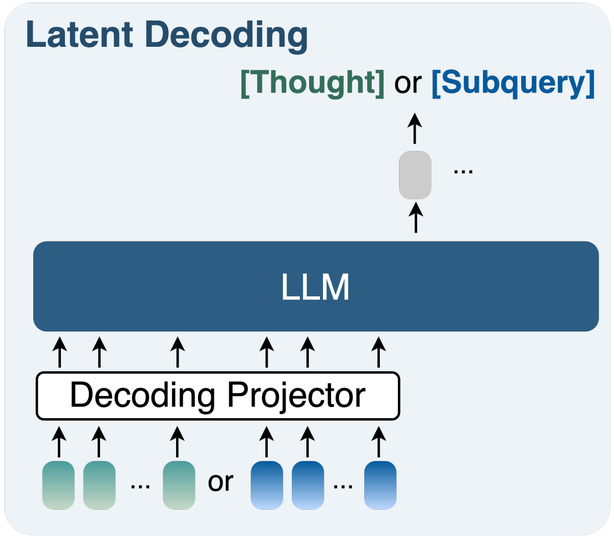

Latent Decoding
[Thought] or [Subquery]
LLM
Decoding Projector
or

LatentRAG

- Supervised finetuning using trajectories generated by an existing text-based agentic RAG method

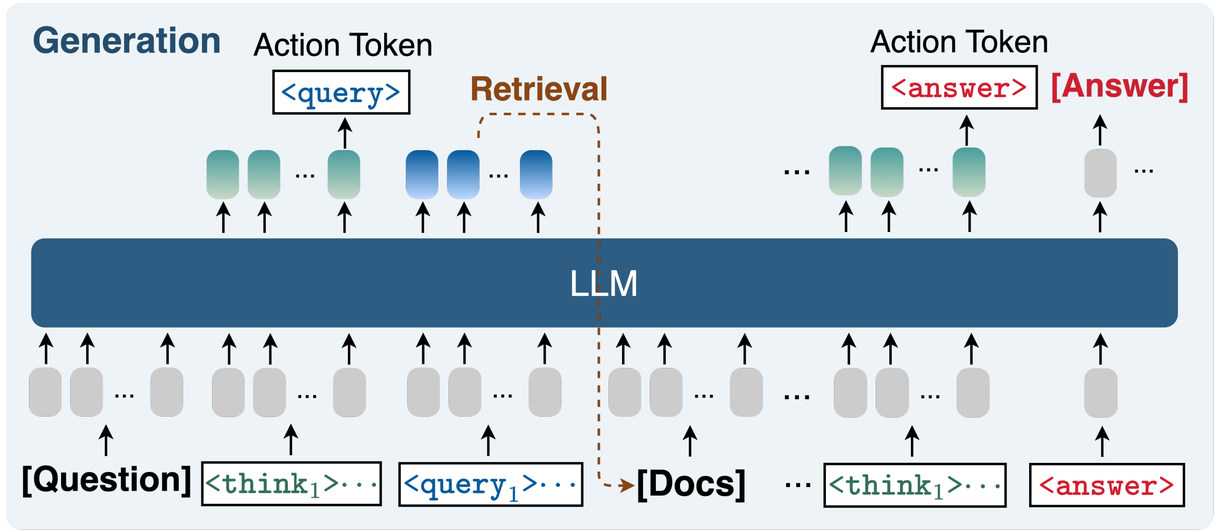


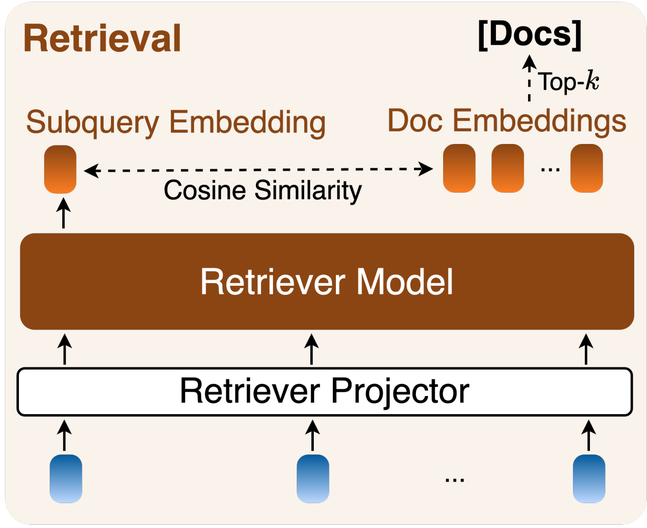


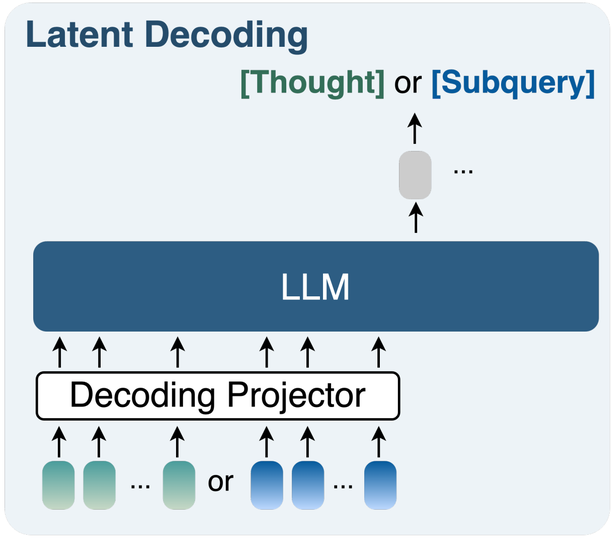


LatentRAG

- Supervised finetuning using trajectories generated by an existing text-based agentic RAG method
- Loss function:

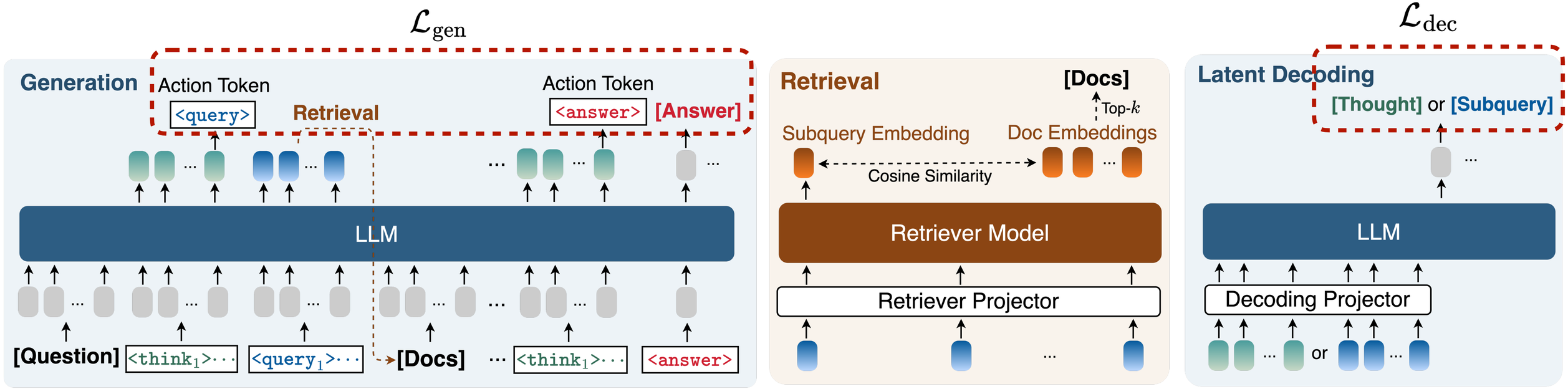






LatentRAG

- Supervised finetuning using trajectories generated by an existing text-based agentic RAG method
- Loss function:
  - For **generation and latent decoding**, we use the standard next token prediction with cross-entropy loss



LatentRAG

- Supervised finetuning using trajectories generated by an existing text-based agentic RAG method
- Loss function:
  - For **generation and latent decoding**, we use the standard next token prediction with cross-entropy loss
  - For **retrieval**, we train the latent subquery to match the retrieval distribution of the reference text subquery via KL divergence

Reference retrieval distribution:

$$p_i(s_t) = \frac{\exp(\cos(\boldsymbol{v}'_{s_t}, \boldsymbol{v}_{d_i})/\beta)}{\sum_{j=1}^{N_d} \exp(\cos(\boldsymbol{v}'_{s_t}, \boldsymbol{v}_{d_j})/\beta)}$$

Text subquery embedding

Document embedding

Latent retrieval distribution:

$$q_i(s_t) = \frac{\exp(\cos(\boldsymbol{v}_{s_t^\ell}, \boldsymbol{v}_{d_i})/\beta)}{\sum_{j=1}^{N_d} \exp(\cos(\boldsymbol{v}_{s_t^\ell}, \boldsymbol{v}_{d_j})/\beta)}$$

Latent subquery embedding

Retrieval loss — KL divergence between both distributions:

$$\mathcal{L}_{\text{ret}} = \frac{1}{\|\mathcal{B}_s\|} \sum_{s_t \in \mathcal{B}_s} \sum_{i=1}^{N_d} p_i(s_t) \log \frac{p_i(s_t)}{q_i(s_t)}$$

LatentRAG

- Supervised finetuning using trajectories generated by an existing text-based agentic RAG method
- Loss function:
  - For **generation and latent decoding**, we use the standard next token prediction with cross-entropy loss
  - For **retrieval**, we train the latent subquery to match the retrieval distribution of the reference text subquery via KL divergence
  - Overall training objective: $\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_{\text{ret}}\mathcal{L}_{\text{ret}} + \mathcal{L}_{\text{dec}}$

Experiments - Setup

- **Datasets**

  We perform experiments on seven public QA datasets based on the Wikipedia corpus (~20 million docs).
  - **General QA datasets:** NQ, TriviaQA, PopQA
  - **Multi-hop QA datasets:** HotpotQA, 2wiki, Musique, Bamboogle

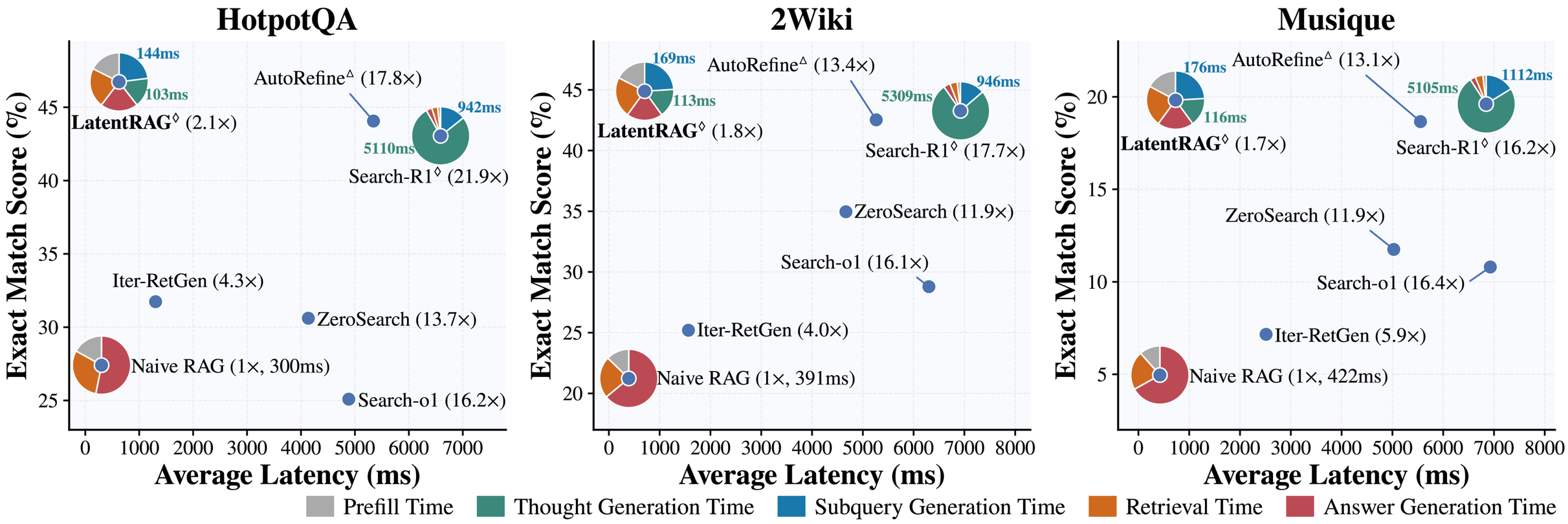
- **Baselines**
  - **Traditional single-step RAG** (NaiveRAG)
  - **Text-based agentic RAG** (Search-R1, AutoRefine, …)

Experiments - Can LatentRAG match text-based agentic RAG performance while achieving significantly lower latency?

Experiments - Can LatentRAG match text-based agentic RAG performance while achieving significantly lower latency?

Comparison of performance and latency on multi-hop QA datasets.

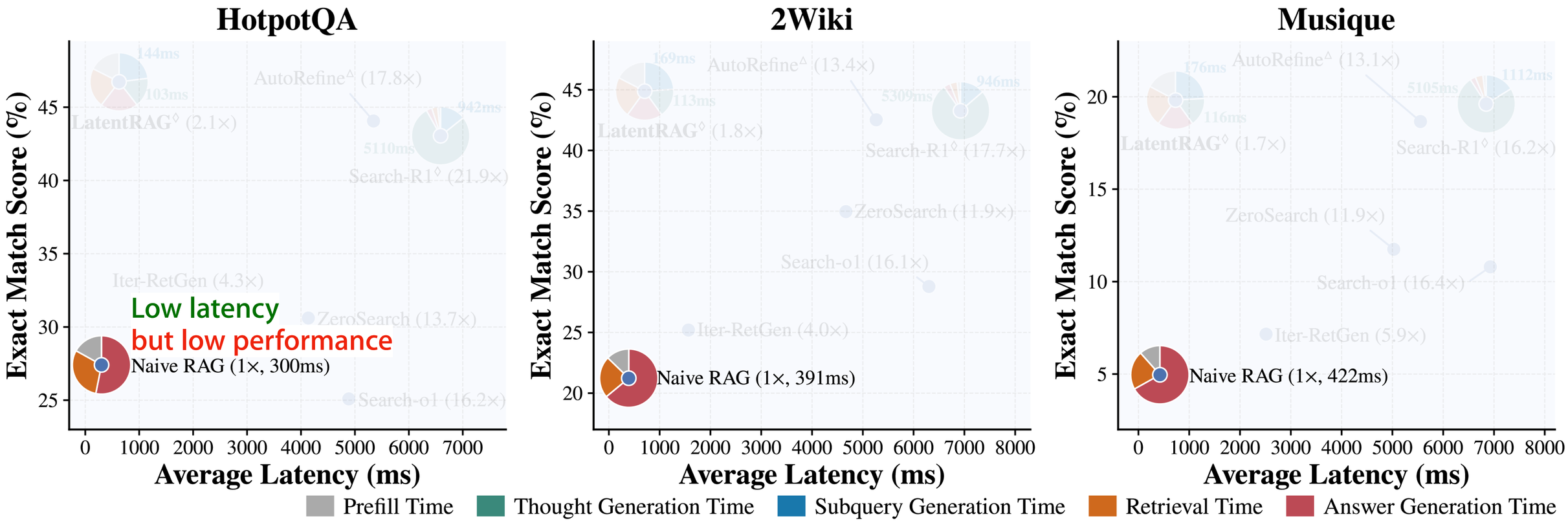
(LLM: Qwen2.5-7B, Retriever: Qwen3-embedding-0.6B)



Experiments - Can LatentRAG match text-based agentic RAG performance while achieving significantly lower latency?

Comparison of performance and latency on multi-hop QA datasets.

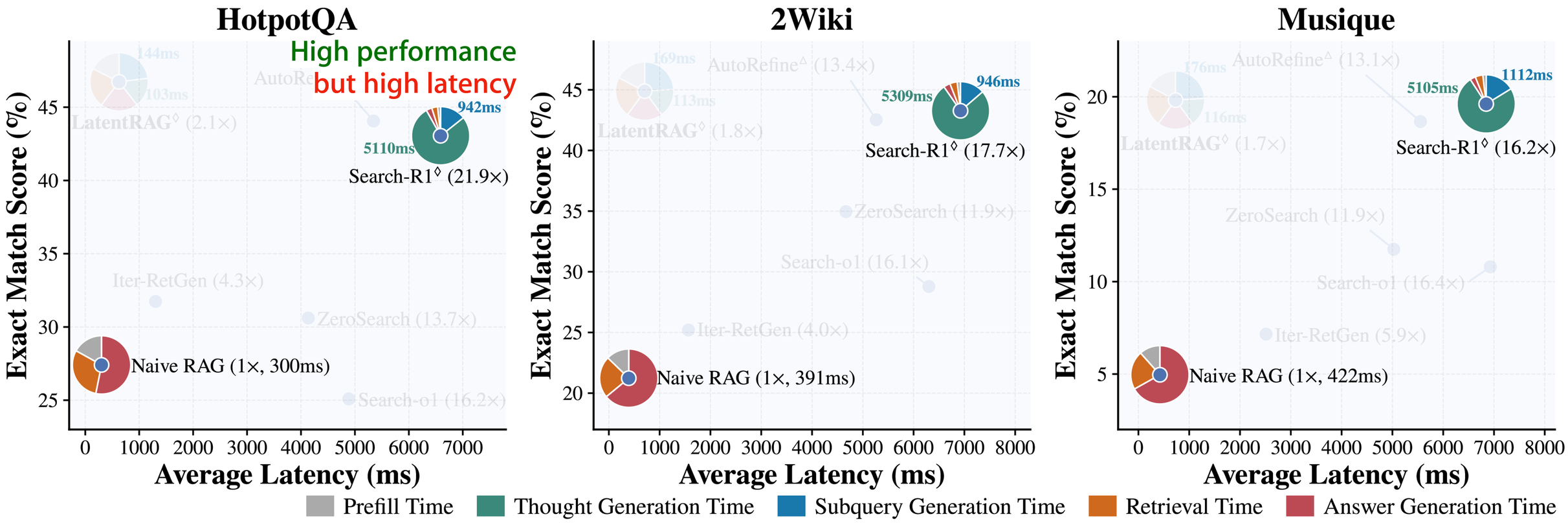
(LLM: Qwen2.5-7B, Retriever: Qwen3-embedding-0.6B)



Experiments - Can LatentRAG match text-based agentic RAG performance while achieving significantly lower latency?

Comparison of performance and latency on multi-hop QA datasets.

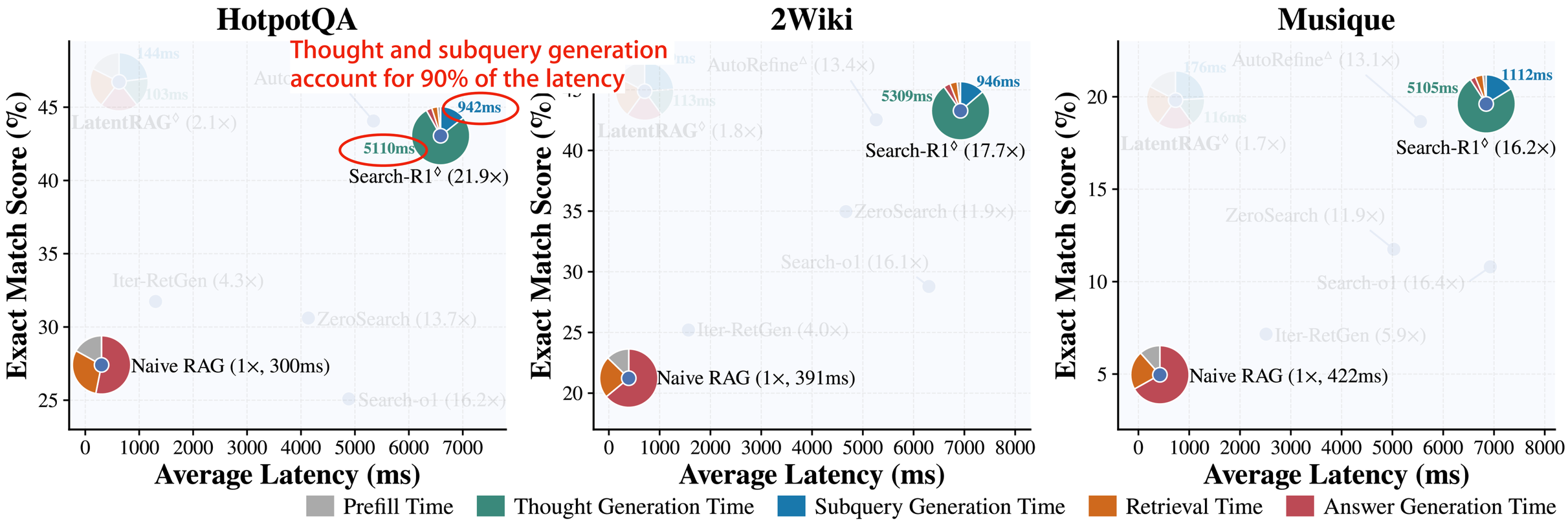
(LLM: Qwen2.5-7B, Retriever: Qwen3-embedding-0.6B)



Experiments - Can LatentRAG match text-based agentic RAG performance while achieving significantly lower latency?

Comparison of performance and latency on multi-hop QA datasets.

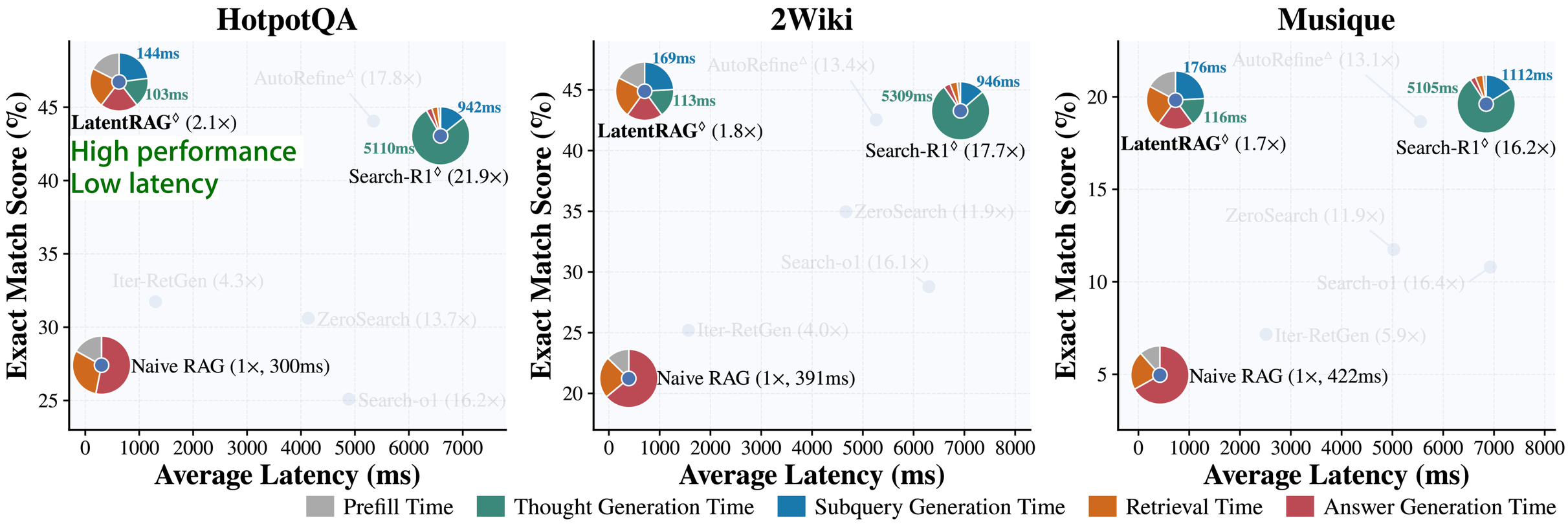
(LLM: Qwen2.5-7B, Retriever: Qwen3-embedding-0.6B)



Experiments - Can LatentRAG match text-based agentic RAG performance while achieving significantly lower latency?

Comparison of performance and latency on multi-hop QA datasets.

(LLM: Qwen2.5-7B, Retriever: Qwen3-embedding-0.6B)

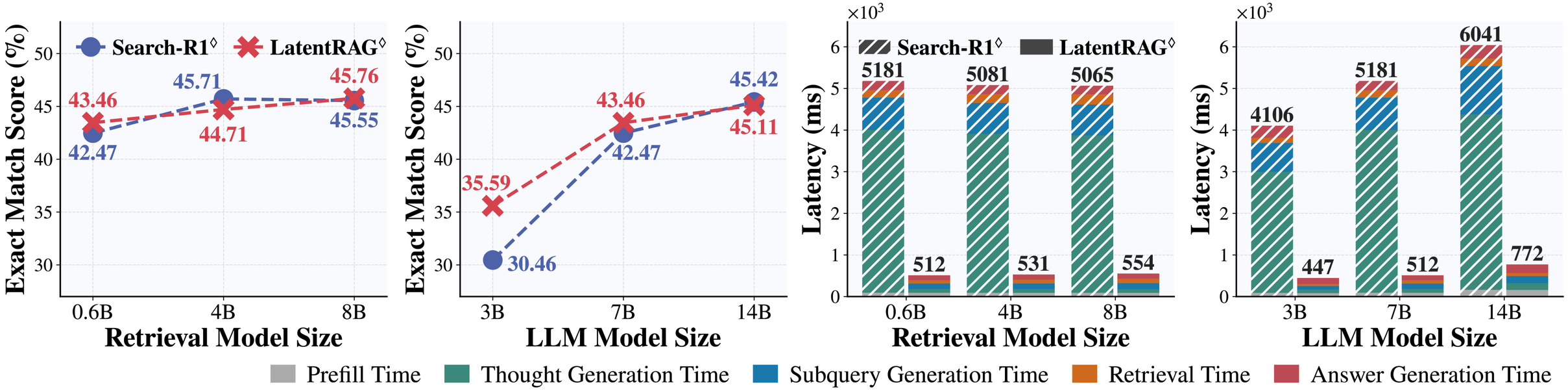


Experiments - How does LatentRAG perform across different model sizes?

Experiments - How does LatentRAG perform across different model sizes?

Results with different model sizes.

(LLM: Qwen2.5-3B/7B/14B, Retriever: Qwen3-embedding-0.6B/4B/8B)

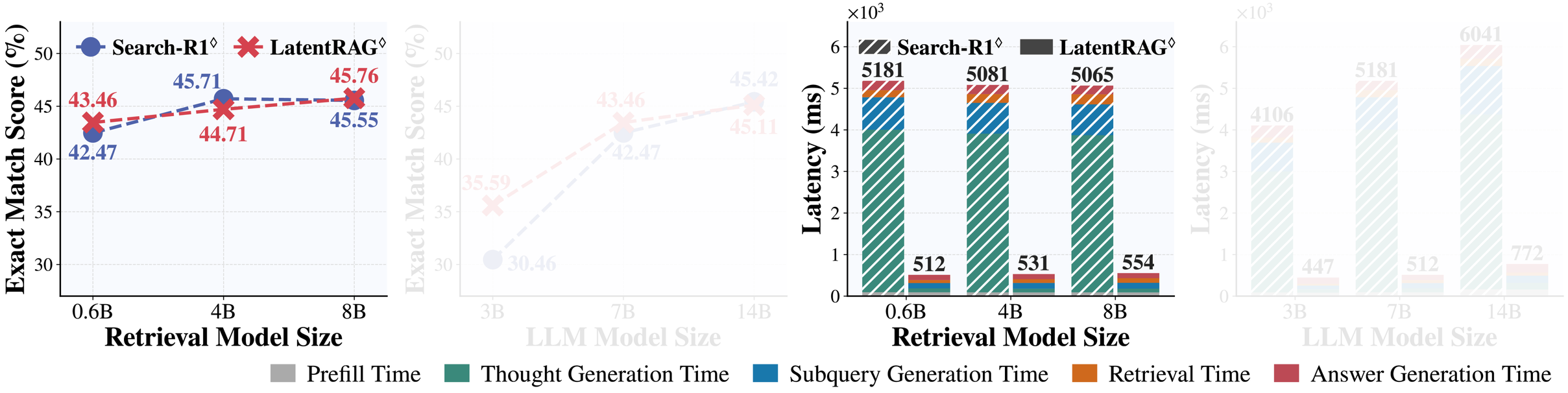


Experiments - How does LatentRAG perform across different model sizes?

Results with different model sizes.

(LLM: Qwen2.5-3B/7B/14B, Retriever: Qwen3-embedding-0.6B/4B/8B)

Increasing retrieval model size has little impact on performance and latency.
LatentRAG matches Search-R1 in performance while significantly reducing latency.

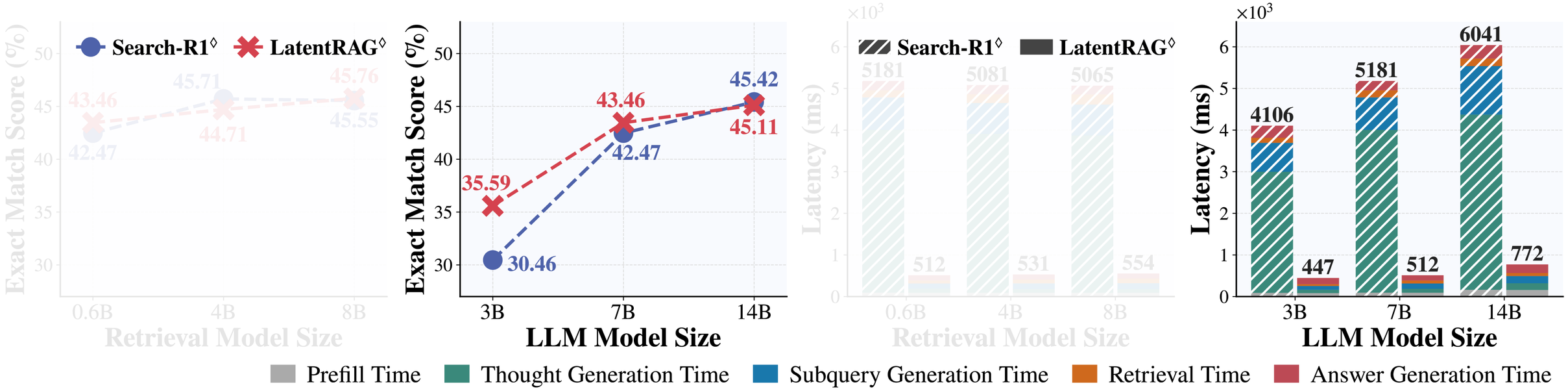


Experiments - How does LatentRAG perform across different model sizes?

Results with different model sizes.

(LLM: Qwen2.5-3B/7B/14B, Retriever: Qwen3-embedding-0.6B/4B/8B)

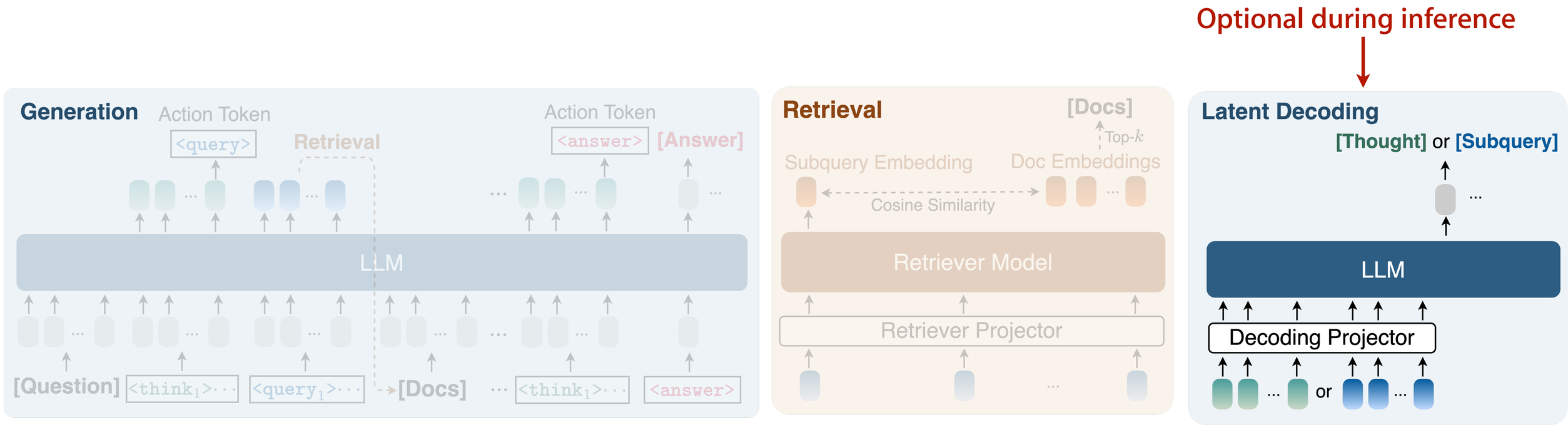
Increasing LLM size improves performance but substantially increases latency.
LatentRAG outperforms Search-R1 on 3B and achieves comparable performance on 7B/14B, while significantly reducing latency.



Experiments - Does LatentRAG with latent decoding still offer a latency advantage?

Experiments - Does LatentRAG with latent decoding still offer a latency advantage?

Latent decoding is an option for **improving transparency** at the cost of **additional latency**.



Experiments - Does LatentRAG with latent decoding still offer a latency advantage?

Latency with and without latent decoding.

Methods	Lat. (ms)↓	Max Len. Ratio(%)
Search-R1◇	5,372	37.74
LatentRAG◇ w/o decoding	593(**-89.0%**)	–
LatentRAG◇ w/ decoding	1,970(**-63.3%**)	33.70
AutoRefine△	4,827	43.14
LatentRAG△ w/o decoding	512(**-89.4%**)	–
LatentRAG△ w/ decoding	2,540(**-47.4%**)	42.13

LatentRAG with latent decoding still reduces latency by 63.3% and 47.4% compared to Search-R1 and AutoRefine, respectively.

Search-R1 and AutoRefine generate textual thoughts and subqueries sequentially at each step, while LatentRAG relies only on latent representations, enabling textural thoughts and subqueries to be decoded in parallel.

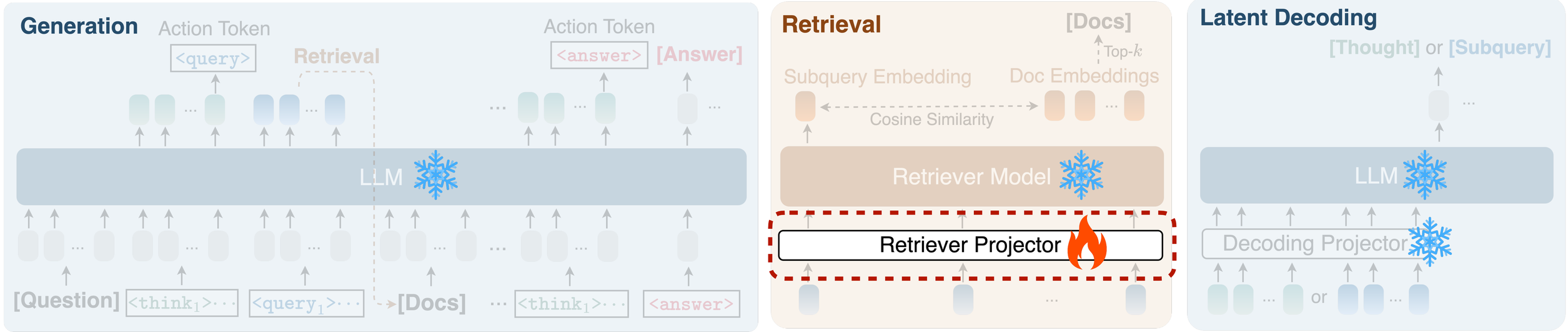
(The latency reduction is related to the max length ratio, which is defined as the ratio between the token count of the longest thought or subquery sequence and the total token count within a decoding batch.)

Experiments - Can a trained LatentRAG model easily adapt to a new retriever?

Experiments - Can a trained LatentRAG model easily adapt to a new retriever?

LatentRAG requires joint finetuning of the LLM and retriever. Adapting a trained LatentRAG model to a new retriever requires additional finetuning.

Can we finetune **only the Retriever Projector** while keeping the LLM and retriever frozen?



Experiments - Can a trained LatentRAG model easily adapt to a new retriever?

Cross-retriever adaptation.

Experimental Setup: We first train LatentRAG with the Qwen3 retriever, then adapt it to the Jina retriever by fine-tuning only the Retriever Projector while keeping the LLM and retriever frozen.

Methods	NQ	TriviaQA	PopQA	HotpotQA	2wiki	Musique	Bamboogle	Average
LatentRAG$^{\diamondsuit}$ (Jina, full data)	47.37	61.25	46.72	47.66	45.06	22.26	43.20	44.79
LatentRAG$^{\diamondsuit}$ (Qwen3→Jina, full data)	47.62	61.25	46.89	47.66	44.51	21.47	41.60	44.43
LatentRAG$^{\diamondsuit}$ (Jina, **5K real data**)	35.12	47.84	34.61	33.99	26.79	10.67	22.40	30.20
LatentRAG$^{\diamondsuit}$ (Qwen3→Jina, **5K real data**)	**46.93**	**58.85**	**45.29**	**46.06**	**43.21**	**19.36**	**40.80**	**42.93**
LatentRAG$^{\diamondsuit}$ (Jina, **5K synthetic data**)	25.18	40.13	30.07	20.65	15.93	3.56	7.20	20.39
LatentRAG$^{\diamondsuit}$ (Qwen3→Jina, **5K synthetic data**)	**45.12**	**57.62**	**44.66**	**43.92**	**43.71**	**17.63**	**39.20**	**41.69**

The adapted model matches the performance of the fully retrained model with the Jina retriever.

Experiments - Can a trained LatentRAG model easily adapt to a new retriever?

Cross-retriever adaptation.

How much data is needed to adapt to a new retriever?

Experiments - Can a trained LatentRAG model easily adapt to a new retriever?

Cross-retriever adaptation.

How much data is needed to adapt to a new retriever?
5K samples are sufficient to recover most of the performance — less than 10% of the full training data.

Methods	NQ	TriviaQA	PopQA	HotpotQA	2wiki	Musique	Bamboogle	Average
LatentRAG$^{\diamondsuit}$ (Jina, full data)	47.37	61.25	46.72	47.66	45.06	22.26	43.20	44.79
LatentRAG$^{\diamondsuit}$ (Qwen3→Jina, full data)	47.62	61.25	46.89	47.66	44.51	21.47	41.60	44.43
LatentRAG$^{\diamondsuit}$ (Jina, **5K real data**)	35.12	47.84	34.61	33.99	26.79	10.67	22.40	30.20
LatentRAG$^{\diamondsuit}$ (Qwen3→Jina, **5K real data**)	**46.93**	**58.85**	**45.29**	**46.06**	**43.21**	**19.36**	**40.80**	**42.93**
LatentRAG$^{\diamondsuit}$ (Jina, **5K synthetic data**)	25.18	40.13	30.07	20.65	15.93	3.56	7.20	20.39
LatentRAG$^{\diamondsuit}$ (Qwen3→Jina, **5K synthetic data**)	**45.12**	**57.62**	**44.66**	**43.92**	**43.71**	**17.63**	**39.20**	**41.69**

Experiments - Can a trained LatentRAG model easily adapt to a new retriever?

Cross-retriever adaptation.

How much data is needed to adapt to a new retriever?
5K synthetic samples are also sufficient to recover most of the performance

Methods	NQ	TriviaQA	PopQA	HotpotQA	2wiki	Musique	Bamboogle	Average
LatentRAG$^{\diamondsuit}$ (Jina, full data)	47.37	61.25	46.72	47.66	45.06	22.26	43.20	44.79
LatentRAG$^{\diamondsuit}$ (Qwen3→Jina, full data)	47.62	61.25	46.89	47.66	44.51	21.47	41.60	44.43
LatentRAG$^{\diamondsuit}$ (Jina, **5K real data**)	35.12	47.84	34.61	33.99	26.79	10.67	22.40	30.20
LatentRAG$^{\diamondsuit}$ (Qwen3→Jina, **5K real data**)	**46.93**	**58.85**	**45.29**	**46.06**	**43.21**	**19.36**	**40.80**	**42.93**
LatentRAG$^{\diamondsuit}$ (Jina, **5K synthetic data**)	25.18	40.13	30.07	20.65	15.93	3.56	7.20	20.39
LatentRAG$^{\diamondsuit}$ (Qwen3→Jina, **5K synthetic data**)	**45.12**	**57.62**	**44.66**	**43.92**	**43.71**	**17.63**	**39.20**	**41.69**

Experiments - Ablation study

- **EM:** Exact match score.
- **Success:** Retrieval success rate, defined as the proportion of samples for which the retrieved documents contain the ground-truth answer.
- **Overlap:** Retrieval overlap, defined as the proportion of retrieved documents that overlap with those retrieved by Search-R1.

Methods	EM(%)$^{\uparrow}$	Success(%)$^{\uparrow}$	Overlap(%)
Search-R1$^{\diamondsuit}$	42.47	60.15	100.00
LatentRAG$^{\diamondsuit}$ (w/ KL loss)	**43.46**	**61.27**	59.41
- (i) Cosine loss	42.55	60.76	68.31
- (ii) InfoNCE loss	41.86	58.60	47.08
- (iii) w/o retriever	41.85	59.07	50.92
- (iv) w/o decoding	40.61	60.64	57.38

Experiments - Ablation study

- **EM:** Exact match score.
- **Success:** Retrieval success rate, defined as the proportion of samples for which the retrieved documents contain the ground-truth answer.
- **Overlap:** Retrieval overlap, defined as the proportion of retrieved documents that overlap with those retrieved by Search-R1.

Methods	EM(%)↑	Success(%)↑	Overlap(%)
Search-R1◇	42.47	60.15	100.00
LatentRAG◇ (w/ KL loss)	**43.46**	**61.27**	59.41
- (i) Cosine loss	42.55	60.76	68.31
- (ii) InfoNCE loss	41.86	58.60	47.08
- (iii) w/o retriever	41.85	59.07	50.92
- (iv) w/o decoding	40.61	60.64	57.38

Ablation1: Choice of retrieval loss (KL v.s. Cosine/InfoNCE)

LatentRAG with KL loss achieves better EM score and retrieval success rate compared to the variants with cosine or InfoNCE alternatives.

The cosine loss yields the highest retrieval overlap ratio, but the performance is lower than that of the KL-based variant, suggesting that overly aligning with the teacher model may limit model capacity and lead to suboptimal performance.

Experiments - Ablation study

- **EM:** Exact match score.
- **Success:** Retrieval success rate, defined as the proportion of samples for which the retrieved documents contain the ground-truth answer.
- **Overlap:** Retrieval overlap, defined as the proportion of retrieved documents that overlap with those retrieved by Search-R1.

Methods	EM(%)↑	Success(%)↑	Overlap(%)
Search-R1◇	42.47	60.15	100.00
LatentRAG◇ (w/ KL loss)	**43.46**	**61.27**	59.41
- (i) Cosine loss	42.55	60.76	68.31
- (ii) InfoNCE loss	41.86	58.60	47.08
- (iii) w/o retriever	41.85	59.07	50.92
- (iv) w/o decoding	40.61	60.64	57.38

Ablation2: Necessity of the Retriever in LatentRAG (Can the LLM directly produce the final subquery embedding?)

Removing the pretrained retrieval model degrades the overall performance and the retrieval performance, highlighting the importance of the inductive bias provided by the pretrained retrieval model.

Experiments - Ablation study

- **EM:** Exact match score.
- **Success:** Retrieval success rate, defined as the proportion of samples for which the retrieved documents contain the ground-truth answer.
- **Overlap:** Retrieval overlap, defined as the proportion of retrieved documents that overlap with those retrieved by Search-R1.

Methods	EM(%)↑	Success(%)↑	Overlap(%)
Search-R1◇	42.47	60.15	100.00
LatentRAG◇ (w/ KL loss)	**43.46**	**61.27**	59.41
- (i) Cosine loss	42.55	60.76	68.31
- (ii) InfoNCE loss	41.86	58.60	47.08
- (iii) w/o retriever	41.85	59.07	50.92
- (iv) w/o decoding	40.61	60.64	57.38

Ablation3: Effectiveness of the latent decoding loss

Latent decoding not only improves transparency, but also encourages the model to learn better semantic latent representations from textual thoughts and subqueries, thereby improving overall performance.

Experiments - Case study

[Question]	What is the place of birth of Clara Novello's father? *(Ground truth answer: London)*

Search-R1$^{\diamond}$:

Step 1

[Thought]	I need to find the place of birth of Clara Novello's father. I'll search for it.
[Subquery]	Clara Novello's father
[Docs]	**Chunk_11245233 [Clara Novello]**: Ivor Novello carried on this family name. Clara Novello Clara Anastasia Novello (10 June 1818 – 12 March 1908) was an acclaimed soprano, the fourth daughter of Vincent Novello, a musician and music publisher, and his wife, Mary Sabilla Hehl. Her acclaimed soprano and pure style made her one of the greatest vocalists, alike in opera, oratorio and on the concert stage, from 1833 onwards. In 1843 she married Count Gigliucci, and retired in 1861. Charles Lamb wrote a poem ("To Clara N.") in her praise. She was born in Oxford Street, London, on 10 June 1818, the fourth daughter ...

[Question]	What is the place of birth of Clara Novello's father? *(Ground truth answer: London)*

LatentRAG$^{\diamond}$:

Step 1

[Thought (dec)]	I need to find the place of birth of Clara Novello's father. I'll search for it.
[Subquery (dec)]	Clara Novello's father
[Docs]	**Chunk_11245233 [Clara Novello]**: Ivor Novello carried on this family name. Clara Novello Clara Anastasia Novello (10 June 1818 – 12 March 1908) was an acclaimed soprano, the fourth daughter of Vincent Novello, a musician and music publisher, and his wife, Mary Sabilla Hehl. Her acclaimed soprano and pure style made her one of the greatest vocalists, alike in opera, oratorio and on the concert stage, from 1833 onwards. In 1843 she married Count Gigliucci, and retired in 1861. Charles Lamb wrote a poem ("To Clara N.") in her praise. She was born in Oxford Street, London, on 10 June 1818, the fourth daughter ...

Experiments - Case study

[Question]	What is the place of birth of Clara Novello's father? *(Ground truth answer: London)*

Search-R1$^{\diamond}$:

Step 2

[Thought]	I found out that Clara Novello's father is Vincent Novello. Now I need to find the place of birth of Vincent Novello.
[Subquery]	place of birth of Vincent Novello
[Docs]	**Chunk_10809160 [Vincent Novello]**: Vincent Novello (6 September 1781 – 9 August 1861), English musician, son of an Italian who married an English wife, was born in London. He was a chorister and organist, but he is best known for bringing to England many works now considered major standards, and with his son he created a major publishing house. As a boy, Novello was a chorister at the Sardinian Embassy Chapel in Duke Street, Lincoln's Inn Fields, where he learnt the organ from Samuel Webbe; and from 1796 to 1822 he became in succession organist of the Sardinian, Spanish (in Manchester Square) and Portuguese ...

[Question]	What is the place of birth of Clara Novello's father? *(Ground truth answer: London)*

LatentRAG$^{\diamond}$:

Step 2

[Thought (dec)]	I found out that Clara Novello's father is Vincent Novello. Now I need to find his place of birth.
[Subquery (dec)]	Vincent Novello place of birth
[Docs]	**Chunk_10809160 [Vincent Novello]**: Vincent Novello (6 September 1781 – 9 August 1861), English musician, son of an Italian who married an English wife, was born in London. He was a chorister and organist, but he is best known for bringing to England many works now considered major standards, and with his son he created a major publishing house. As a boy, Novello was a chorister at the Sardinian Embassy Chapel in Duke Street, Lincoln's Inn Fields, where he learnt the organ from Samuel Webbe; and from 1796 to 1822 he became in succession organist of the Sardinian, Spanish (in Manchester Square) and Portuguese ...

Experiments - Case study

[Question]	What is the place of birth of Clara Novello's father? *(Ground truth answer: London)*

Search-R1$^{\diamond}$:

Step 3

[Thought]	I found out that the place of birth of Vincent Novello is London. Now I can provide the answer.
[Answer]	London

[Question]	What is the place of birth of Clara Novello's father? *(Ground truth answer: London)*

LatentRAG$^{\diamond}$:

Step 3

[Thought (dec)]	I found out that the place of birth of Vincent Novello is London. Now I can provide the answer.
[Answer]	London

Summary

- We propose LatentRAG, an efficient agentic RAG framework that performs reasoning and retrieval in the latent space.
- LatentRAG achieves efficiency comparable to traditional RAG, while the performance is similar to explicit agentic RAG.

Thanks for your attention!

Check our paper for more details!

