NEURAL INFORMATION
PROCESSING SYSTEMS

The Atlas of In-Context Learning: How Attention Heads Shape In-Context Retrieval Augmentation

Patrick Kahardipraja*, Reduan Achtibat*, Thomas Wiegand, Wojciech Samek, Sebastian Lapuschkin

*Equal contribution

Fraunhofer HHI

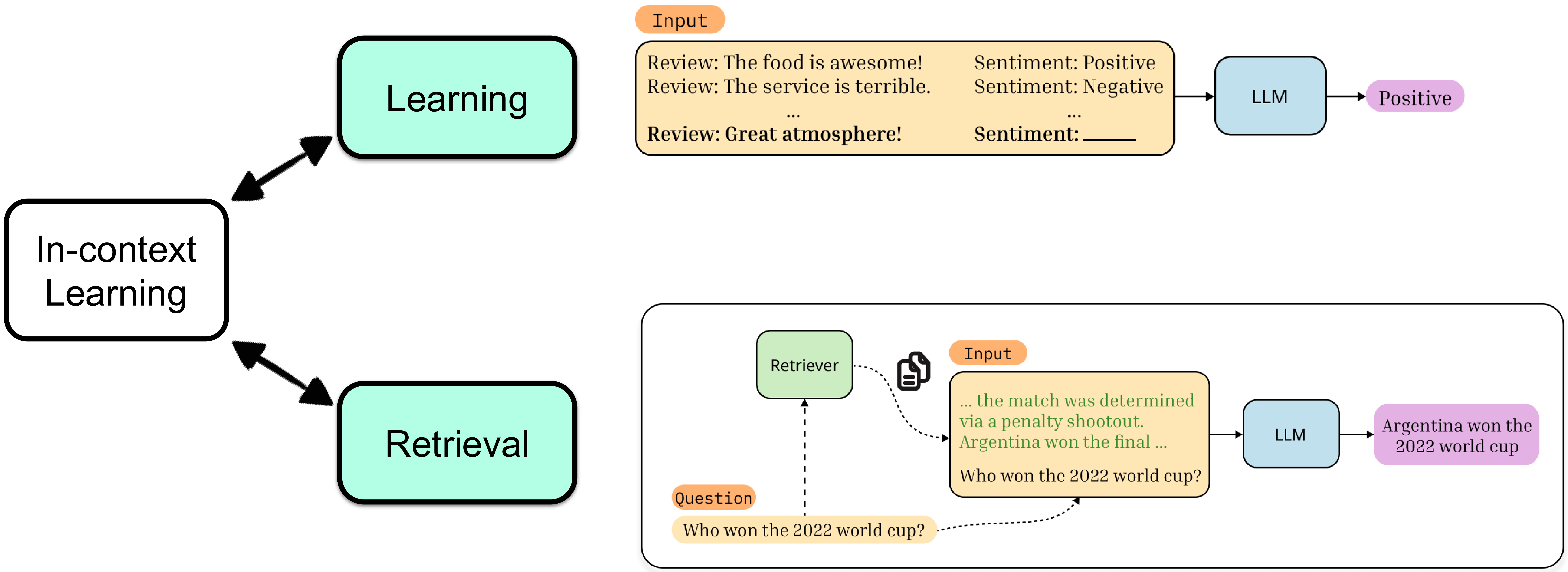
TECHNISCHE UNIVERSITÄT BERLIN

BIFOLD

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

39th Conference on Neural Information Processing Systems (NeurIPS 2025)

In-Context Learning in LLMs



In-context Learning

Learning

Retrieval

Input

Review: The food is awesome! Sentiment: Positive
Review: The service is terrible. Sentiment: Negative
... ...
Review: Great atmosphere! Sentiment: _____

LLM

Positive

Retriever

Input

... the match was determined via a penalty shootout. Argentina won the final ...

Who won the 2022 world cup?

LLM

Argentina won the 2022 world cup

Question

Who won the 2022 world cup?

ICL can be seen as combination of learning and retrieval

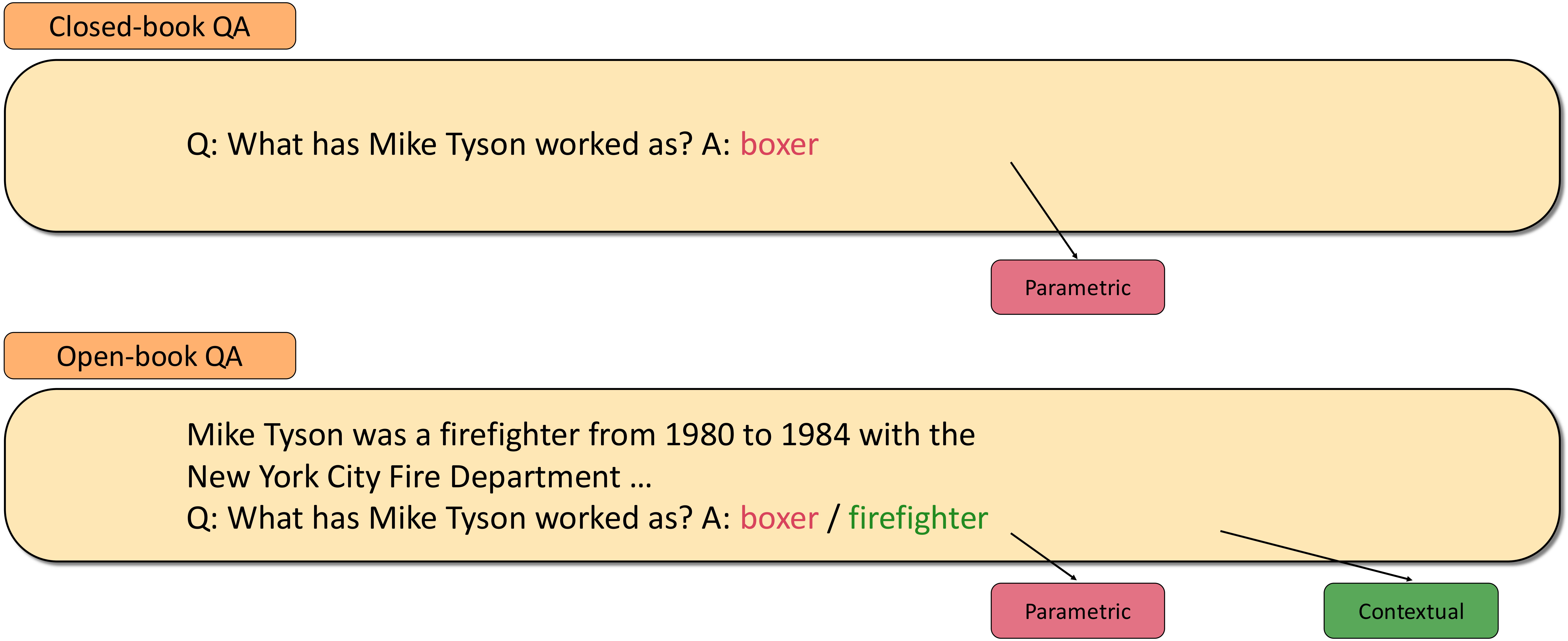
How Does In-Context Retrieval Augmentation Work?

- ICL has been mostly studied from meta-learning perspective, for instance as implicit fine-tuning (Akyürek et al., 2023; Dai et al., 2023)

- However, how ICL works for knowledge retrieval is understudied

- We explore this question **under retrieval augmentation paradigm with focus on QA**

Localization of In-Context and Parametric Heads



Closed-book QA

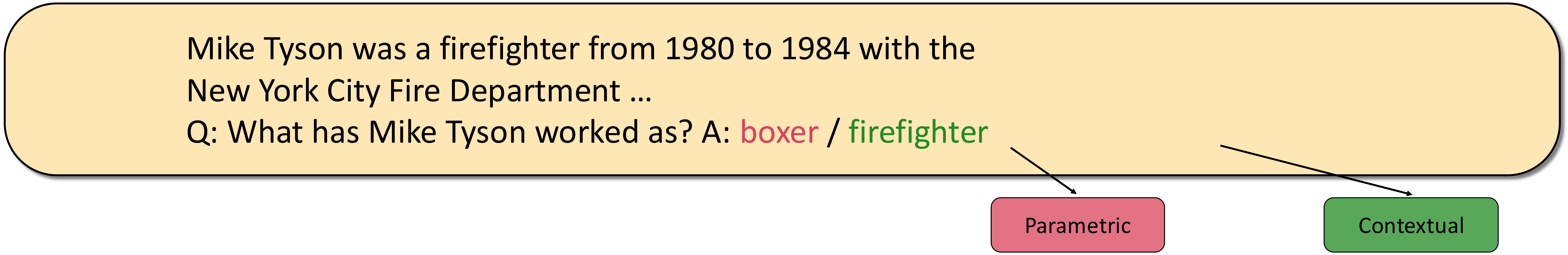
Q: What has Mike Tyson worked as? A: boxer

Open-book QA

Mike Tyson was a firefighter from 1980 to 1984 with the
New York City Fire Department …
Q: What has Mike Tyson worked as? A: boxer / firefighter

Hypothesis: Information in prompts and relational knowledge are processed by different heads

Localization of In-Context and Parametric Heads



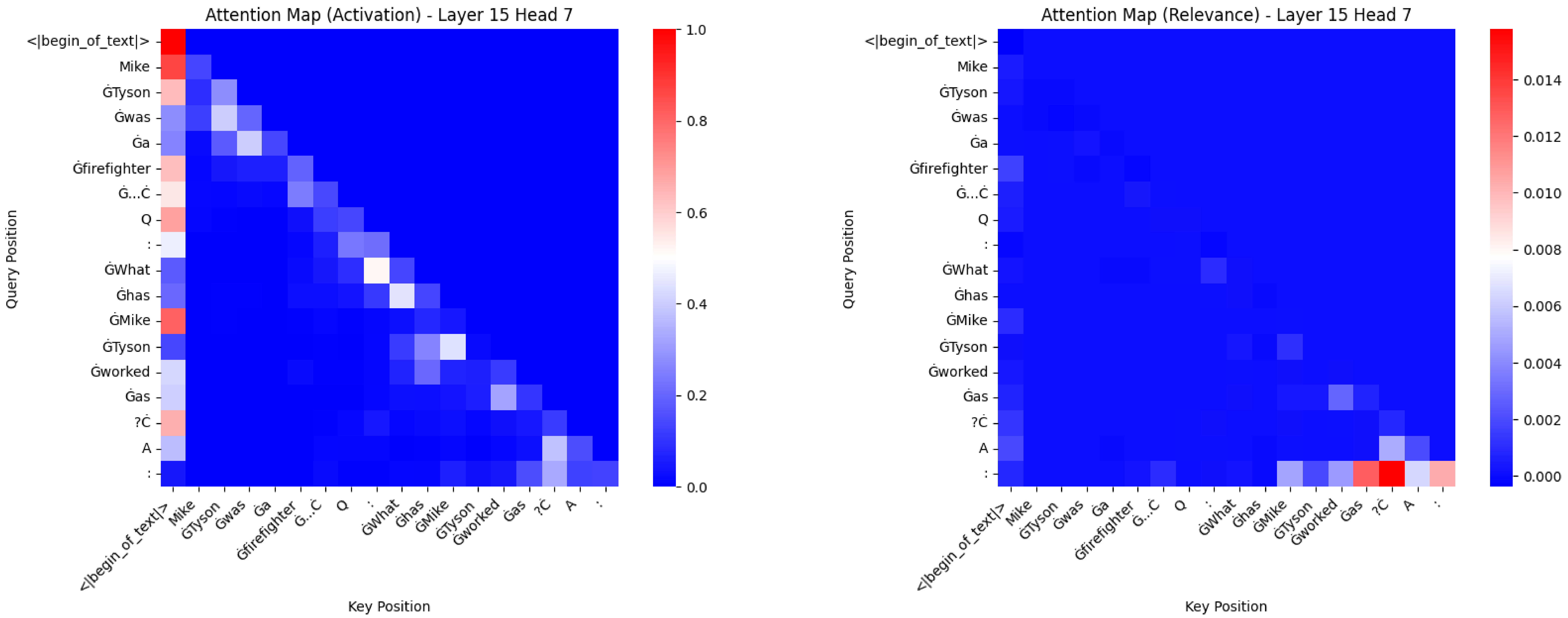
Mike Tyson was a firefighter from 1980 to 1984 with the New York City Fire Department …

Q: What has Mike Tyson worked as? A: boxer / firefighter

- Compute relevance (Bach et al., 2015; Achtibat et al., 2024) of each head for open-book and closed-book QA, determining their activity

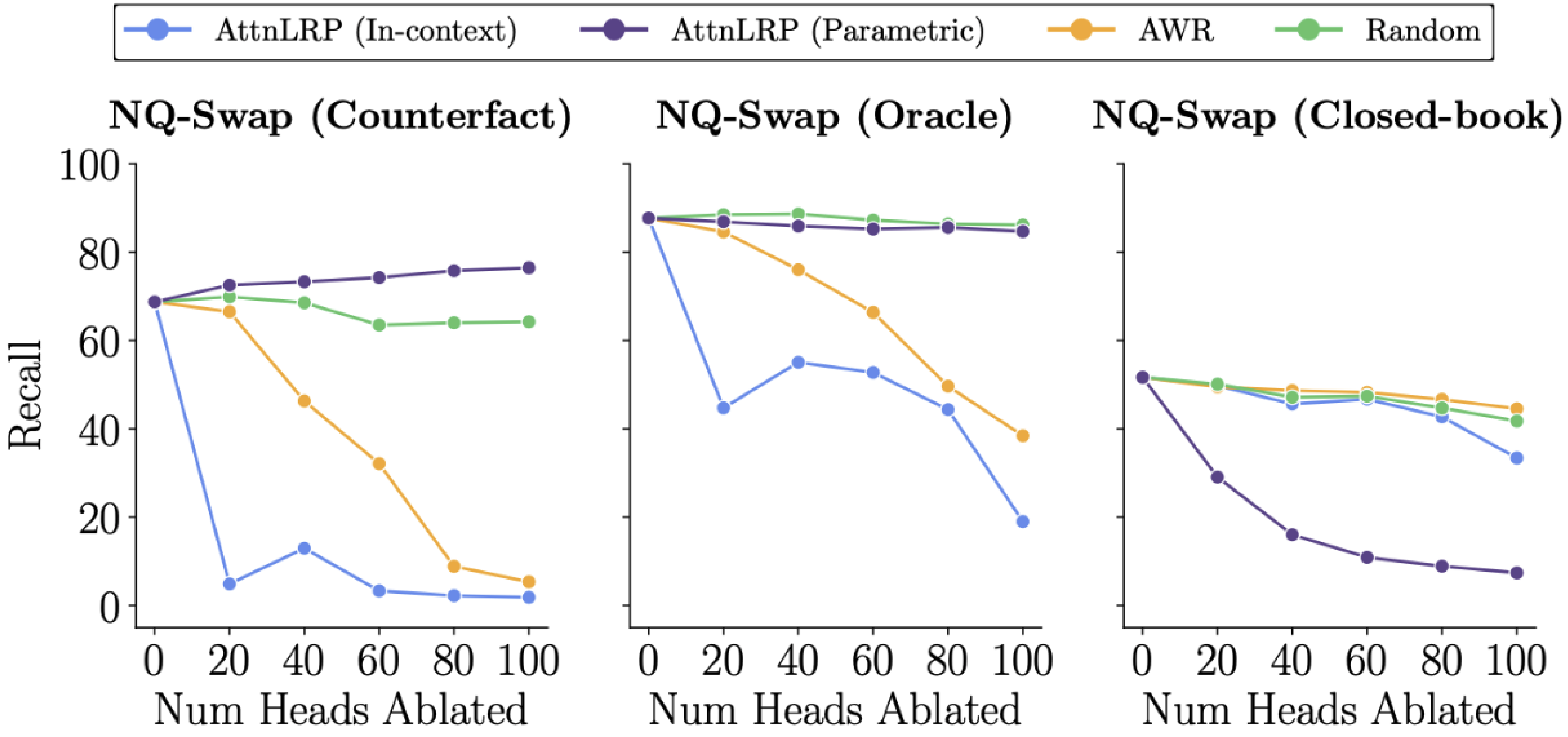
- This allows to determine if each head belongs to in-context or parametric, revealing important heads for in-context retrieval augmentation

Why Relevance Instead of Activation?



Attention activation can be unfaithful (Wiegreffe and Pinter, 2019; Jain and Wallace, 2019), exacerbated by “sinks” (Xiao et al., 2024)

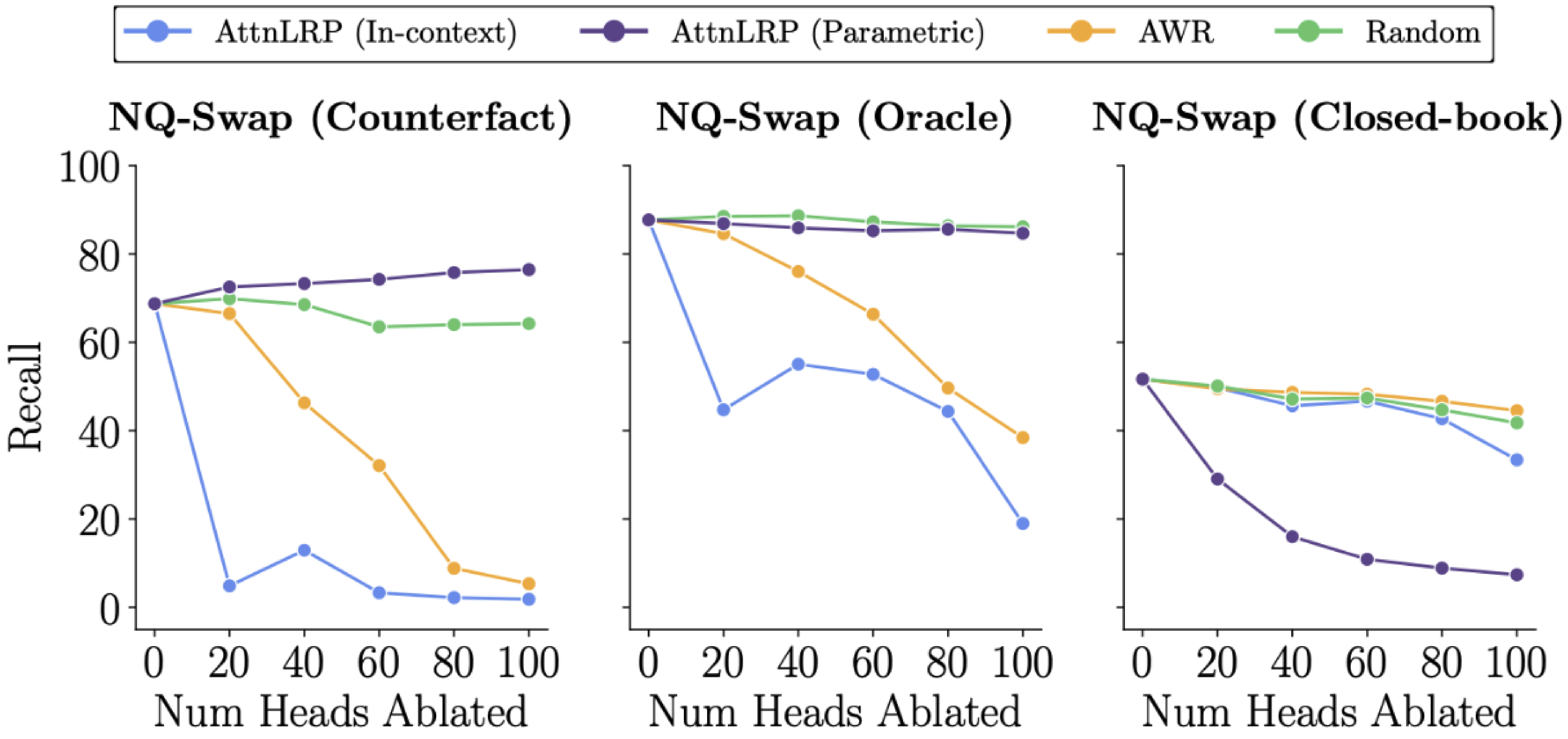
Does Heads Ablation Affect QA Performance? **Yes!**



Removal of in-context heads leads to a more drastic decrease compared to AWR heads

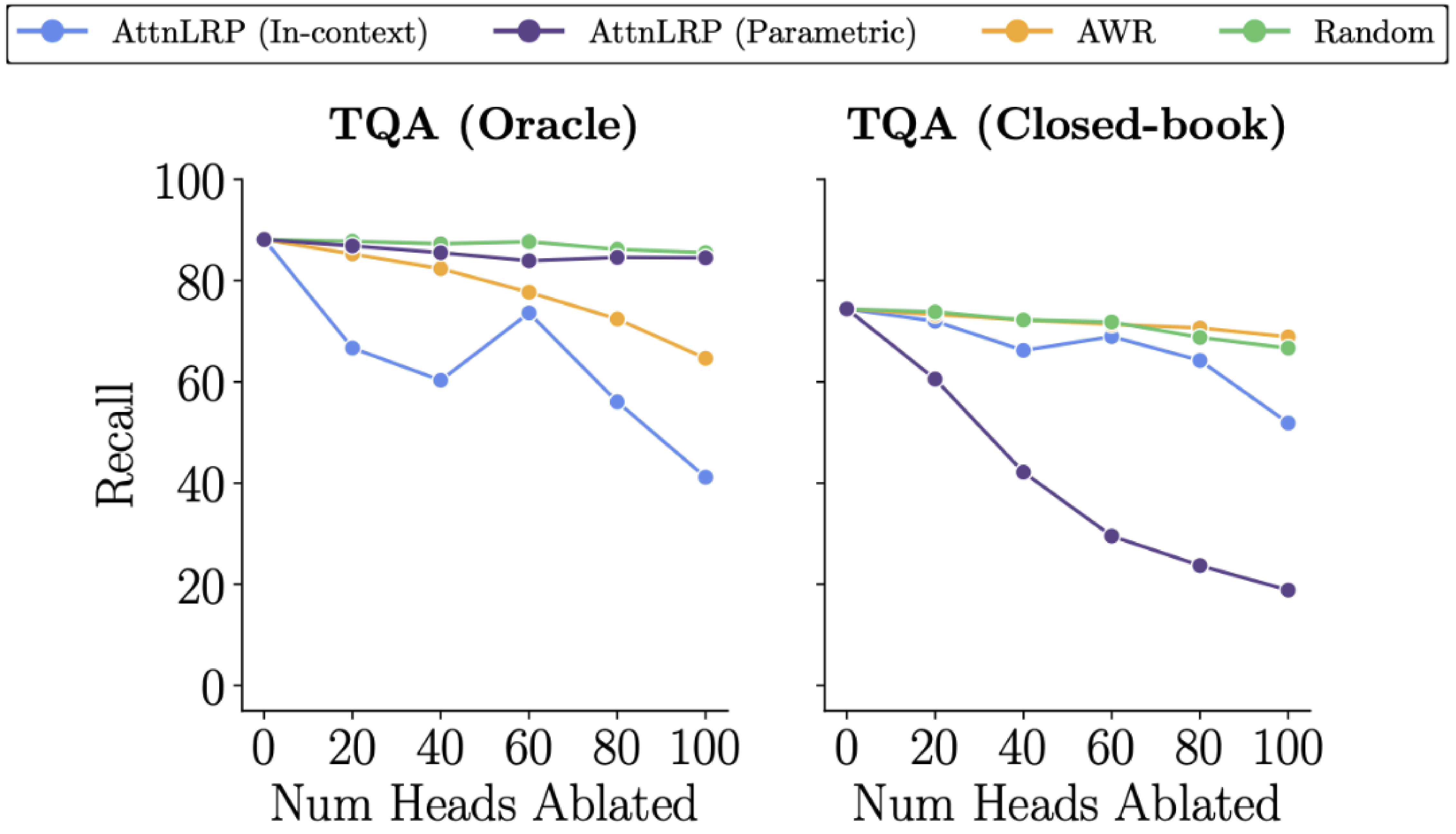
Note: AWR heads are based on attention weights (Wu et al., 2025)

Does Heads Ablation Affect QA Performance? **Yes!**



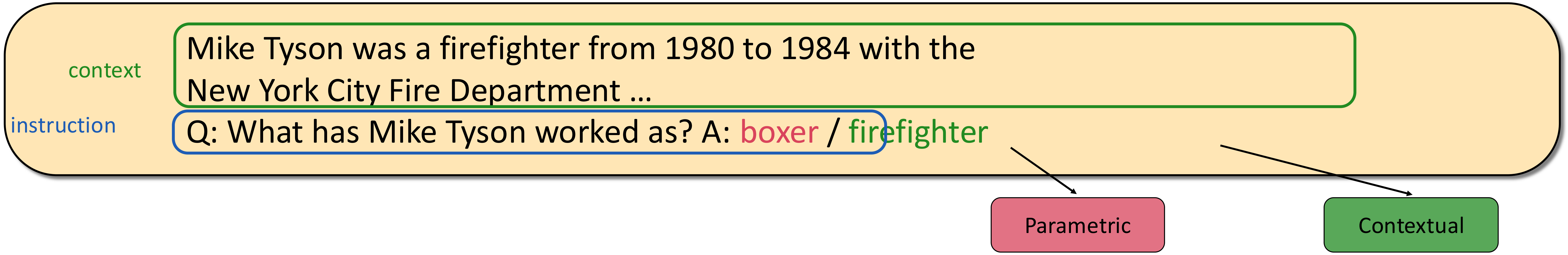
- Ablate parametric heads in open-book → little to no influence
- Ablate in-context heads in closed-book → decrease in performance

Does Their Effect Generalize? **Yes!**



- The identified in-context and parametric heads are transferrable to TriviaQA
- Performance drops still hold!

Are Prompt Components Processed Similarly?



context

Mike Tyson was a firefighter from 1980 to 1984 with the New York City Fire Department …

instruction

Q: What has Mike Tyson worked as? A: boxer / firefighter

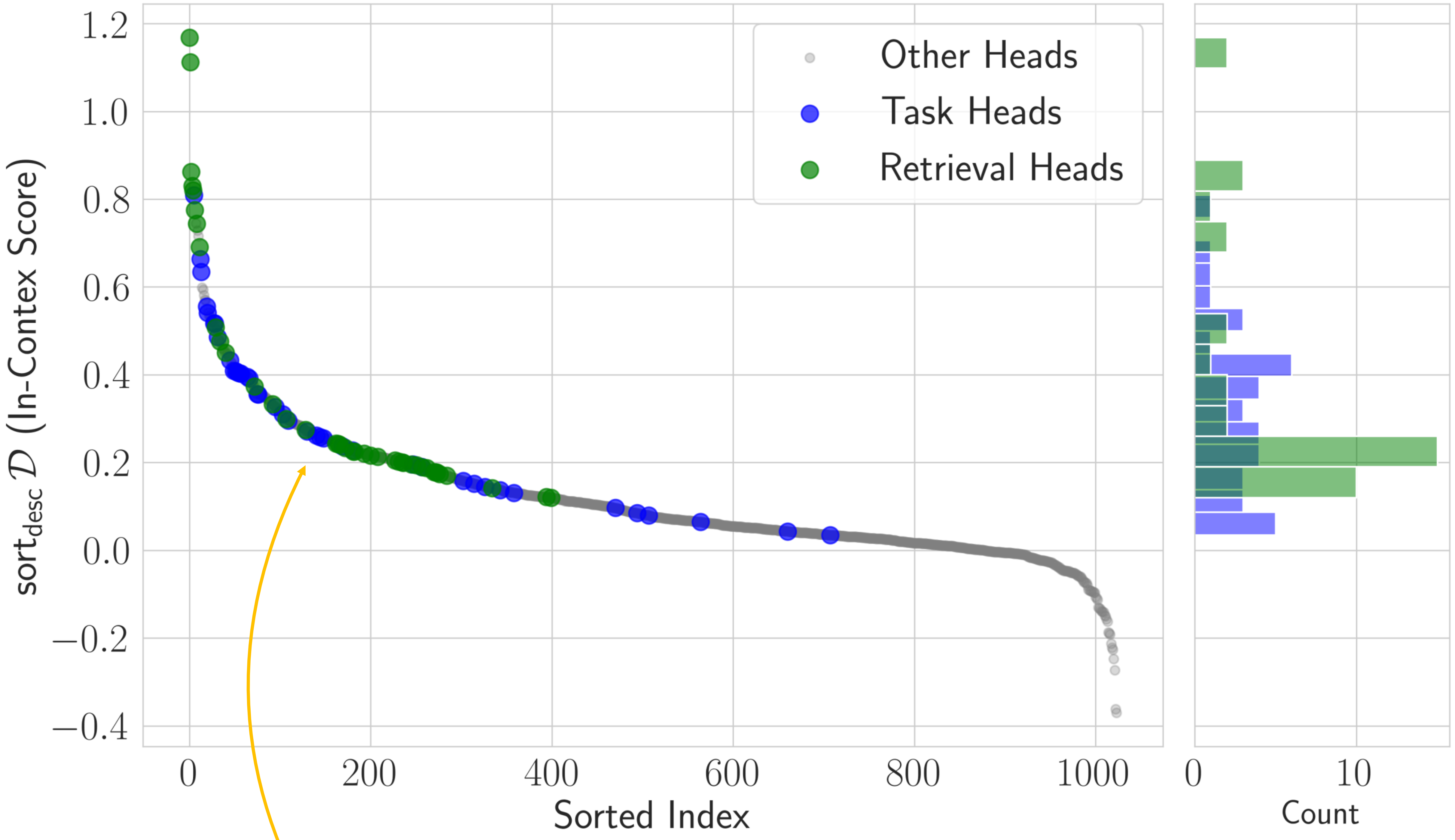
Parametric

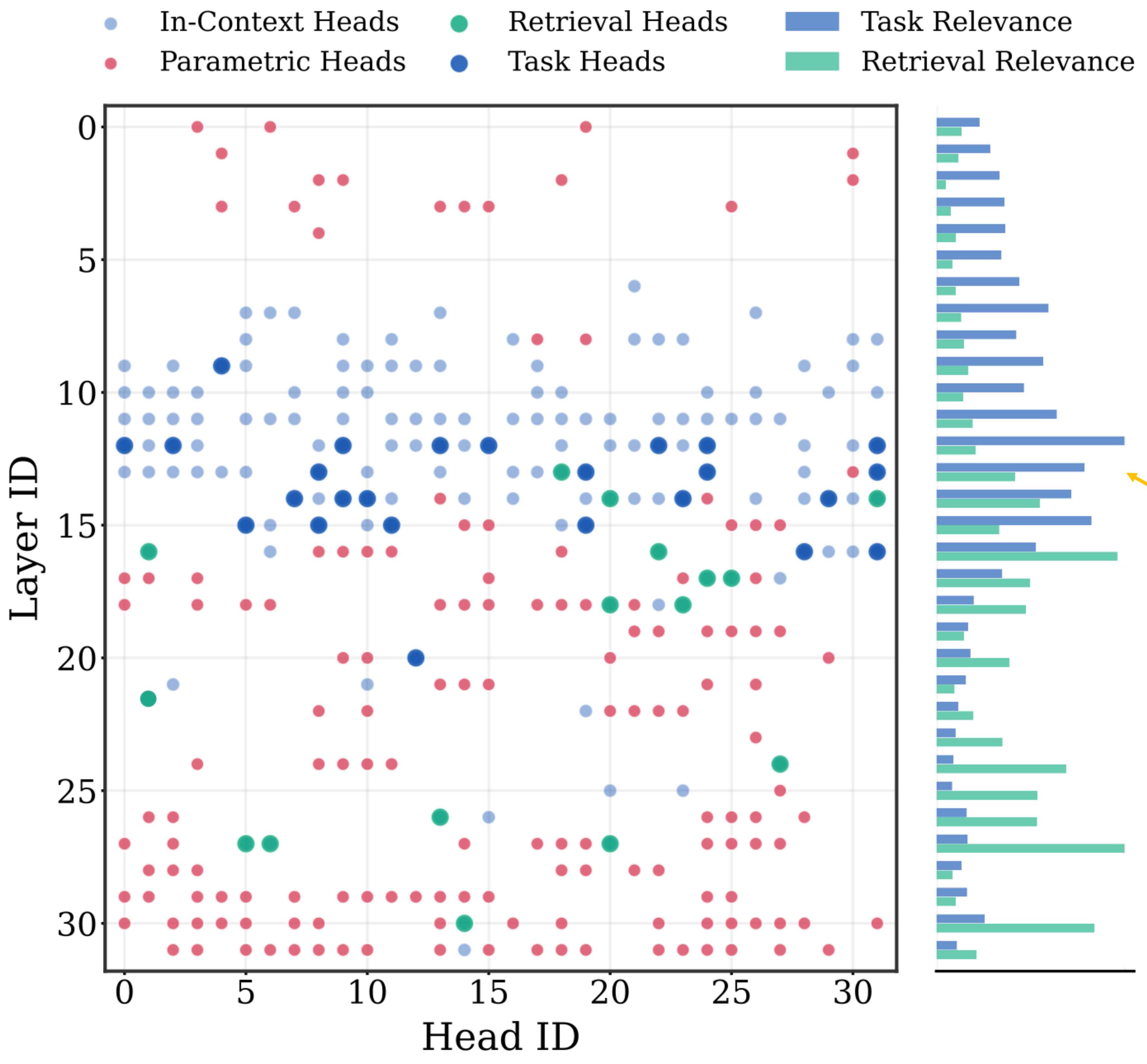
Contextual

Hypothesis: In-context heads process instructions and retrieval of relevant information differently

- Aggregate relevance wrt. question tokens and answer tokens within the context, then rank heads to obtain task and retrieval heads

How Do In-Context Heads Specialize?

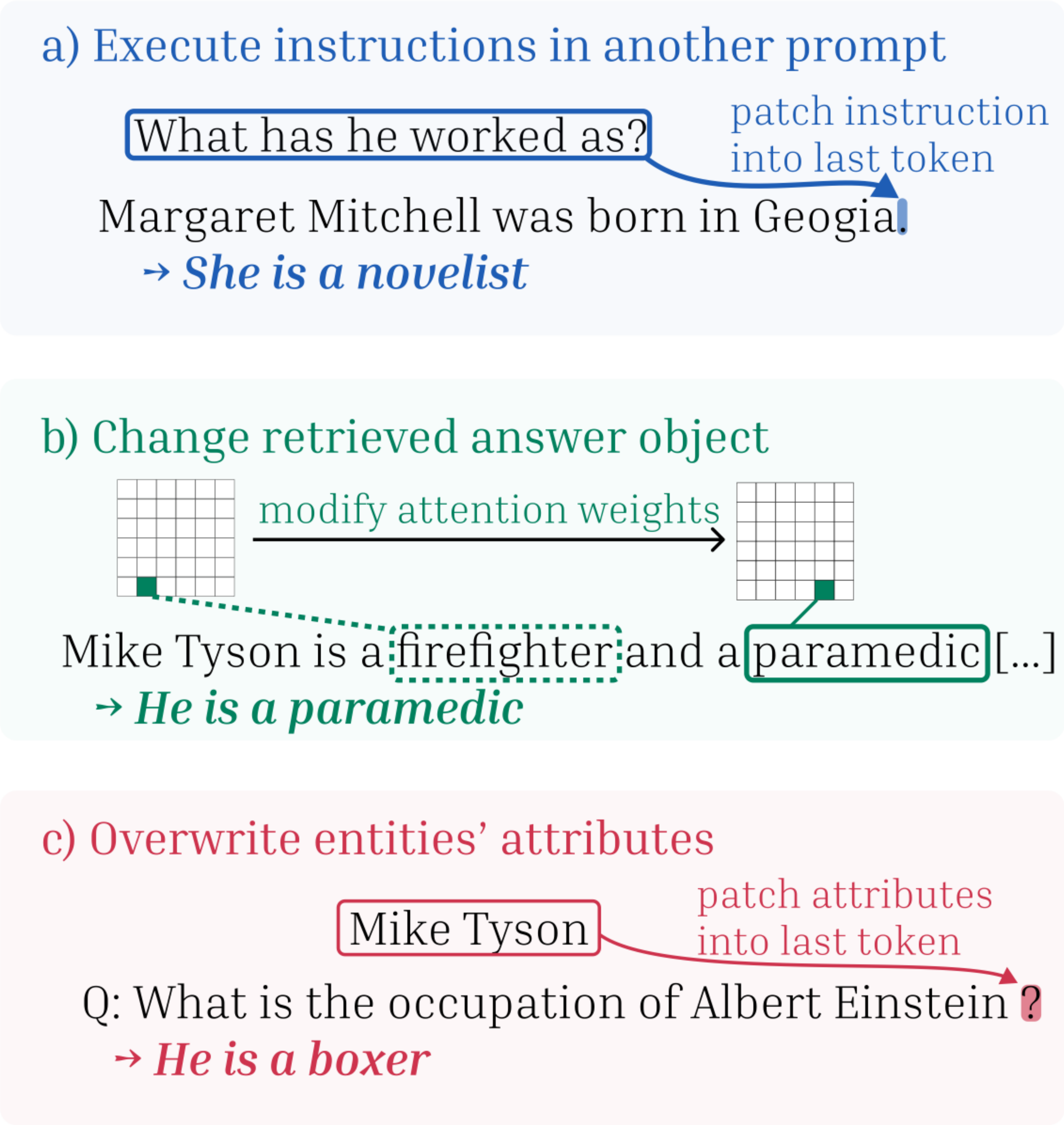




Top in-context heads mainly composed of task and retrieval heads

Instruction-following emerge in the middle, answer retrieval occurs later

Causal Effects of In-Context and Parametric Heads



a) Execute instructions in another prompt

What has he worked as?

patch instruction into last token

Margaret Mitchell was born in Geogia.

→ *She is a novelist*

b) Change retrieved answer object

modify attention weights

Mike Tyson is a firefighter and a paramedic [...]

→ *He is a paramedic*

c) Overwrite entities' attributes

Mike Tyson

patch attributes into last token

Q: What is the occupation of Albert Einstein ?

→ *He is a boxer*

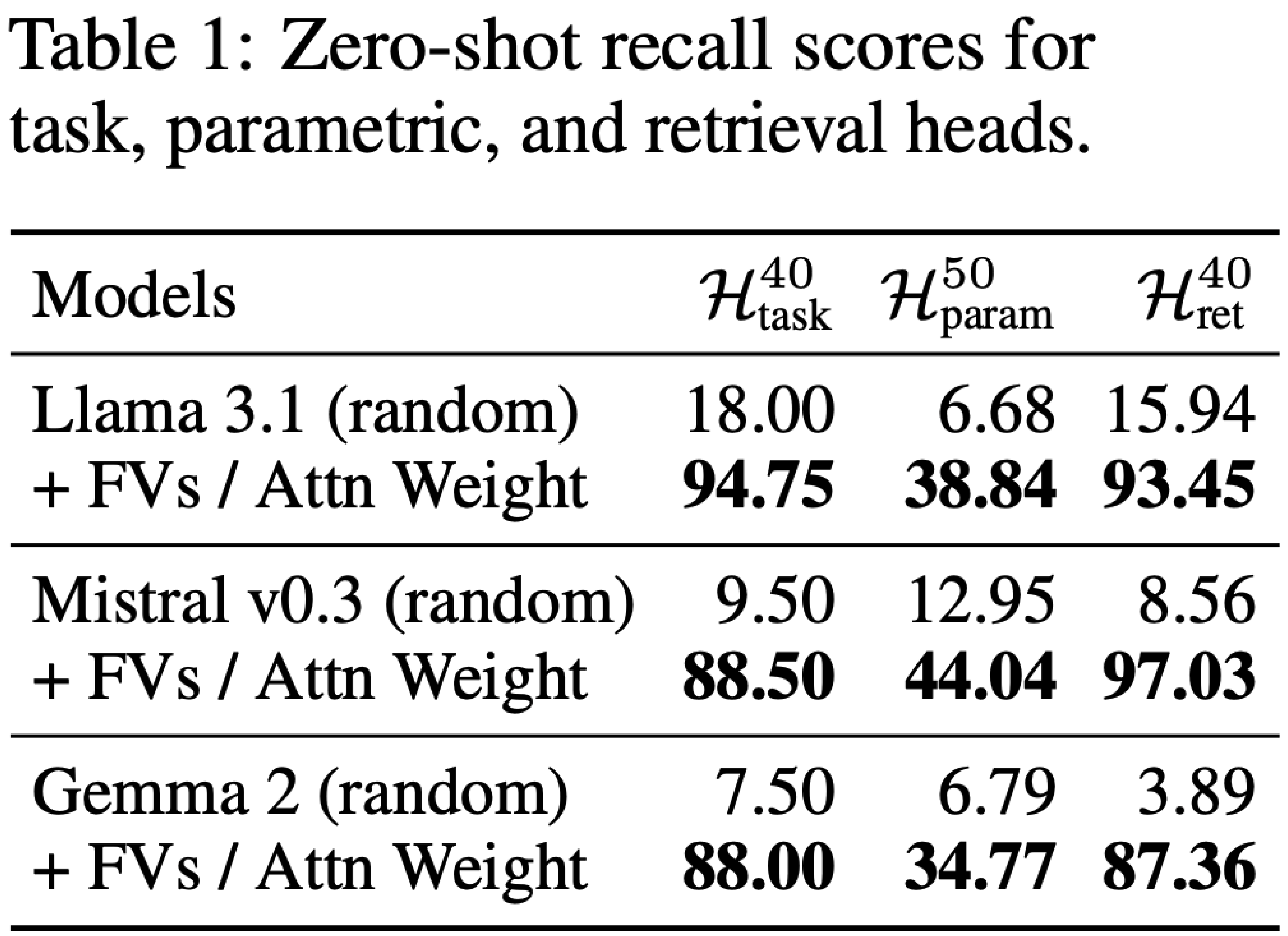


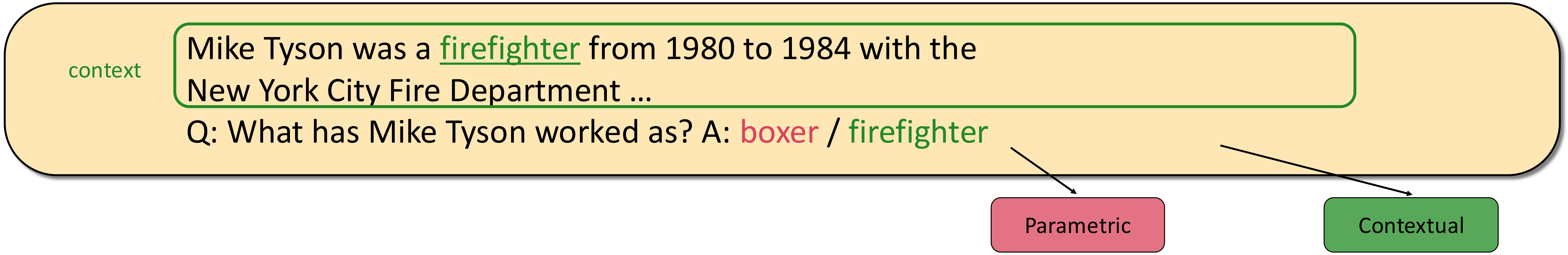
Table 1: Zero-shot recall scores for task, parametric, and retrieval heads.

Models	$\mathcal{H}^{40}_{\text{task}}$	$\mathcal{H}^{50}_{\text{param}}$	$\mathcal{H}^{40}_{\text{ret}}$
Llama 3.1 (random)	18.00	6.68	15.94
+ FVs / Attn Weight	**94.75**	**38.84**	**93.45**
Mistral v0.3 (random)	9.50	12.95	8.56
+ FVs / Attn Weight	**88.50**	**44.04**	**97.03**
Gemma 2 (random)	7.50	6.79	3.89
+ FVs / Attn Weight	**88.00**	**34.77**	**87.36**

Task and parametric heads compress to FVs, retrieval heads change focus wrt. weights

Causal effect on answer generation, inducing specific, targeted functions

Tracking Knowledge Provenance with Identified Heads



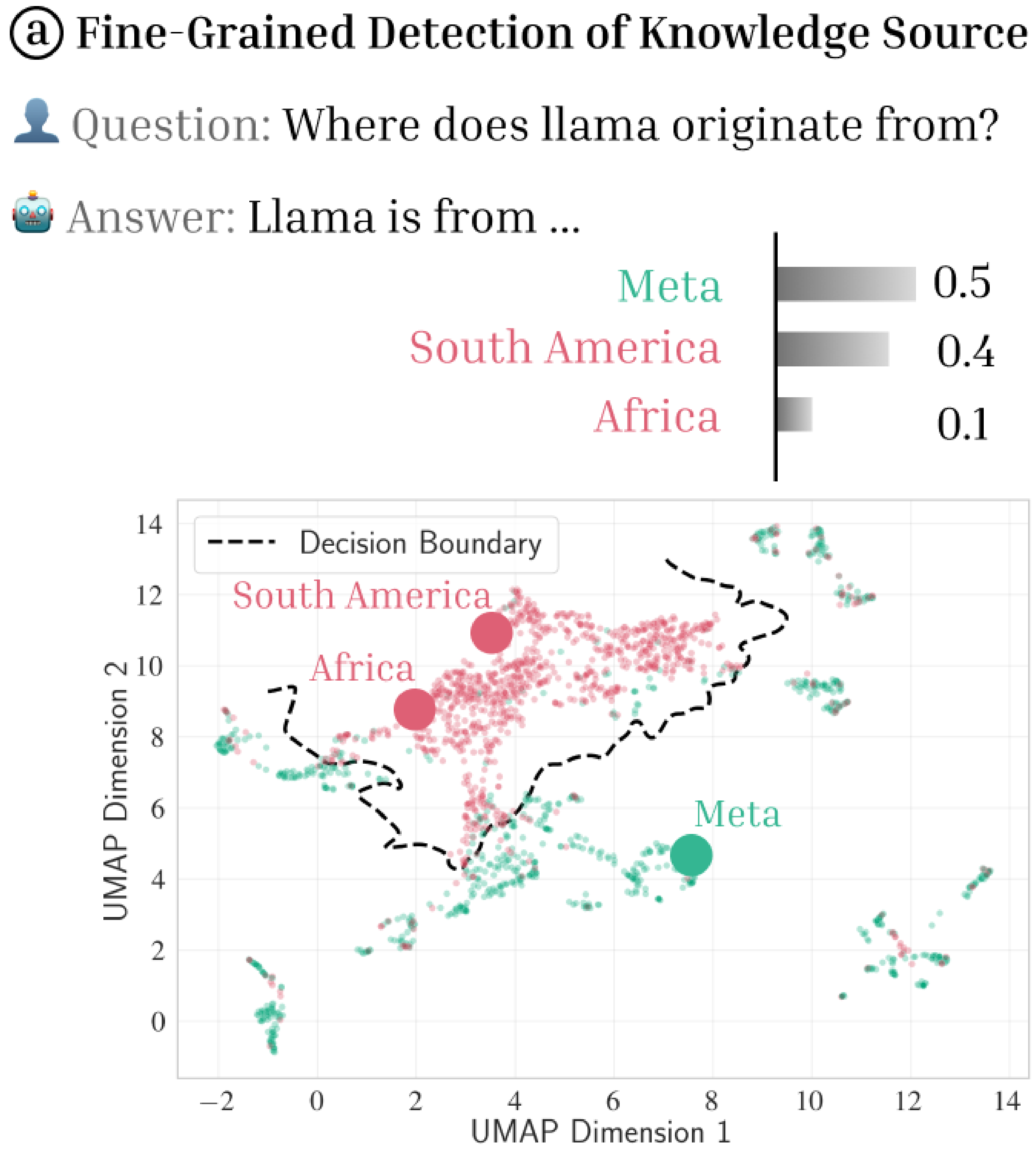
Mike Tyson was a firefighter from 1980 to 1984 with the New York City Fire Department …

Q: What has Mike Tyson worked as? A: boxer / firefighter

Hypothesis: Retrieval heads can be used to track which knowledge is being used

- Train a probe to check if answer comes from parametric knowledge or context
- Aggregate retrieval heads to pinpoint location of answer tokens

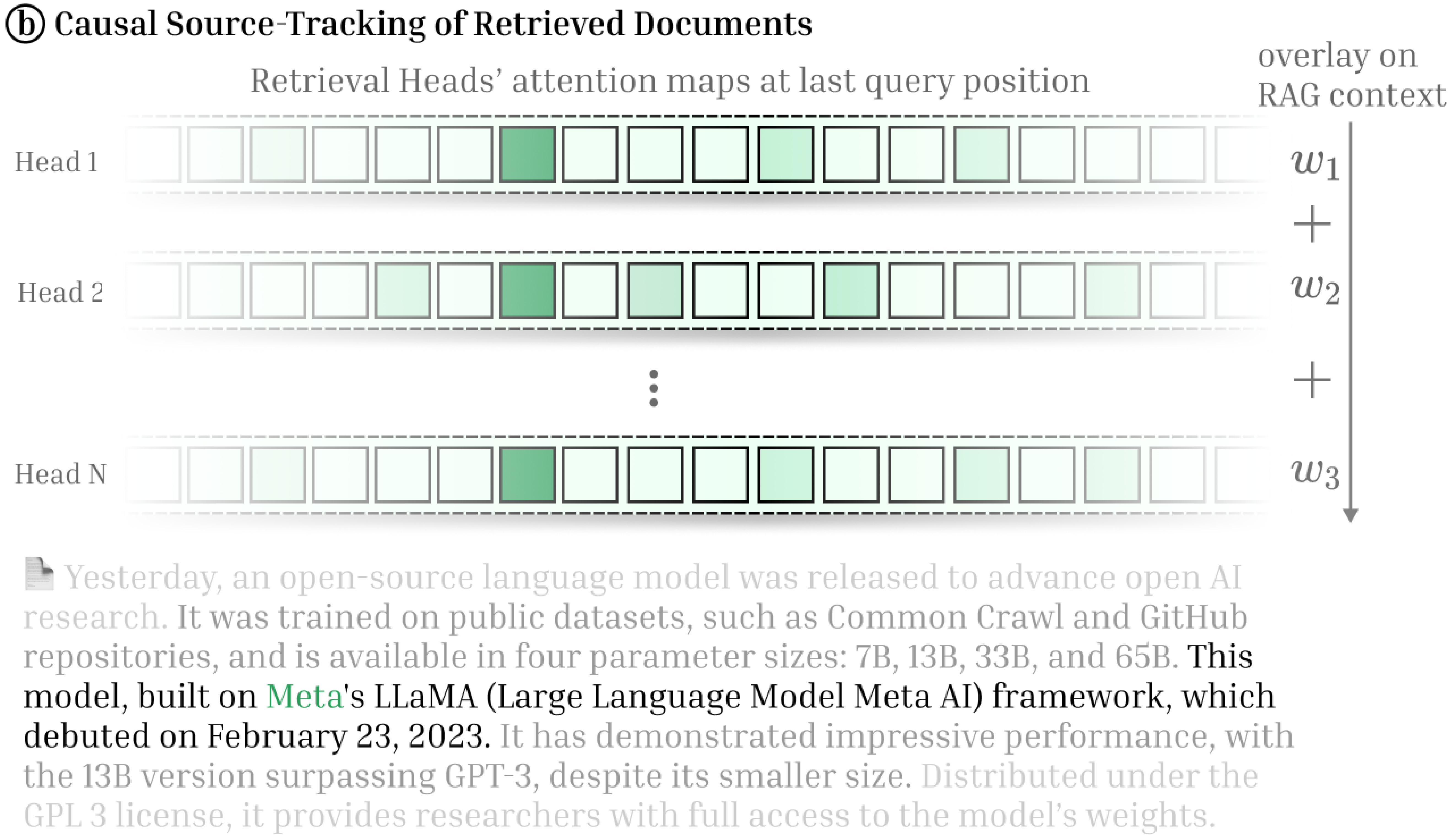
Tracking Knowledge Provenance with Identified Heads



ⓐ Fine-Grained Detection of Knowledge Source

Question: Where does llama originate from?

Answer: Llama is from ...



ⓑ Causal Source-Tracking of Retrieved Documents

Yesterday, an open-source language model was released to advance open AI research. It was trained on public datasets, such as Common Crawl and GitHub repositories, and is available in four parameter sizes: 7B, 13B, 33B, and 65B. This model, built on Meta's LLaMA (Large Language Model Meta AI) framework, which debuted on February 23, 2023. It has demonstrated impressive performance, with the 13B version surpassing GPT-3, despite its smaller size. Distributed under the GPL 3 license, it provides researchers with full access to the model's weights.

UMAP projection shows contextual and parametric answers are separable

Weighted aggregation of retrieval heads able to pinpoint span of retrieved answer

How Retrieval Heads Distinguish and Localize Answer?

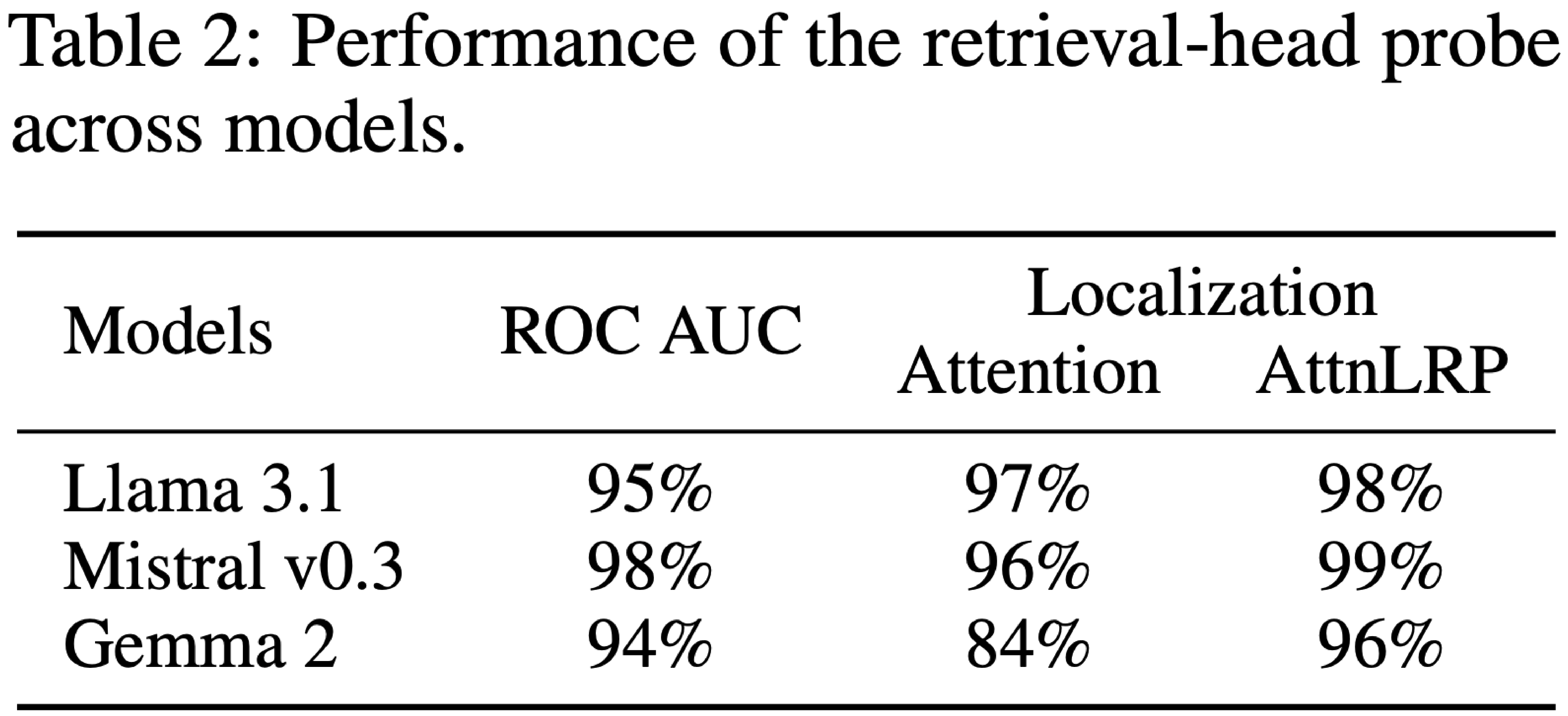


Table 2: Performance of the retrieval-head probe across models.

Models	ROC AUC	Localization	
		Attention	AttnLRP
Llama 3.1	95%	97%	98%
Mistral v0.3	98%	96%	99%
Gemma 2	94%	84%	96%

- Retrieval heads reliably distinguish contextual from parametric predictions
- They can also be used to accurately localize answer tokens

Takeaways

- In-context and parametric heads operate on input prompt distinctly for retrieval-augmented QA

- In-context heads may specialize into task or retrieval heads, and shape model's representations along with parametric heads

- Retrieval heads can efficiently track for knowledge provenance

Paper: https://arxiv.org/abs/2505.15807