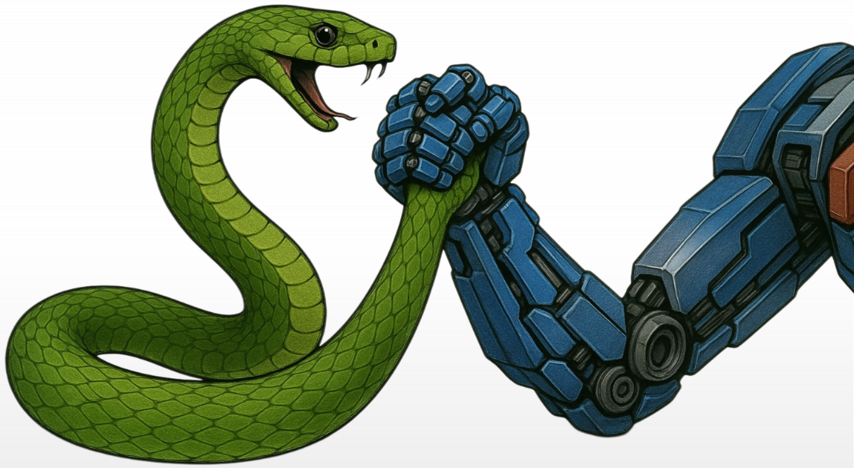
# On the Transformer-SSM Gap

And the Role of the Gather-and-Aggregate Mechanism



Aviv Bick

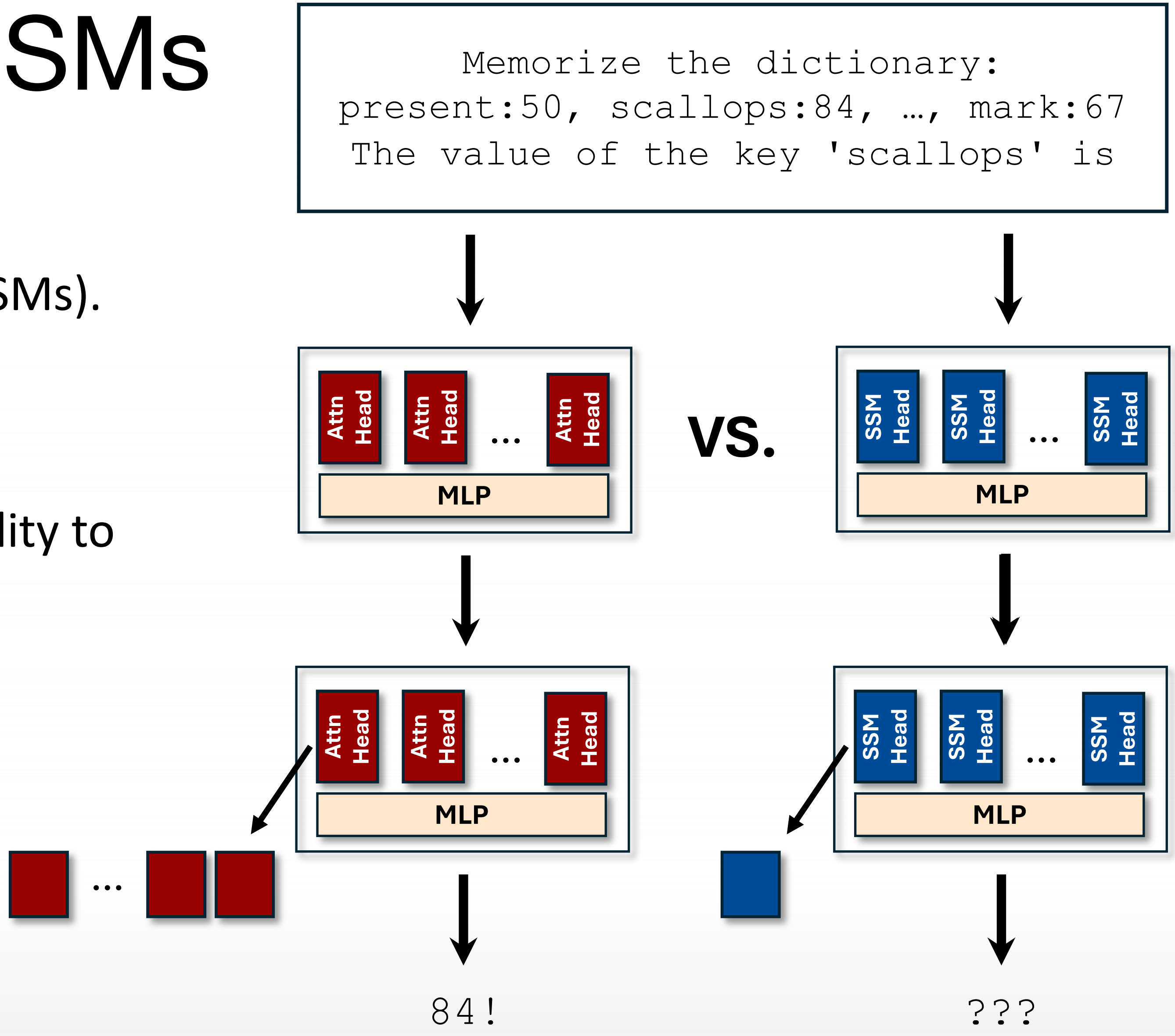
Carnegie Mellon University

# Transformers vs. SSMs

There is a performance gap between Transformers and State-Space Models (SSMs).

- Mathematical reasoning, coding, etc.

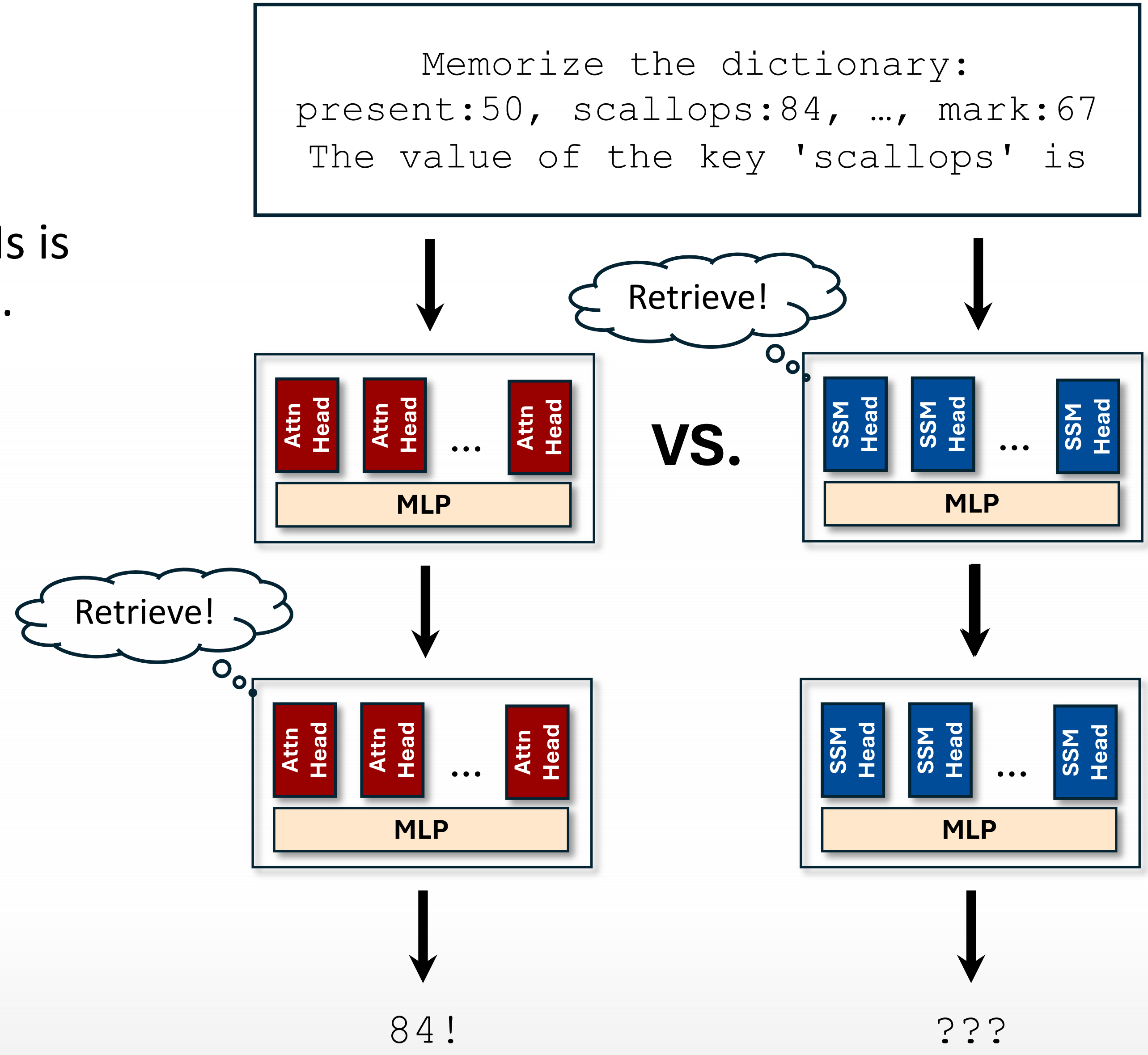
This gap has been linked to a model's ability to do **in-context retrieval** [Arora et al.]




Arora et al., "Zoology: Measuring and Improving Recall in Efficient Language Models"

# Outline

1. Retrieval in both Transformers and SSMs is performed similarly, in just a few heads.

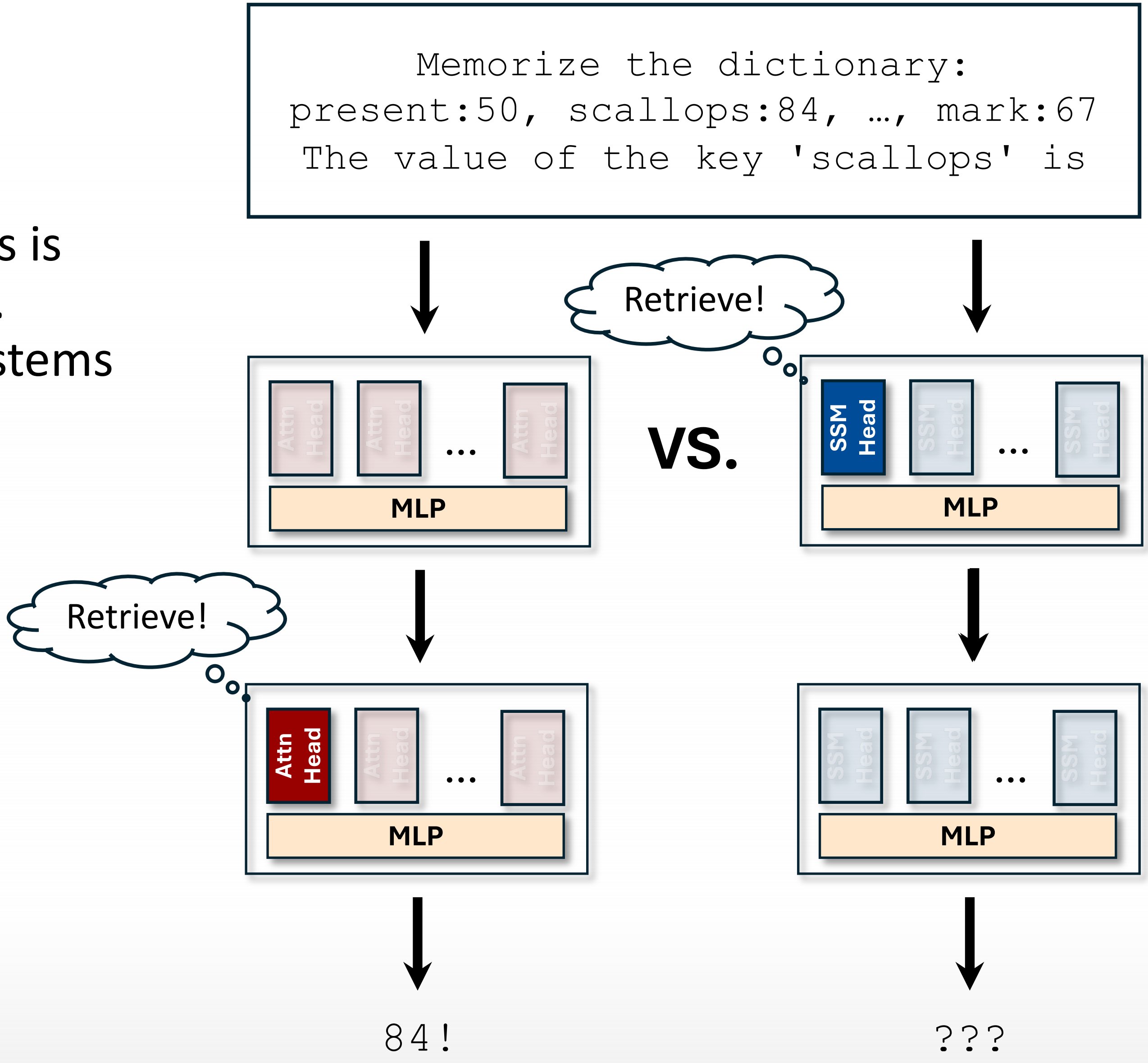


# Outline

1. Retrieval in both Transformers and SSMs is performed similarly, in just a few heads.
   ⟹ Transformer-SSM performance gap stems from these heads

2. SSMs approximate these heads weakly

3. Hybrid models close the gap!



# Case Study: MMLU Benchmark

MMLU requires extensive **knowledge** across 57 different fields.

SSMs have the knowledge but struggle with MMLU [Waleffe et al.]

How is MMLU different from other benchmarks? It’s in the format

```
___ is the central node of 802.11 wireless operations.
A.  WPA
B.  Access Point
C.  WAP
D.  Access Port
Answer:
```

Waleffe et al., “An Empirical Study of Mamba-based Language Models”

# Case Study: MMLU Benchmark

MMLU requires extensive **knowledge** across 57 different fields.

SSMs have the knowledge but struggle with MMLU [Waleffe et al.]

How is MMLU different from other benchmarks? It's in the format

```
___ is the central node of
802.11 wireless operations.
A.  WPA
B.  Access Point
C.  WAP
D.  Access Port
Answer: WPA
```

Classic format

vs.

```
___ is the central node of
802.11 wireless operations.
A.  WPA
B.  Access Point
C.  WAP
D.  Access Port
Answer: B
```

MMLU format

Waleffe et al., "An Empirical Study of Mamba-based Language Models"

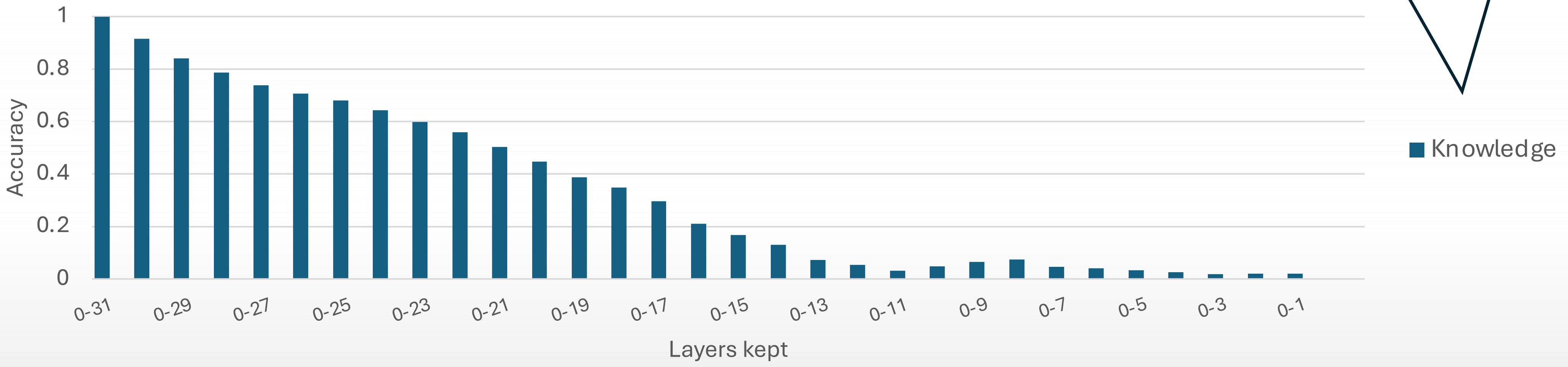
# Case Study: MMLU Benchmark

**Gradual Pruning.** Prune layers from the end of Llama-3.1-8B

After each prune, we measure how much knowledge is retained

- Knowledge extraction is distributed

<u>Minimal Retrieval Tasks</u>
ARC-Challenge, ARC-Easy, PIQA, Winogrande, OpenBookQA, HellaSwag



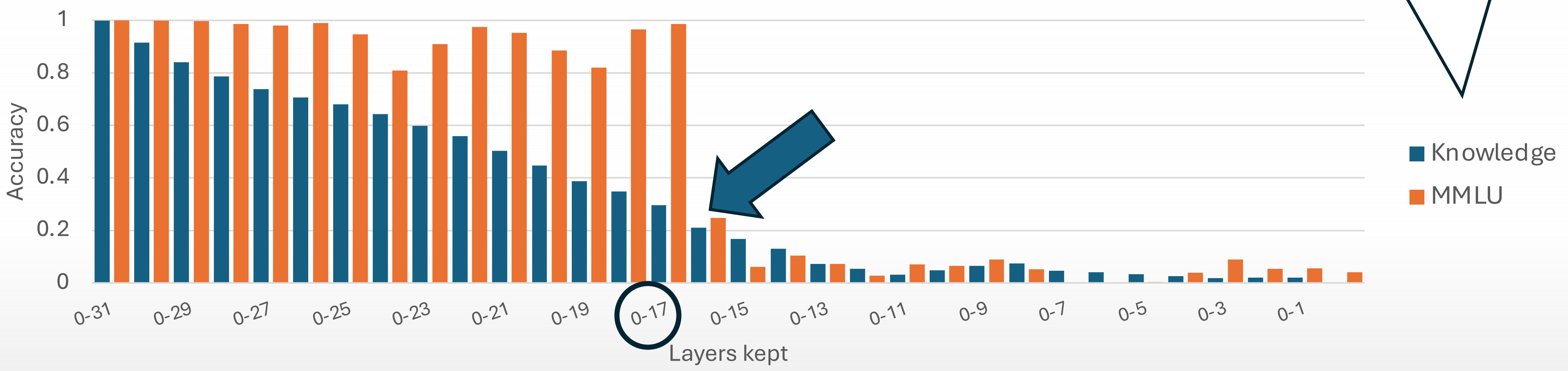
# Case Study: MMLU Benchmark

**Gradual Pruning.** Prune layers from the end of Llama-3.1-8B

After each prune, we measure how much knowledge is retained

- Knowledge extraction is distributed
- `L17` removal significantly harms MMLU

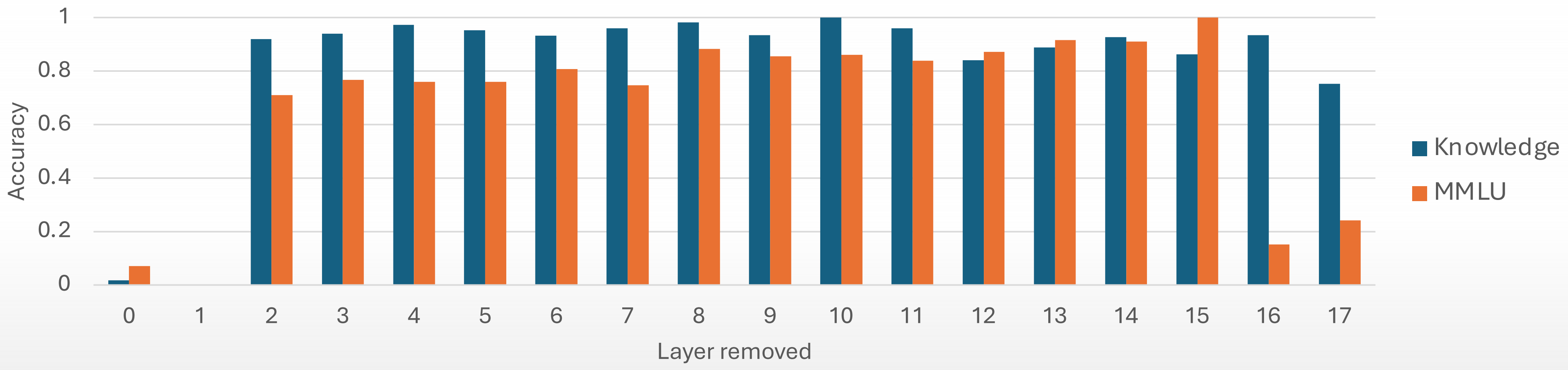
<u>Minimal Retrieval Tasks</u>
ARC-Challenge, ARC-Easy, PIQA, Winogrande, OpenBookQA, HellaSwag



# Case Study: MMLU Benchmark

**Individual Pruning**. Remove layer, evaluate, and reinsert

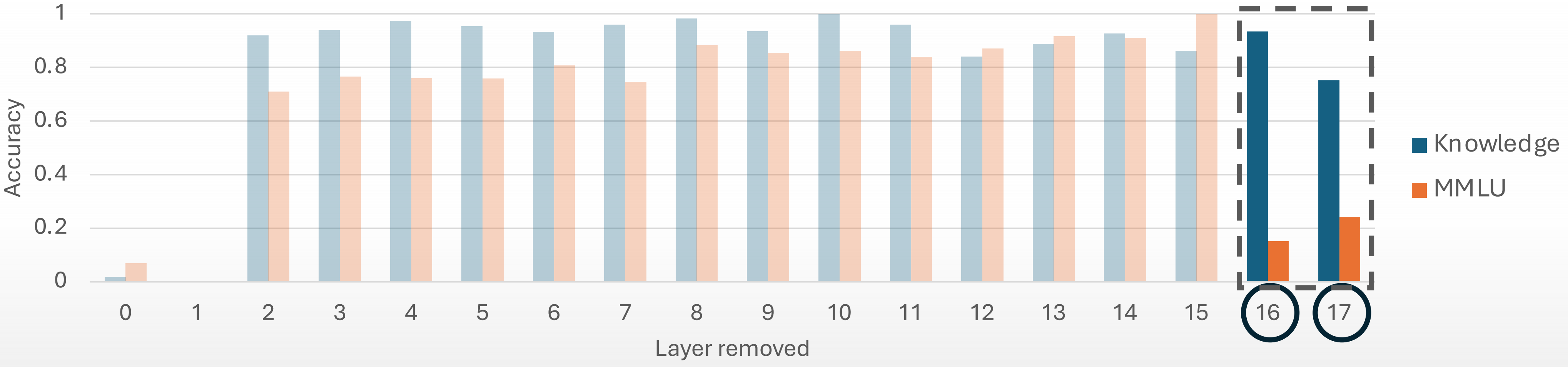
- We first remove all layers above `L17` from Llama-3.1-8B



# Case Study: MMLU Benchmark

**Individual Pruning**. Remove layer, evaluate, and reinsert

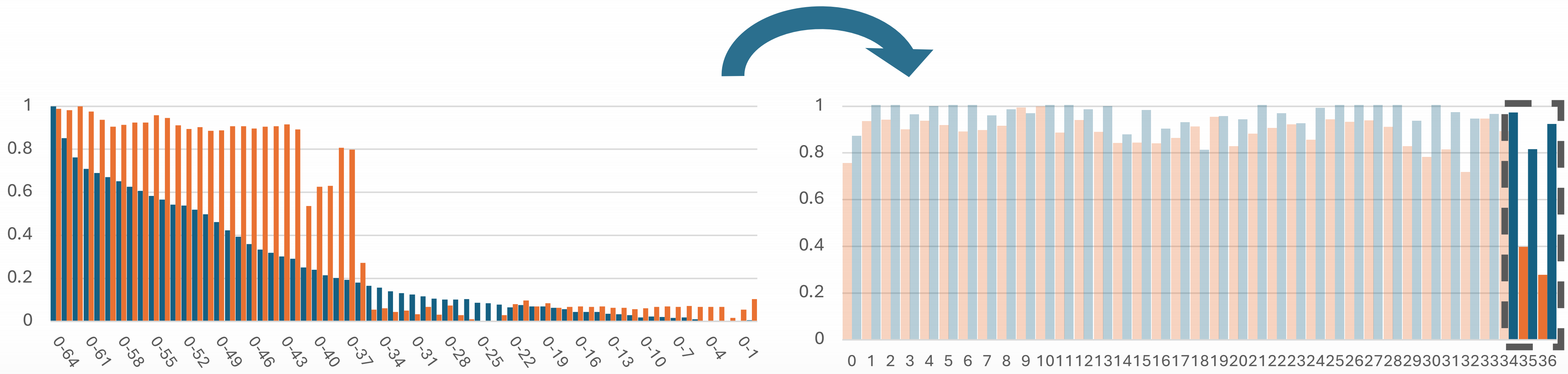
- We first remove all layers above `L17` from Llama-3.1-8B
- `L16` & `L17` removal significantly harms MMLU



# Case Study: MMLU Benchmark

Same goes for Falcon-Mamba-7B (based on Mamba-1).

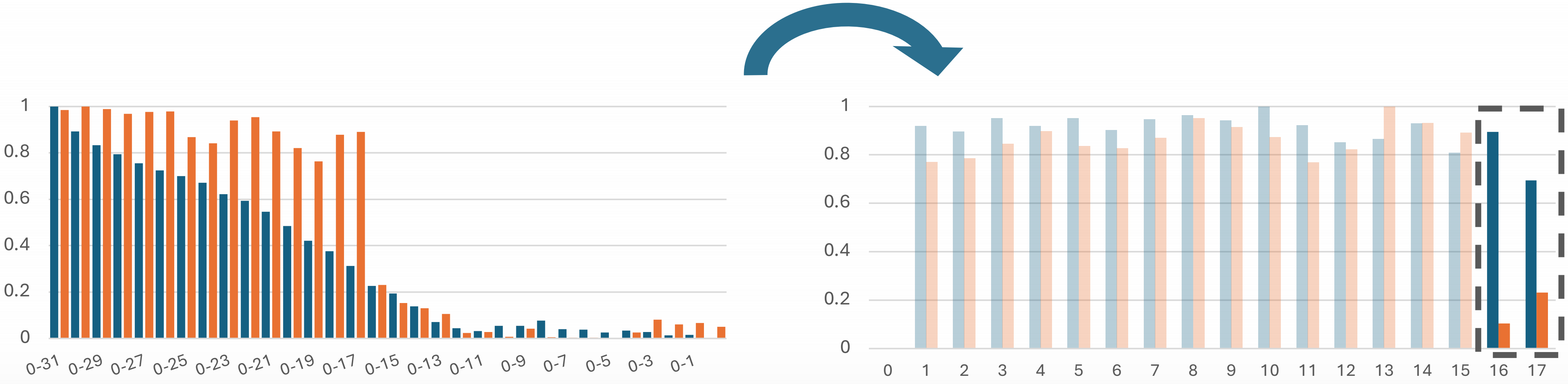
- `L35` & `L36` removal significantly harms MMLU



# Case Study: MMLU Benchmark

Same goes for Llamba-8B (based on Mamba-2).

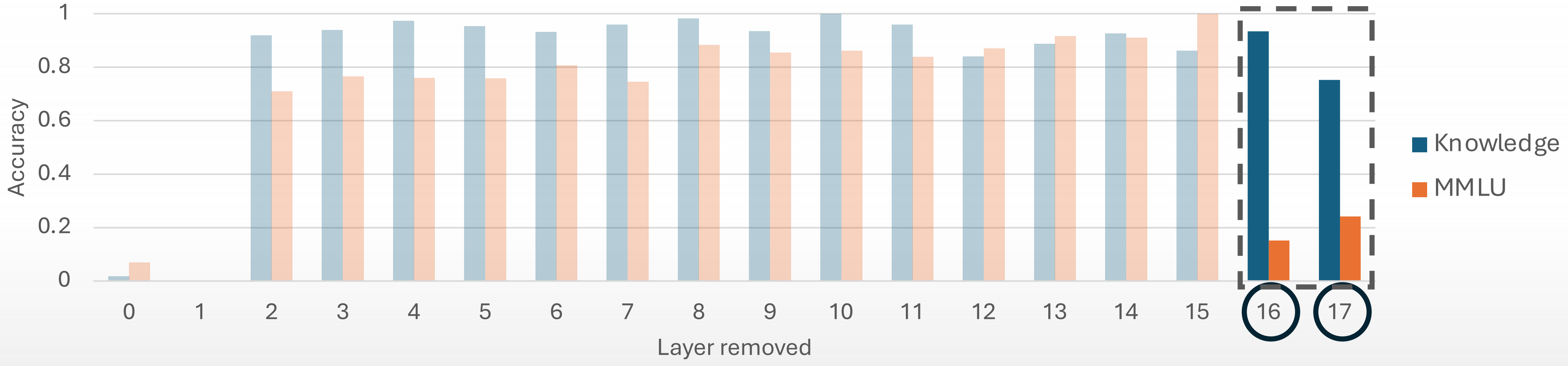
- `L16` & `L17` removal significantly harms MMLU



# Case Study: MMLU Benchmark

What exactly is happening in those two layers?

We probe Llama-3.1-8B's heads.



# Case Study: MMLU Benchmark

**Heads Pruning**. Keeping heads whose removal hurts MMLU

- `L16H22` and `L17H24` are part of a mechanism for MMLU.
- What's so important about `L16H22` and `L17H24` ?

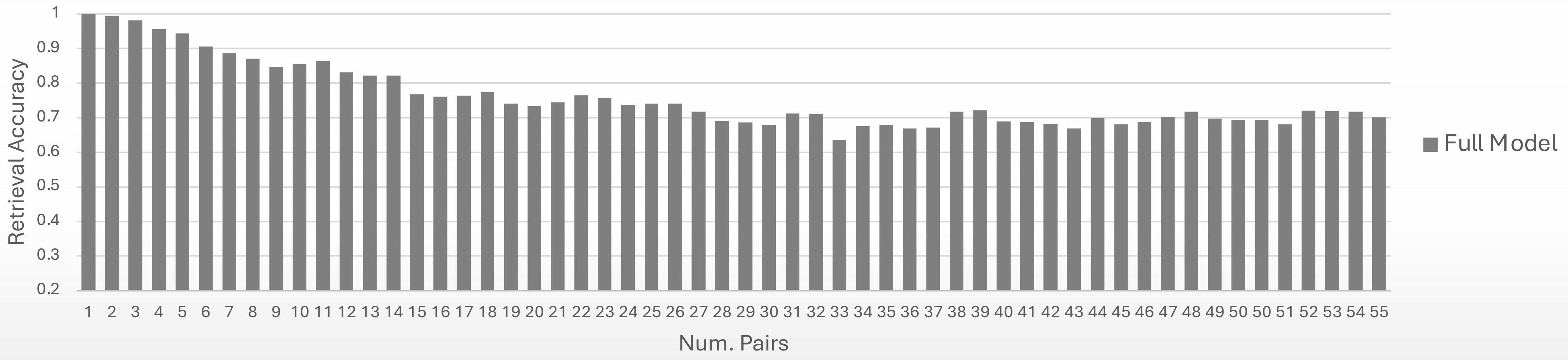
| Heads Kept in a Layer | | | Metrics (%) | |
|---|---|---|---|---|
| **0-15** | **16** | **17** | **MMLU** | **Knowledge Tasks** |
| 0,1,...,31 | 22 | 24 | **66.32** | 39.09 |
| 0,1,...,31 | ∅ | 24 | 24.36 | 39.18 |
| 0,1,...,31 | 22 | ∅ | 25.59 | |
| 0,1,...,31 | ∅ | ∅ | 25.56 | |

= Random Guess

# Retrieval in MMLU

We test Llama-3.1-8B on KV-Retrieval with growing dictionary sizes.

```
Memorize the
dictionary:
present:50
scallops:84
…
psychiatry:67
The value of the key
'scallops' is
```

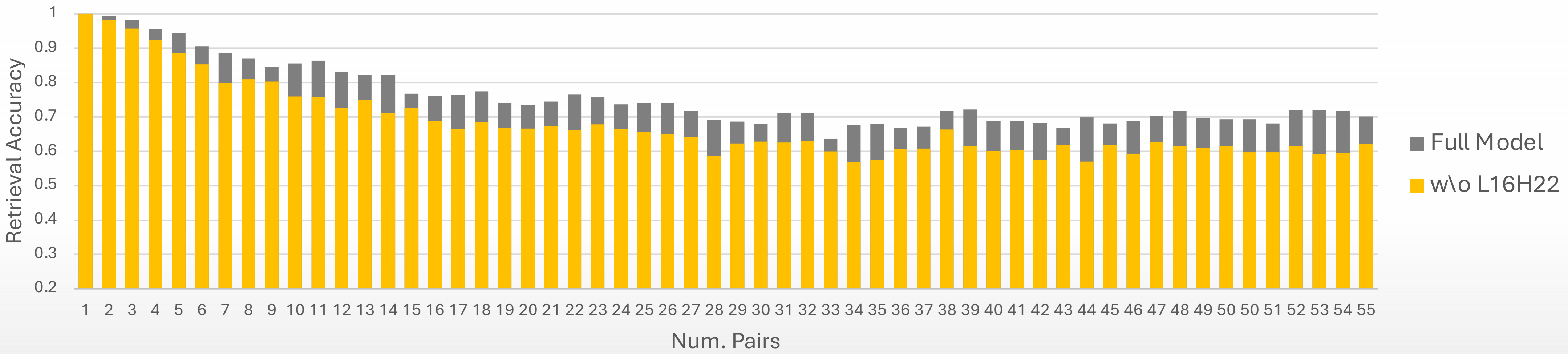


# Retrieval in MMLU

We test Llama-3.1-8B on KV-Retrieval with growing dictionary sizes.

- `L16H22` removal causes a constant drop.

```
Memorize the
dictionary:
present:50
scallops:84
…
psychiatry:67
The value of the key
'scallops' is
```



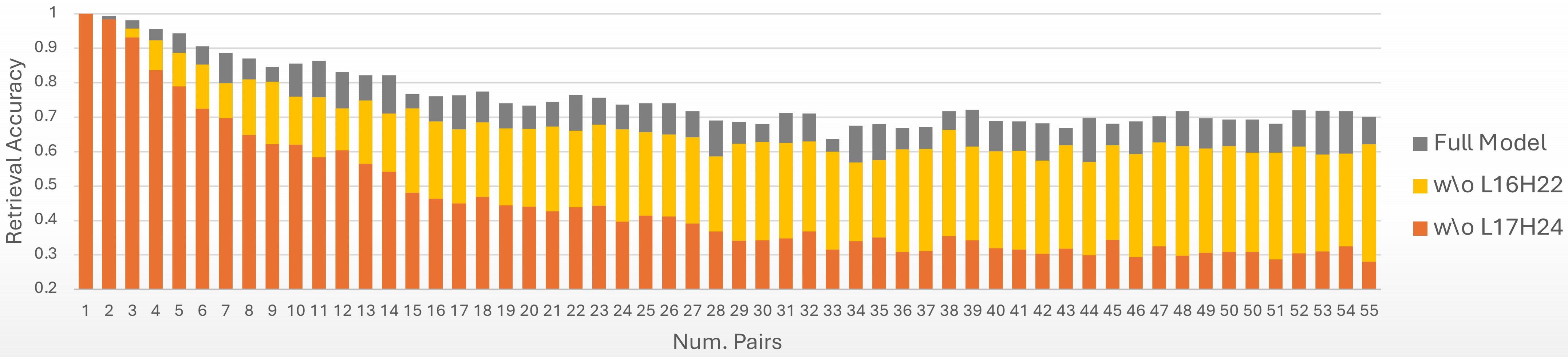
# Retrieval in MMLU

We test Llama-3.1-8B on KV-Retrieval with growing dictionary sizes.

- `L16H22` removal causes a constant drop.
- `L17H24` removal causes drops as complexity increases.

⇒ `L16H22` & `L17H24` are part of a retrieval mechanism.

```
Memorize the
dictionary:
present:50
scallops:84
…
psychiatry:67
The value of the key
'scallops' is
```



# Retrieval in MMLU

We test Llama-3.1-8B on KV-Retrieval with growing dictionary sizes.

- `L16H22` removal causes a constant drop.
- `L17H24` removal causes drops as complexity increases.

⇒ `L16H22` & `L17H24` are part of a retrieval mechanism.

| Heads Kept in a Layer | | | Metrics (%) | |
|---|---|---|---|---|
| **0-15** | **16** | **17** | **MMLU** | **Knowledge Tasks** |
| 0,1,...,31 | 22 | 24 | **66.32** | 39.09 |
| 0,1,...,31 | ∅ | 24 | 24.36 | 39.18 |
| 0,1,...,31 | 22 | ∅ | 25.59 | 39.21 |
| 0,1,...,31 | ∅ | ∅ | 25.56 | 39.21 |

MMLU difficulty is more **retrieval** than knowledge

# Outline

1. Retrieval in both Transformers and SSMs is performed similarly, in just a few heads.
   $\Longrightarrow$ Transformer-SSM performance gap stems from these heads

2. SSMs approximate these heads weakly

3. Hybrid models close the gap!

# Outline

1. Retrieval in both Transformers and SSMs is performed similarly, in just a few heads.
   $\Longrightarrow$ Transformer-SSM performance gap stems from these heads

How do `L16H22` and `L17H24` perform it?

- They implement a Gather-and-Aggregate mechanism.

# Gather-and-Aggregate

Two heads collaborate to retrieve:

- **Gather Head** condenses token segments (e.g., `L16H22`),

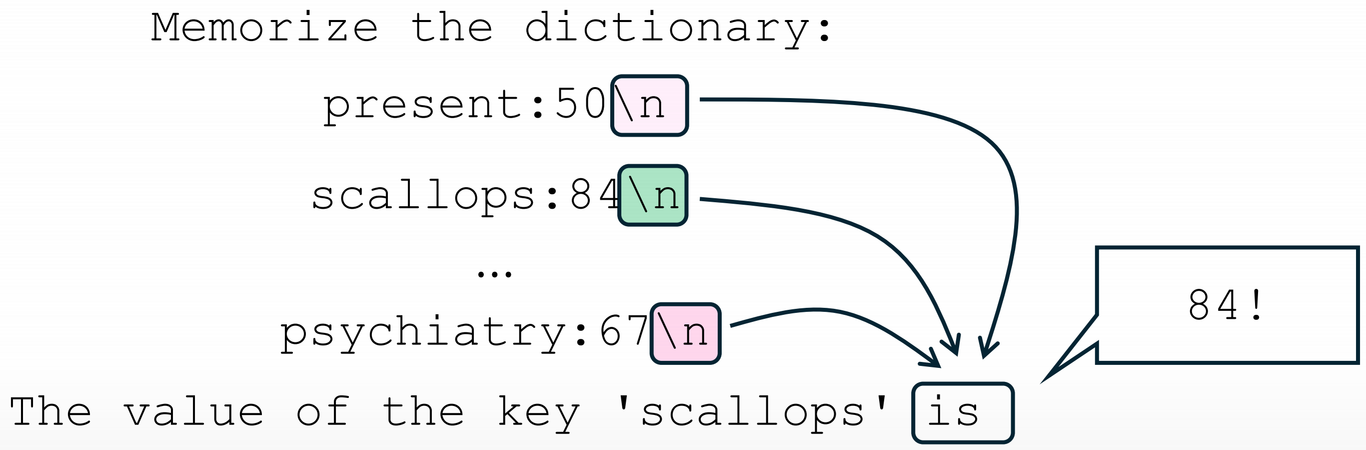
```
Memorize the dictionary:
present:50\n
scallops:84\n
...
psychiatry:67\n
The value of the key 'scallops' is
```

**Gather Head**

# Gather-and-Aggregate

Two heads collaborate to retrieve:

- **Gather Head** condenses token segments (e.g., `L16H22`),
- **Aggregate Head** integrates them into representation (e.g., `L17H24`).




**Aggregate Head**

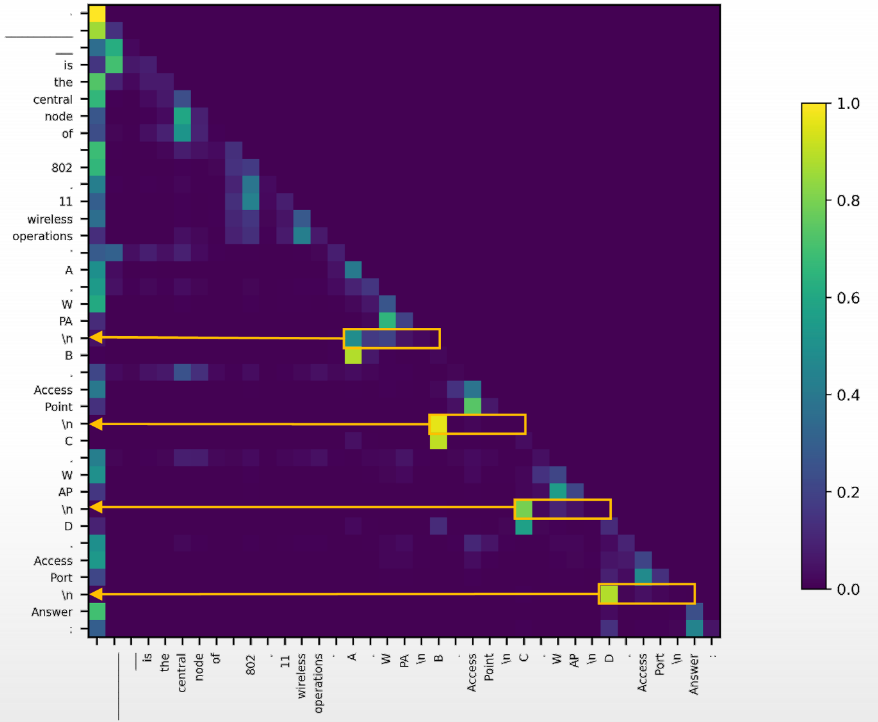
# Gather-and-Aggregate

Two heads collaborate to retrieve:

- **Gather Head** condenses token segments (e.g., `L16H22`),
- **Aggregate Head** integrates them into representation (e.g., `L17H24`).

```
___ is the central node of 802.11
wireless operations.\n
A. WPA\nB. Access Point\nC. WAP\n
D. Access Port\nAnswer :
```

**Gather Head**



# Gather-and-Aggregate

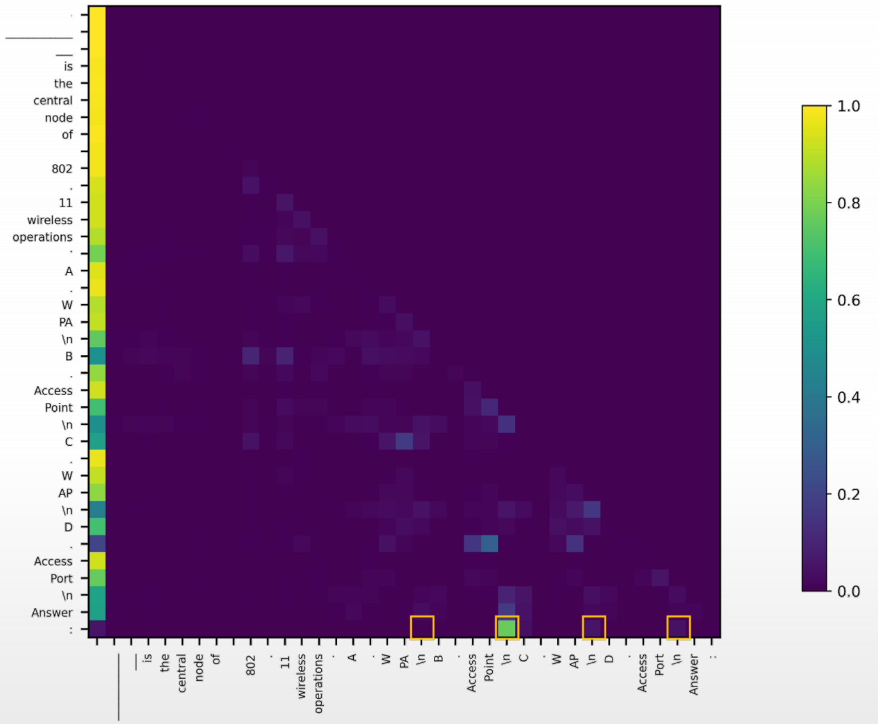
Two heads collaborate to retrieve:

- **Gather Head** condenses token segments (e.g., `L16H22`),
- **Aggregate Head** integrates them into representation (e.g., `L17H24`).

```
___ is the central node of 802.11
wireless operations.\n
A. WPA\nB. Access Point\nC. WAP\n
D. Access Port\nAnswer:
```

**Aggregate Head**

B!

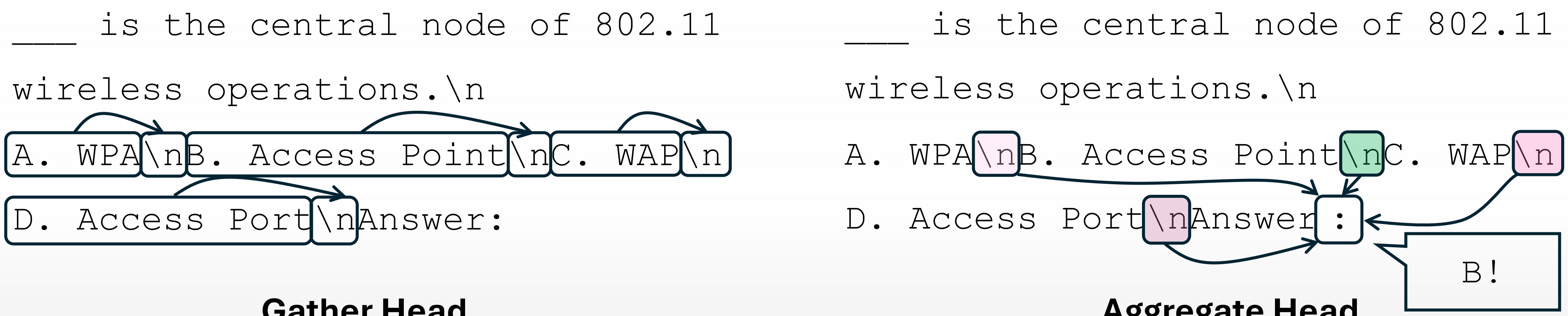


# Gather-and-Aggregate

Two heads collaborate to retrieve:

- **Gather Head** condenses token segments (e.g., `L16H22`),
- **Aggregate Head** integrates them into representation (e.g., `L17H24`).

➢ "Content Gatherer" and "Correct Letter" Heads [Lieberum et al.]




**Gather Head** **Aggregate Head**

Lieberum et al., "Does Circuit Analysis Interpretability Scale?"

# Gather-and-Aggregate

**Retrieval (and G&A) are implicitly involved in many tasks**

- We iteratively ablate each head, measure KV-Retrieval, and reinsert it to rank importance
- Removing top G&A heads impairs retrieval-heavy tasks, while knowledge remains stable

| MODEL | #HEADS | MMLU ACC ↑ | LAMB. PPL ↓ | GSM8K ACC ↑ | SWDE ACC ↑ | BBH ACC ↑ | KNOWLEDGE ACC ↑ |
|---|---|---|---|---|---|---|---|
| Llama-3B | 0 | 60.3 (+0.0%) | 4.8 (+0.0%) | 28.7 (+0.0%) | 85.8 (+0.0%) | 38.2 (+0.0%) | 60.5 (+0.0%) |
| | 10 | 53.1 (-12.0%) | 6.5 (+35.7%) | 17.4 (-39.4%) | 81.9 (-4.5%) | 33.4 (-12.6%) | 59.4 (-1.8%) |
| | 20 | 32.2 (-46.6%) | 8.8 (+82.8%) | 9.1 (-68.2%) | 57.5 (-33.0%) | 27.7 (-27.5%) | 58.7 (-3.0%) |
| | 30 | 29.9 (-50.4%) | 10.1 (+109%) | 5.6 (-80.5%) | 47.5 (-44.6%) | 25.4 (-33.5%) | 58.0 (-4.1%) |
| Llama-8B | 0 | 68.1 (+0.0%) | 3.4 (+0.0%) | 27.3 (+0.0%) | 90.8 (+0.0%) | 45.1 (+0.0%) | 68.5 (+0.0%) |
| | 10 | 61.9 (-9.1%) | 4.2 (+22.0%) | 21.7 (-20.5%) | 87.3 (-3.9%) | 37.7 (-16.5%) | 67.1 (-2.0%) |
| | 20 | 38.1 (-44.0%) | 6.8 (+98.6%) | 9.4 (-65.6%) | 79.5 (-12.4%) | 29.2 (-35.2%) | 64.8 (-5.4%) |
| | 30 | 38.7 (-43.2%) | 7.3 (+115%) | 7.8 (-71.4%) | 74.0 (-18.5%) | 29.0 (-35.7%) | 64.4 (-6.0%) |

# Gather-and-Aggregate

**Retrieval (and G&A) can be triggered by task format**

- We compare ARC-Challenge in chat vs. completion modes
- Chat requires more reasoning, boosting accuracy
- Removing G&A heads hurts chat more, reducing it to completion-level performance

| Model | #Removed Heads | ARC-C (Chat) acc ↑ | ARC-C (Regular) acc ↑ |
|---|---|---|---|
| Llama-3B | 0 | 76.8 (+0.0%) | 45.5 (+0.0%) |
| | 10 | 72.2 (-6.0%) | 43.6 (-4.2%) |
| | 20 | 50.0 (-34.9%) | 42.0 (-7.7%) |
| | 30 | 43.2 (-43.8%) | 41.9 (-7.9%) |
| Llama-8B | 0 | 84.3 (+0.0%) | 54.9 (+0.0%) |
| | 10 | 77.1 (-8.5%) | 51.6 (-6.0%) |
| | 20 | 49.3 (-41.5%) | 47.3 (-13.8%) |
| | 30 | 53.6 (-36.4%) | 47.9 (-12.8%) |

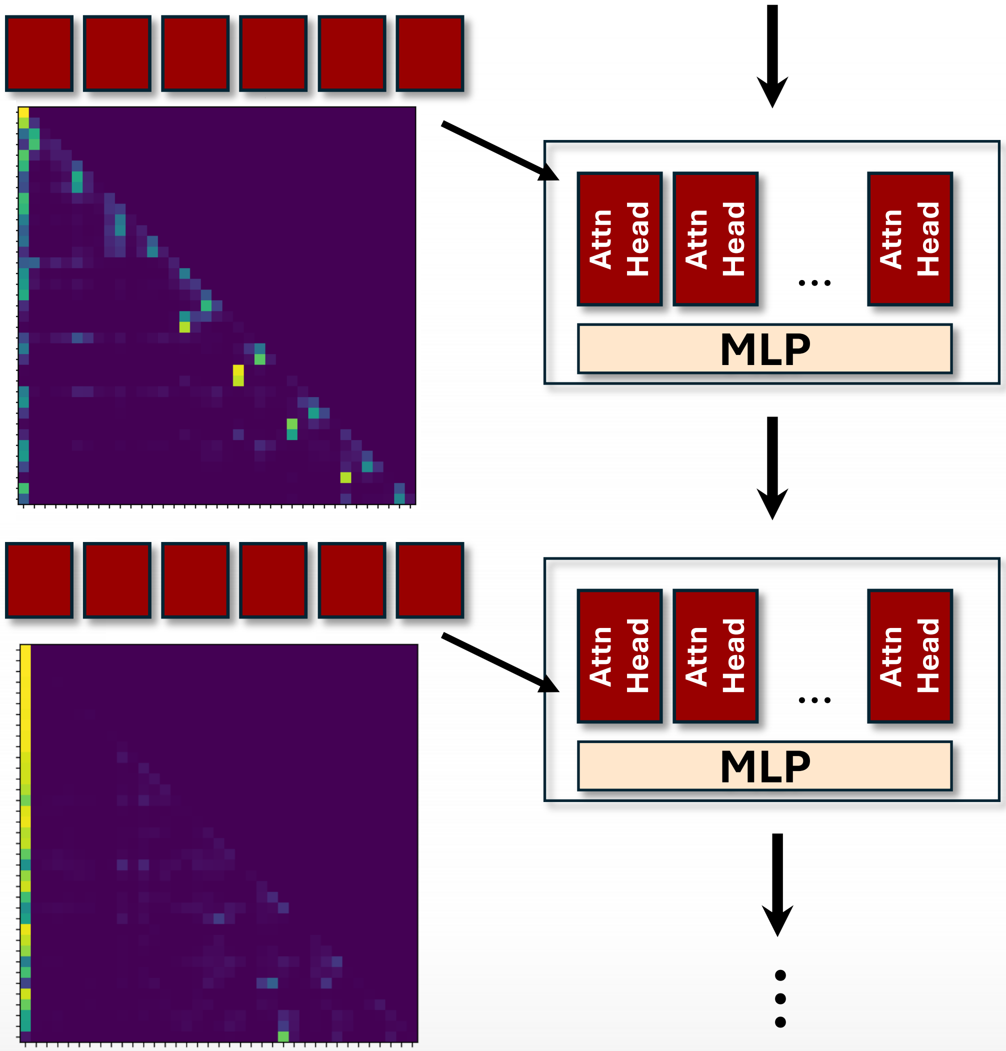
# Gather-and-Aggregate

**A mechanistic view of attention-based retrieval**

- Attention retrieves well by caching history (intuitive)
- Mechanistically, this enables sharp, noise-free G&A mappings

**Not all heads retrieve**

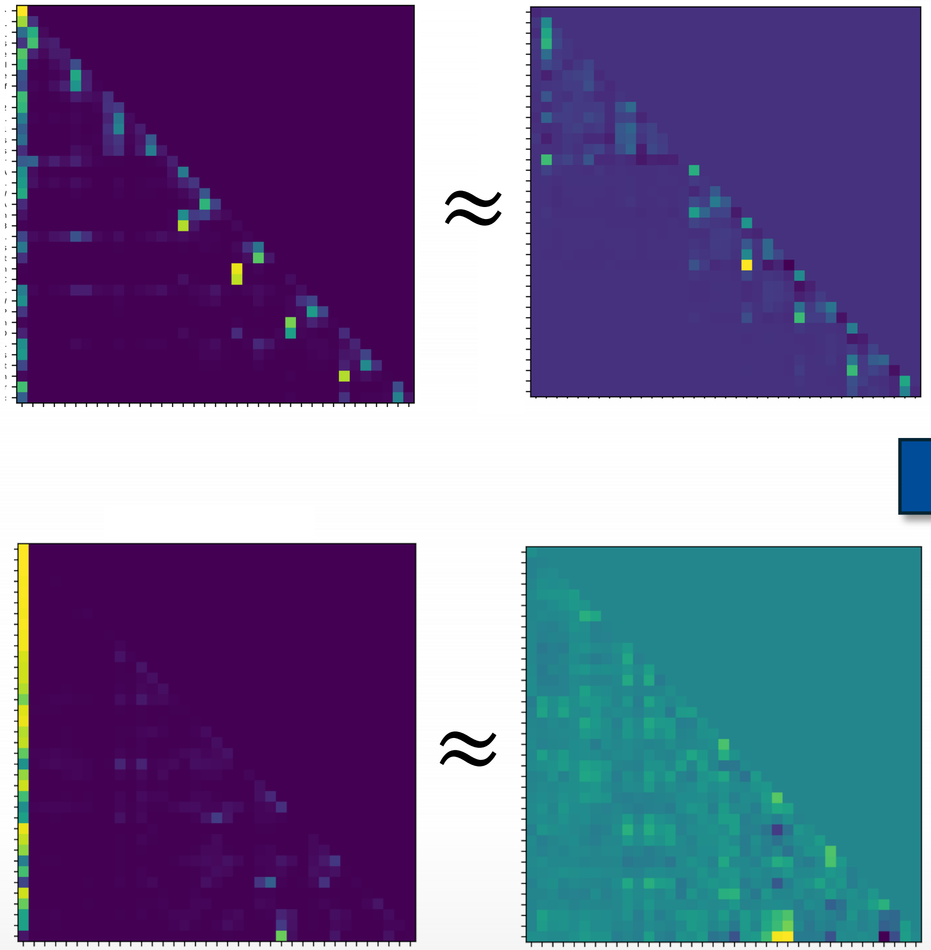
- Only a few key heads drive this behavior
- These heads are critical across many tasks



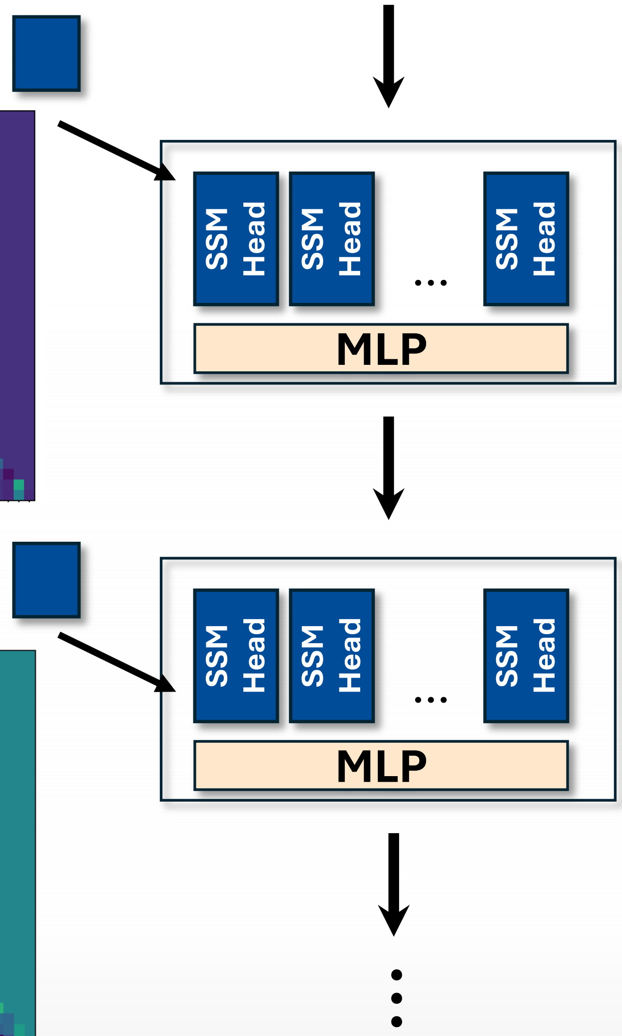
# Gather-and-Aggregate

**What about SSMs?**

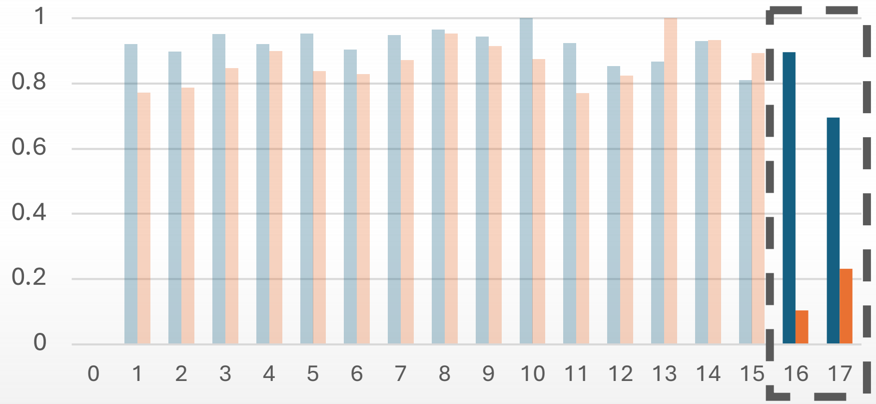
- Visually resemble G&A heads
- But they are noisy…
- Do they implement G&A?







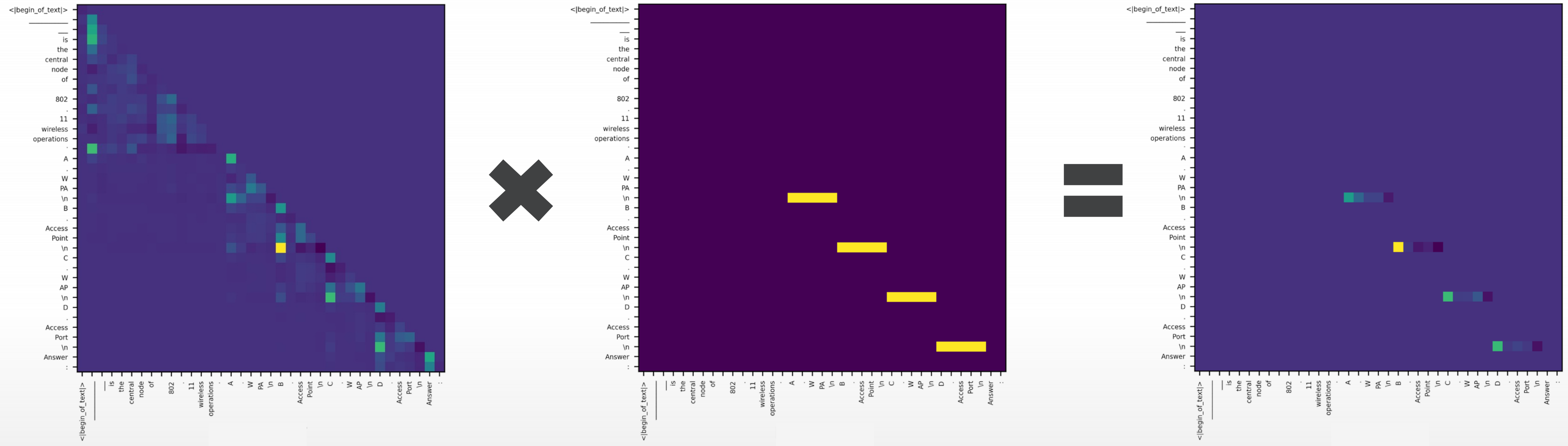




# Gather-and-Aggregate

**Masking shows SSMs use G&A.**

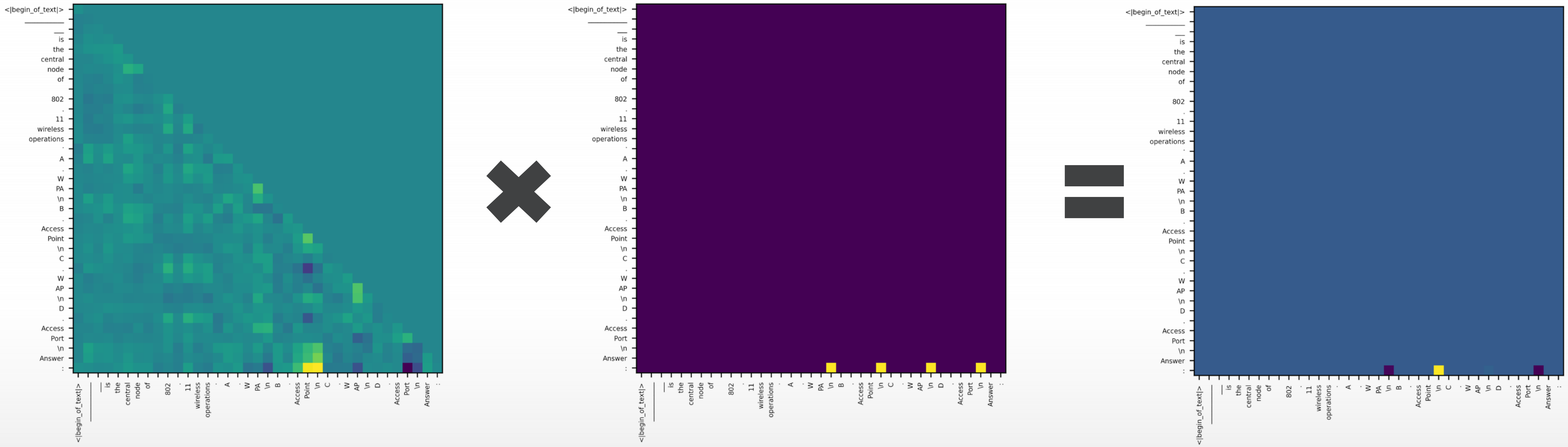
- A custom mask is generated for each MMLU sample.
- For the **Gather head**, we unmask the answer segments.



# Gather-and-Aggregate

**Masking shows SSMs use G&A.**

- A custom mask is generated for each MMLU sample.
- For the **Aggregate head**, we unmask the summary tokens.

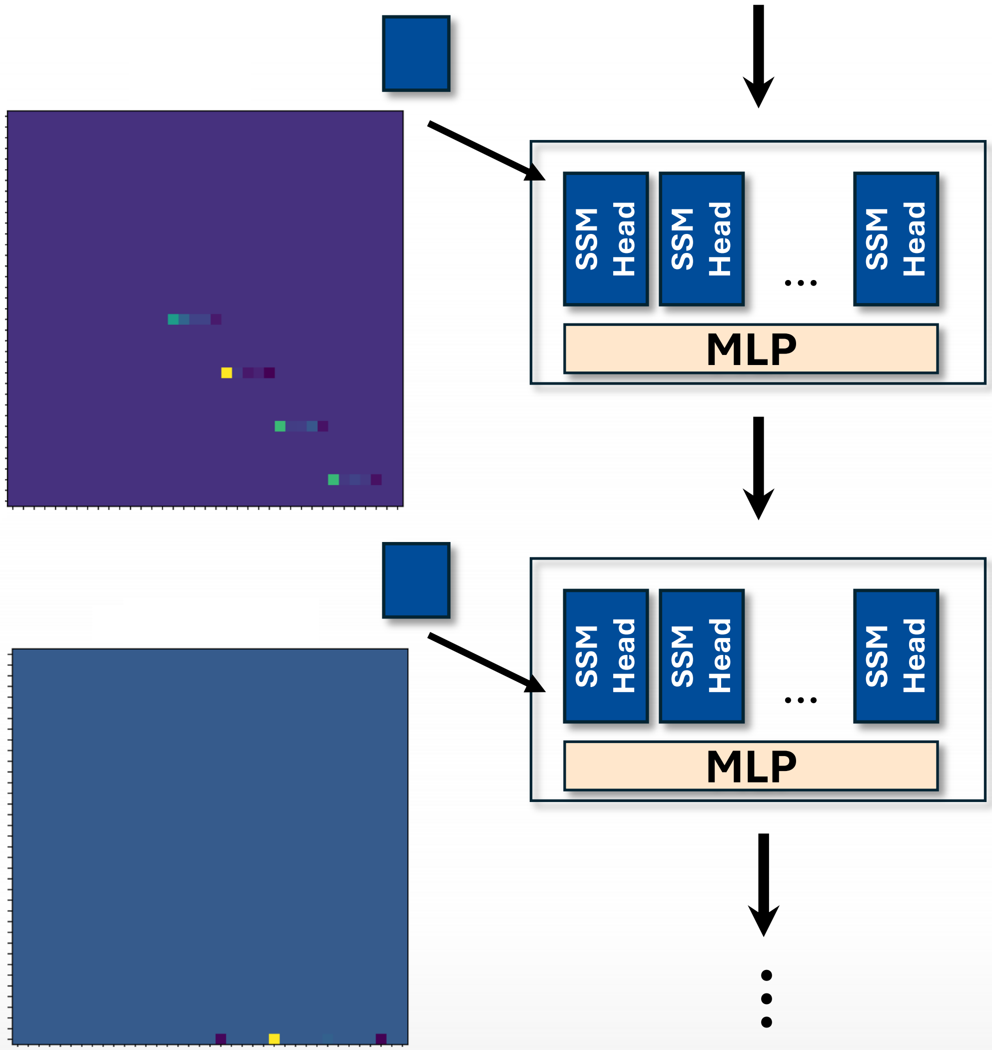


# Gather-and-Aggregate

**Masking shows SSMs use G&A.**

- Recall: Fully masking G&A drops MMLU to near-random.
- Preserving only the G&A pattern (with mask) keeps MMLU high.

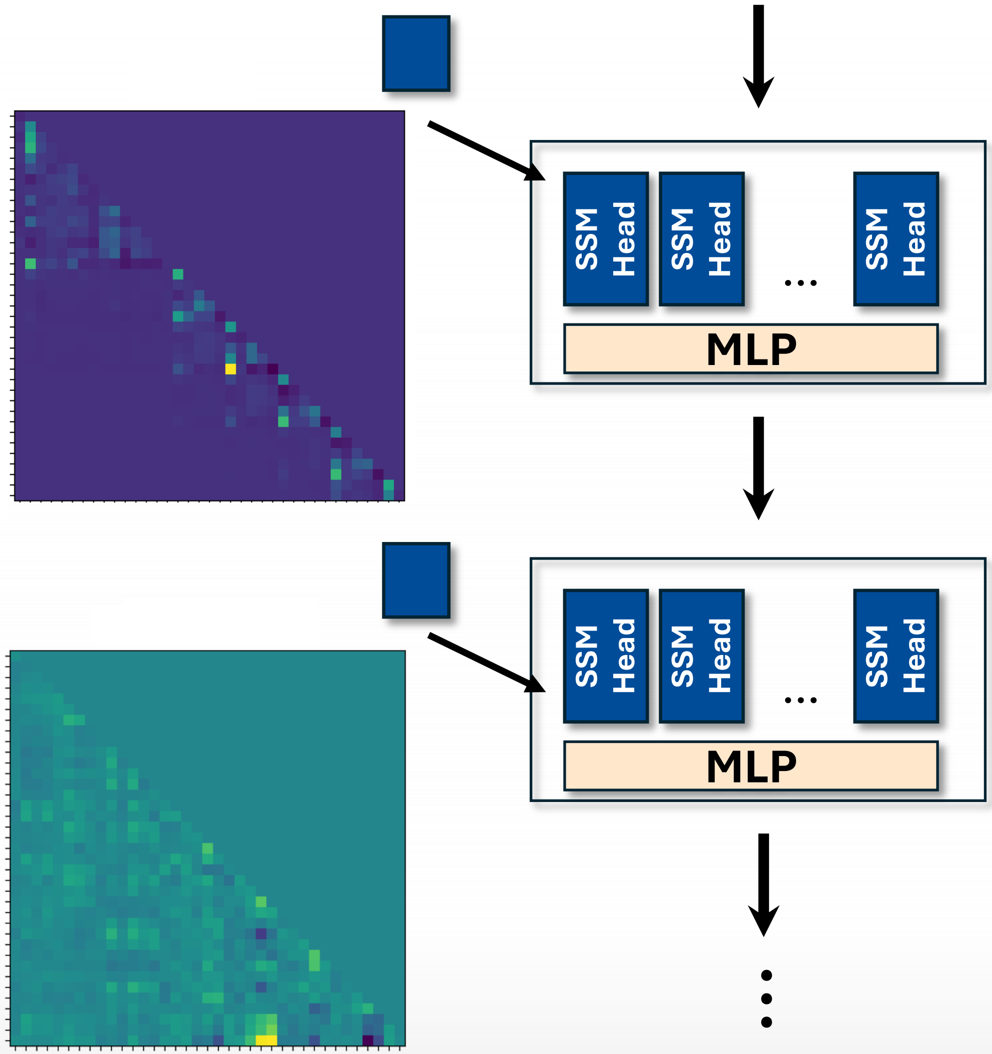
⇒ SSMs develop G&A too!



# SSMs struggle with G&A

**A mechanistic view of SSM-based retrieval**

- Hidden states compress history into one evolving representation
- SSMs implement smoother version of G&A
- This adds noise, reducing G&A power



# SSMs struggle with G&A

**SSM-based G&A has higher redundancy:**

- SSMs are less sensitive to G&A ablation than attention models.
- SSM models compensate for weaker G&A

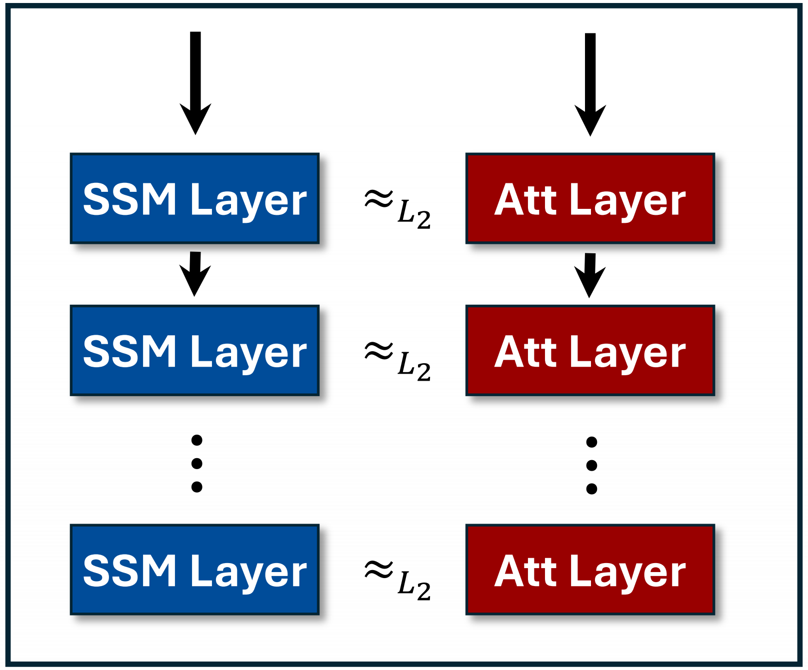
| Model | #Heads | MMLU ACC ↑ | LAMB. PPL ↓ | SWDE ACC ↑ | BBH ACC ↑ | Knowledge ACC ↑ |
|---|---|---|---|---|---|---|
| Llama-3B (Transformer) | 0 | 60.3 (+0.0%) | 4.8 (+0.0%) | 85.8 (+0.0%) | 38.2 (+0.0%) | 60.5 (+0.0%) |
| | 10 | 53.1 (-12.0%) | 6.5 (+35.7%) | 81.9 (-4.5%) | 33.4 (-12.6%) | 59.4 (-1.8%) |
| | 20 | 32.2 (-46.6%) | 8.8 (+82.8%) | 57.5 (-33.0%) | 27.7 (-27.5%) | 58.7 (-3.0%) |
| | 30 | 29.9 (-50.4%) | 10.1 (+109%) | 47.5 (-44.6%) | 25.4 (-33.5%) | 58.0 (-4.1%) |
| Llamba-3B (SSM) | 0 | 52.5 (+0.0%) | 3.6 (+0.0%) | 21.3 (+0.0%) | 9.2 (+0.0%) | 63.8 (+0.0%) |
| | 10 | 42.6 (-18.9%) | 5.2 (+44.4%) | 18.6 (-12.7%) | 9.0 (-2.2%) | 63.7 (-0.2%) |
| | 20 | 41.3 (-21.3%) | 8.2 (+128%) | 18.1 (-15.0%) | 9.0 (-2.2%) | 63.1 (-1.1%) |
| | 30 | 41.2 (-21.5%) | 9.1 (+153%) | 18.1 (-15.0%) | 9.0 (-2.2%) | 62.6 (-1.9%) |

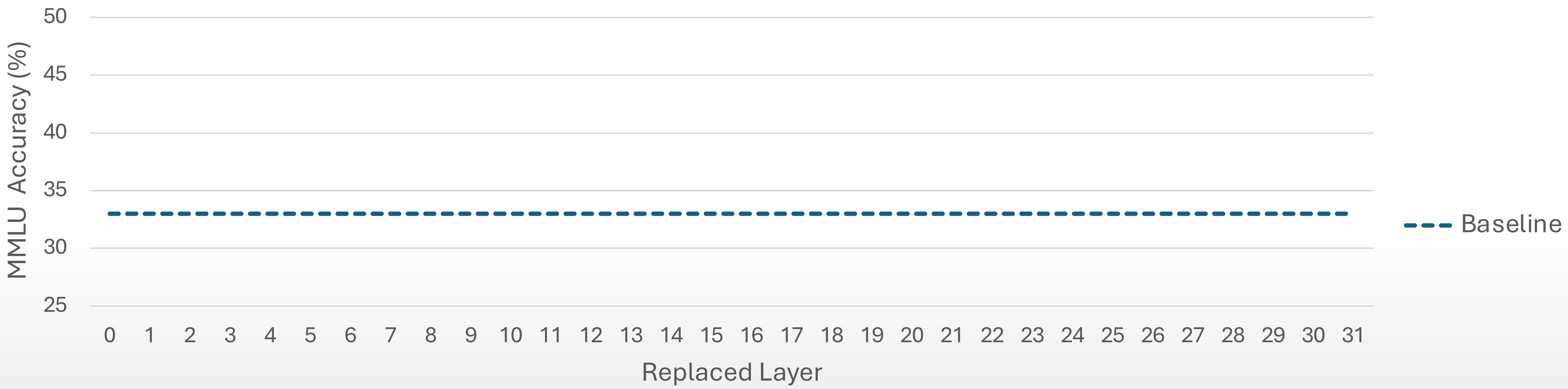
# SSMs struggle with G&A

**SSM-based G&A struggle to match attention:**

- After alignment, each SSM layer mimics its corresponding attention layer.
- **Baseline**: MMLU is **33%** and knowledge is 69%

Layer-to-Layer Distillation [Bick et al. ]



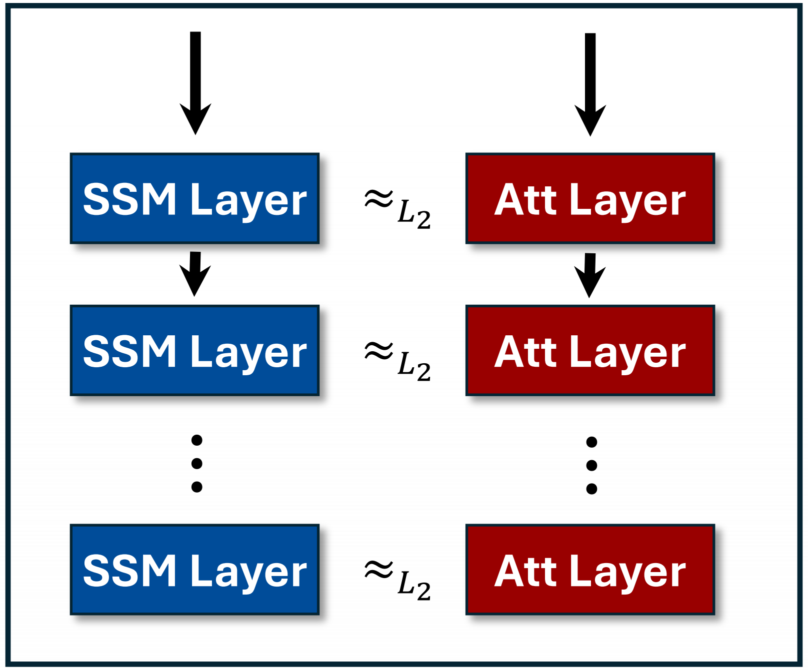

Bick et al., "Transformers to SSMs: Distilling Quadratic Knowledge to Subquadratic Models"

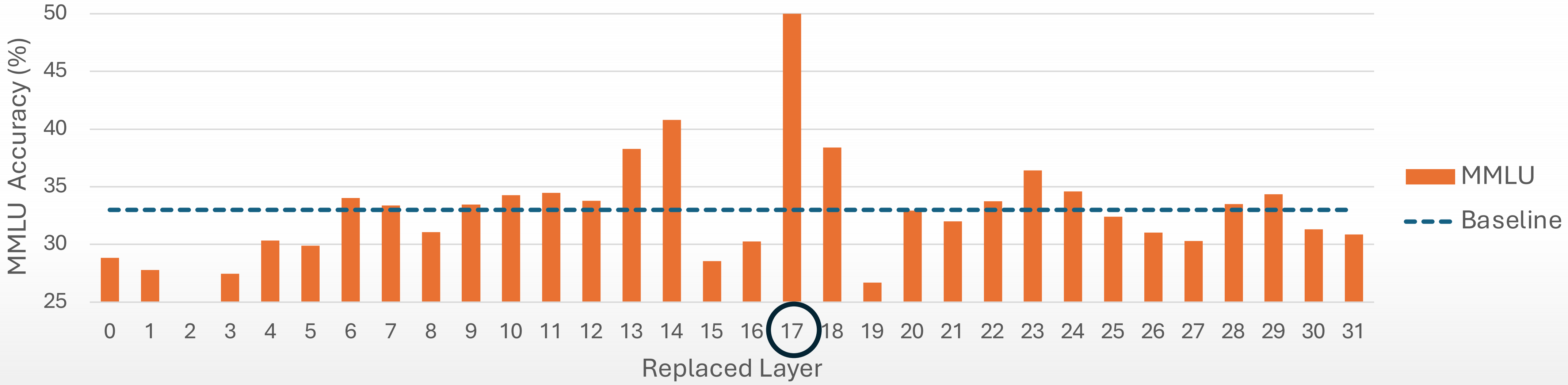
# SSMs struggle with G&A

**SSM-based G&A struggle to match attention:**

- After alignment, each SSM layer mimics its corresponding attention layer.
- **Baseline**: MMLU is 33% and knowledge is 69%
- **Replacing L17**: MMLU is 50% and knowledge remains 69%

Layer-to-Layer Distillation [Bick et al. ]




Bick et al., "Transformers to SSMs: Distilling Quadratic Knowledge to Subquadratic Models"

# Outline

1. Retrieval in both Transformers and SSMs is performed similarly, in just a few heads.
   $\Longrightarrow$ Transformer-SSM performance gap stems from these heads

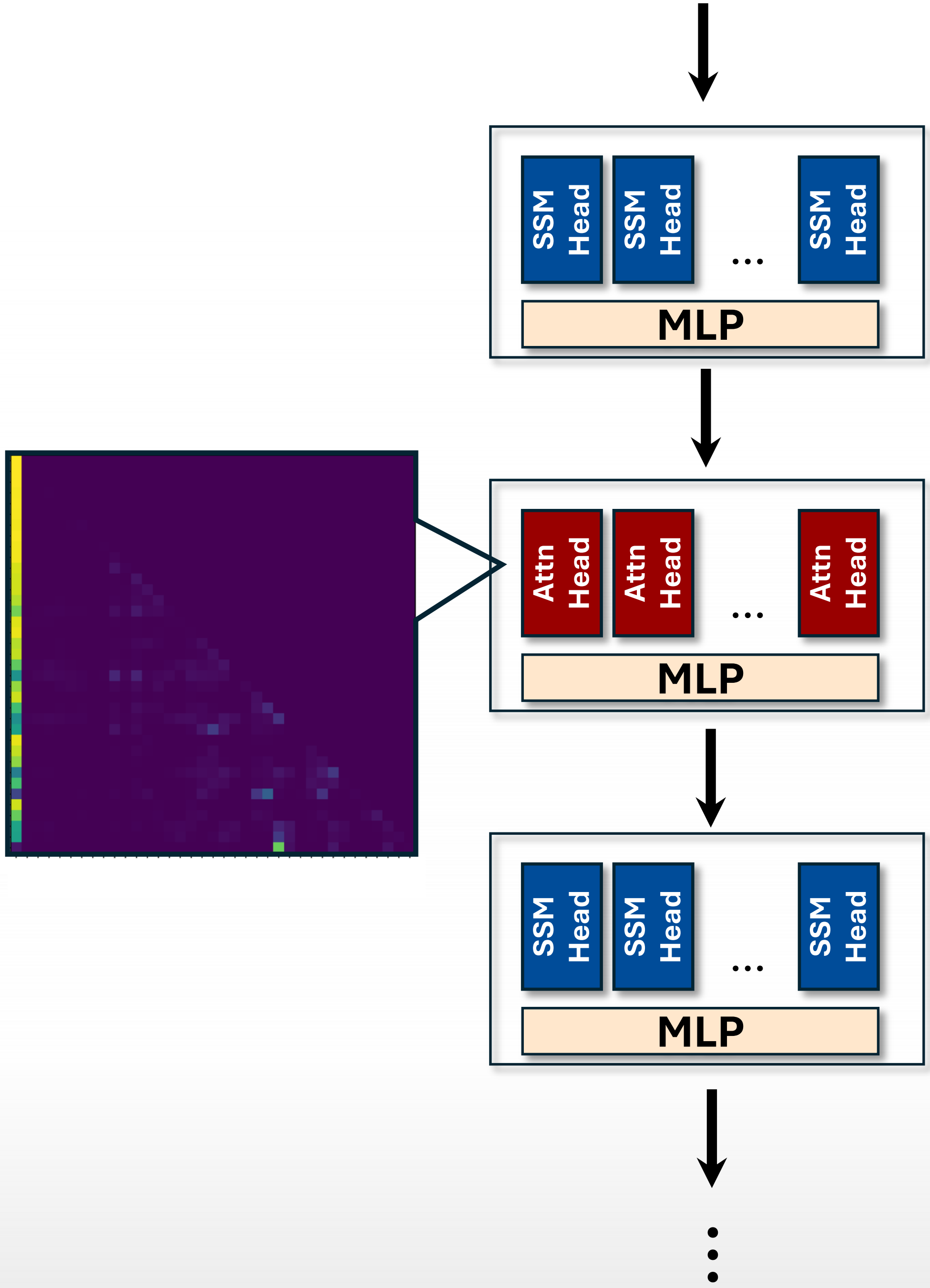
2. SSMs approximate these heads weakly

3. Hybrid models close the gap!

# Hybrid Models

**Hybrid models overcome SSMs' retrieval limits**

- A few attention layers interleaved with mostly SSMs
- Attention handles aggregation
- SSMs handle language modeling and knowledge



# Hybrid Models

**Attention handles aggregation:**

- Attention-based Aggregates are masked, with SSMs left untouched
- Knowledge tasks remain stable
- Retrieval-heavy tasks drop sharply

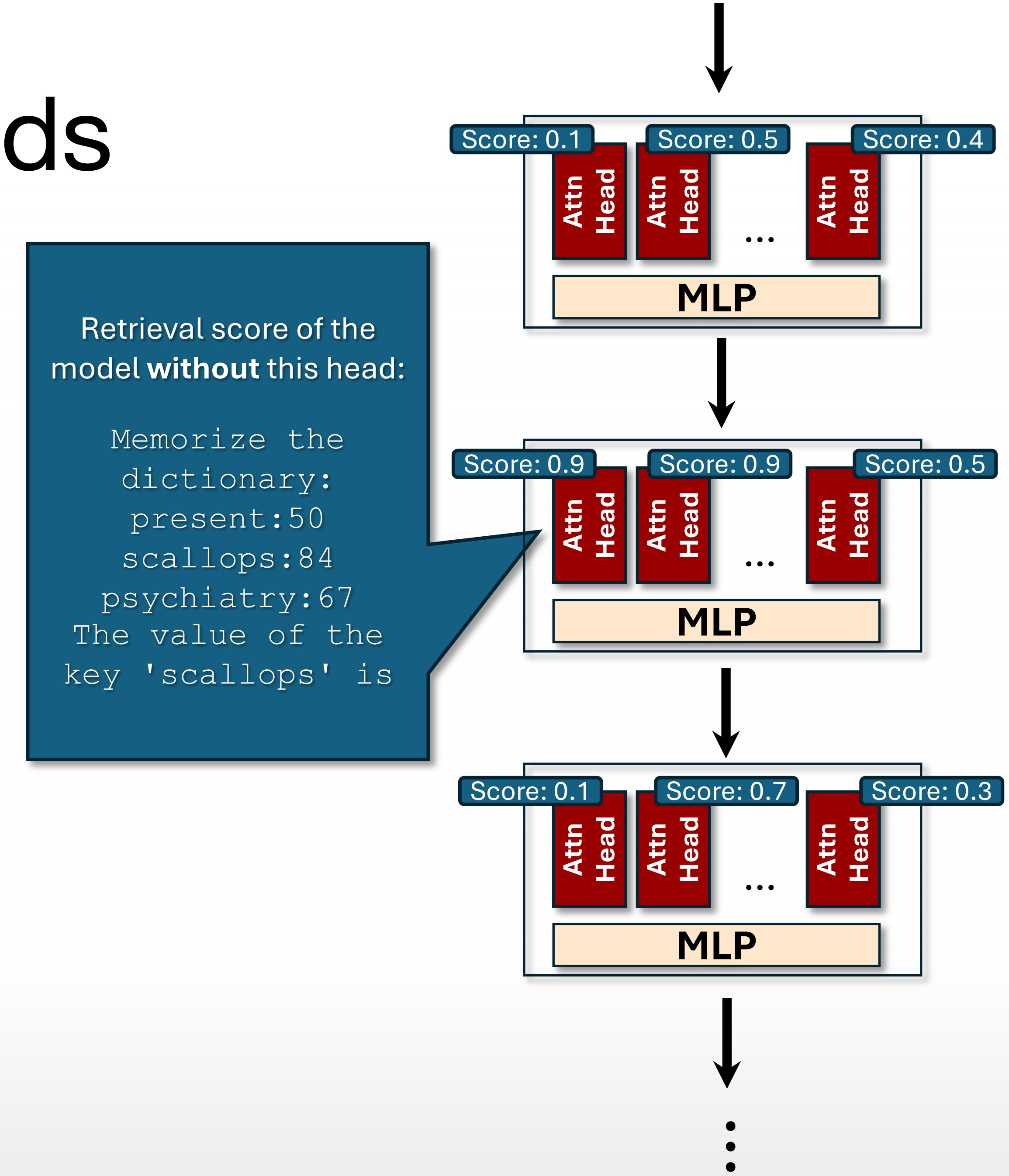
| Model | #Heads | MMLU ACC ↑ | LAMB. PPL ↓ | GSM8K ACC ↑ | SWDE ACC ↑ | BBH ACC ↑ | Knowledge ACC ↑ |
|---|---|---|---|---|---|---|---|
| Zamba2-2.7B | 0 | 55.7 (+0.0%) | 4.2 (+0.0%) | 57.4 (+0.0%) | 89.5 (+0.0%) | 30.6 (+0.0%) | 66.8 (+0.0%) |
| | 10 | 42.4 (-23.9%) | 12.8 (+204%) | 24.7 (-57.0%) | 84.3 (-5.8%) | 25.5 (-16.7%) | 64.8 (-3.0%) |
| | 20 | 37.2 (-33.2%) | 22.2 (+428%) | 6.5 (-88.7%) | 74.4 (-16.9%) | 17.4 (-43.1%) | 62.6 (-6.3%) |
| Zamba2-7B | 0 | 65.1 (+0.0%) | 3.1 (+0.0%) | 60.5 (+0.0%) | 91.7 (+0.0%) | 33.0 (+0.0%) | 70.6 (+0.0%) |
| | 20 | 57.3 (-12.0%) | 5.2 (+67.7%) | 27.6 (-54.4%) | 75.1 (-18.1%) | 28.9 (-12.4%) | 67.5 (-4.4%) |
| | 40 | 50.6 (-22.3%) | 9.5 (+206%) | 14.9 (-75.4%) | 41.2 (-55.1%) | 21.7 (-34.2%) | 67.0 (-5.1%) |
| | 60 | 36.2 (-44.4%) | 19.8 (+538%) | 7.2 (-88.1%) | 39.6 (-56.8%) | 15.9 (-51.8%) | 66.5 (-5.8%) |

# Retrieval-Guided Hybrids

**A better strategy to merge their strengths?**

**Distillation:** Keep attention only where needed

1. Evaluate each ablated model on synthetic KV-Retrieval
2. Sort heads by ablation score

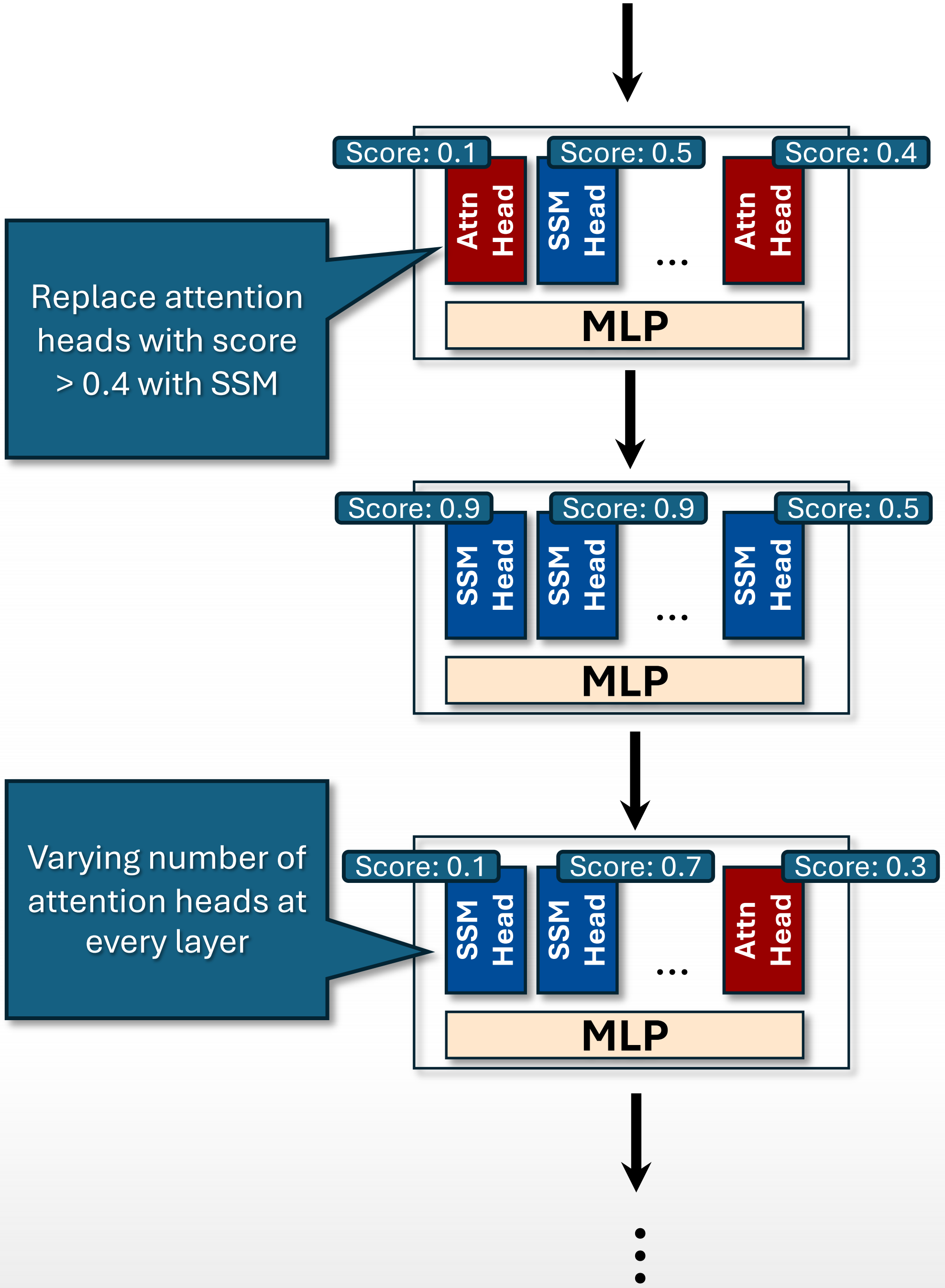


# Retrieval-Guided Hybrids

**A better strategy to merge their strengths?**

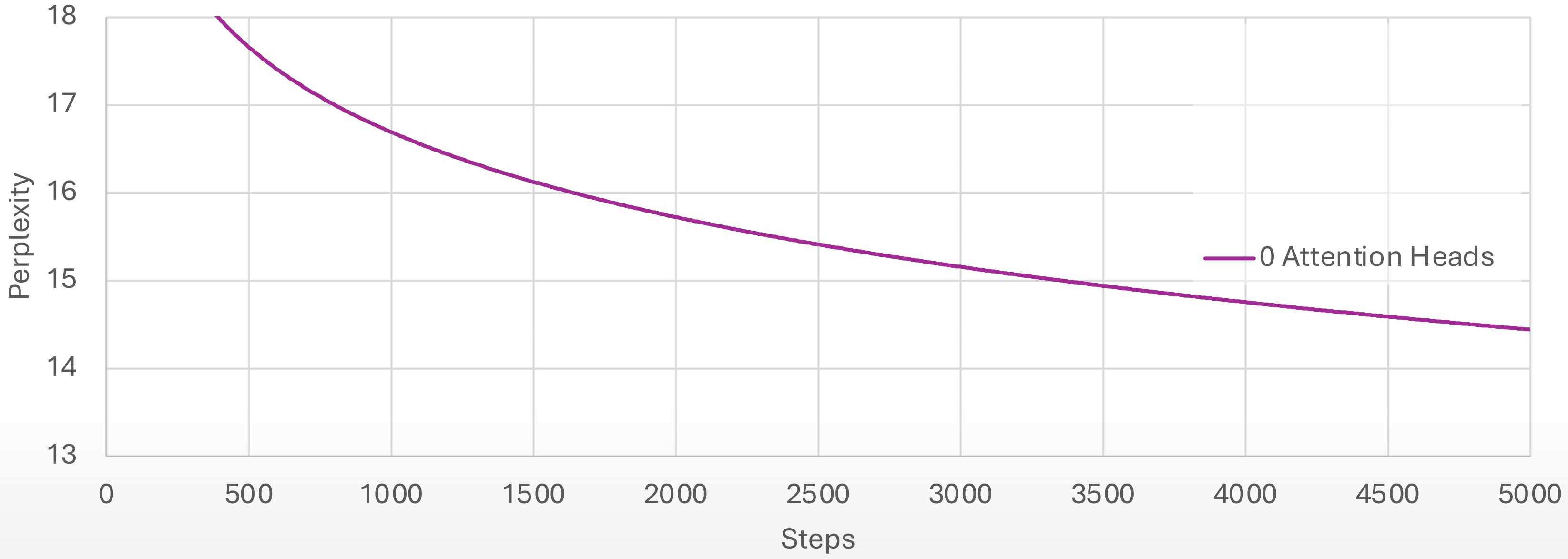
**Distillation:** Keep attention only where needed

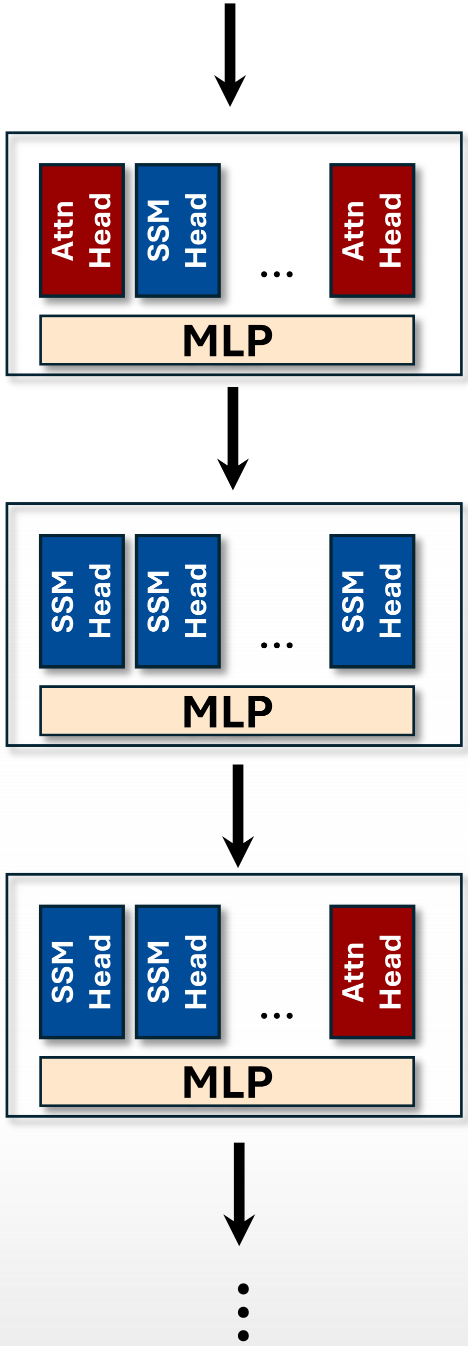
1. Evaluate each ablated model on synthetic KV-Retrieval
2. Sort heads by ablation score
3. Retain heads with largest performance drops (they're most critical for retrieval)



# Retrieval-Guided Hybrids

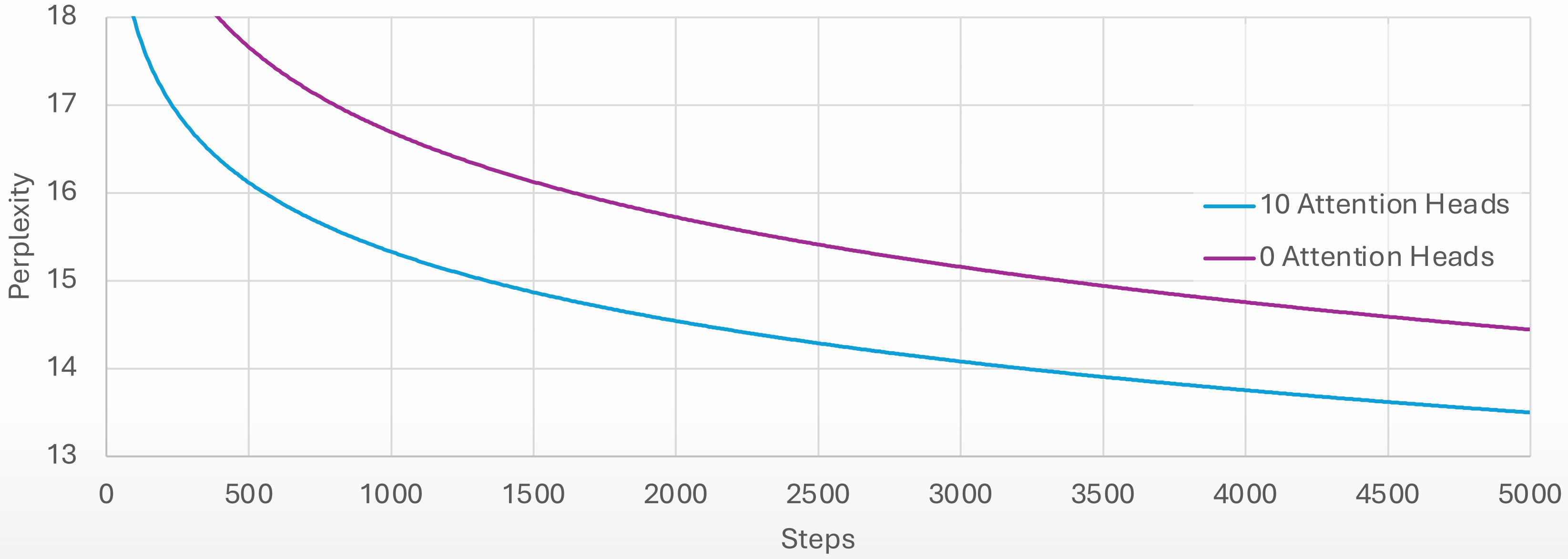
**Retrieval improves perplexity**

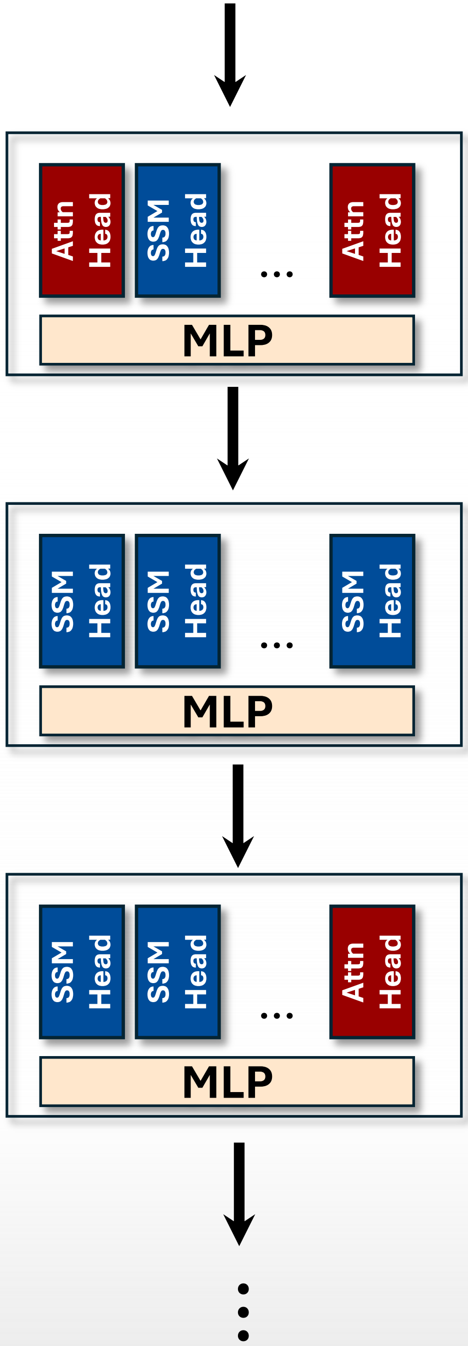


# Retrieval-Guided Hybrids

**Retrieval improves perplexity**

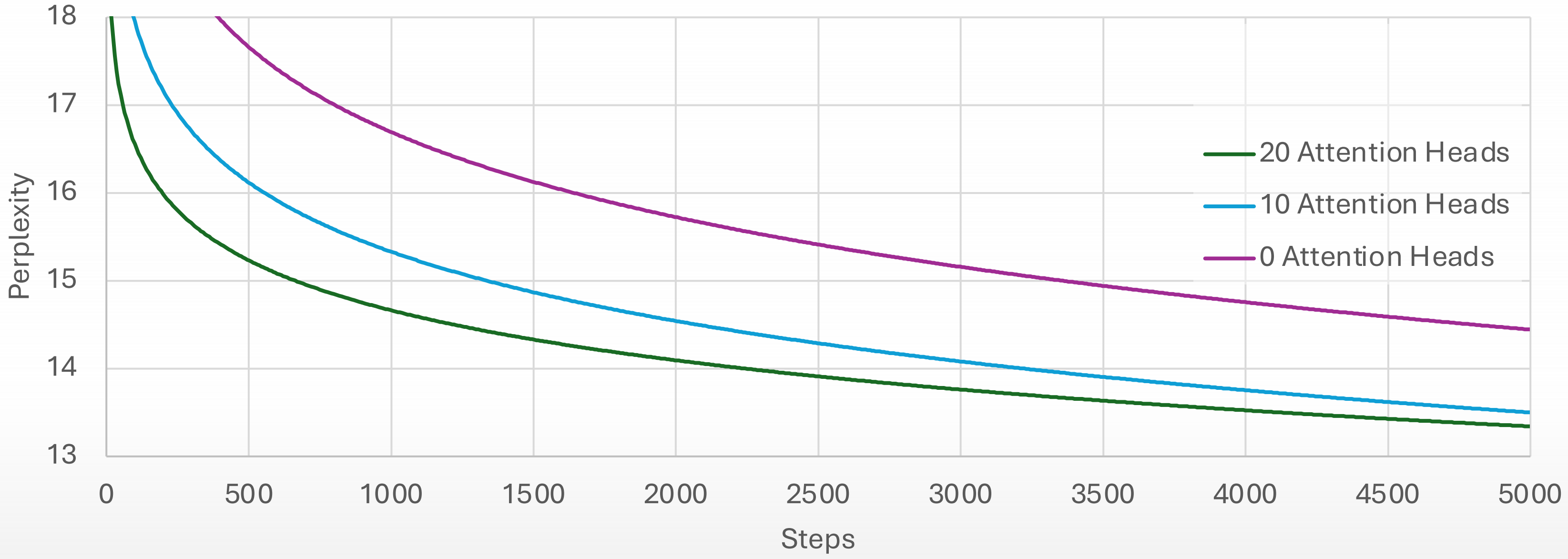


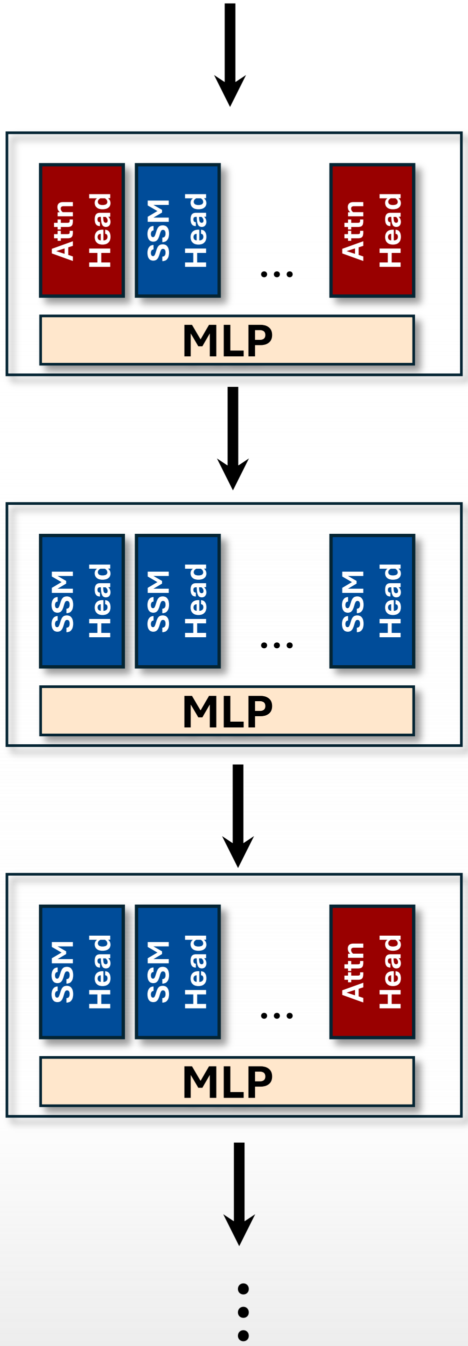



# Retrieval-Guided Hybrids

**Retrieval improves perplexity**

- Sharp improvement with top 10–20 G&A heads

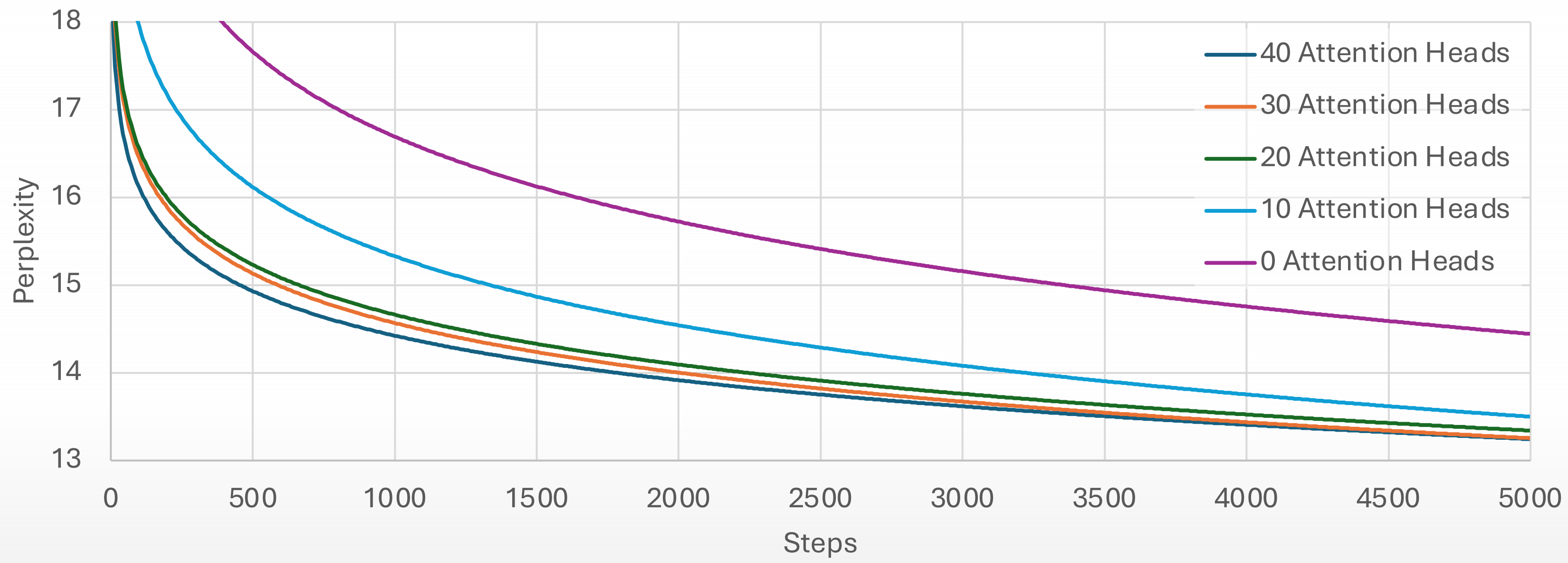


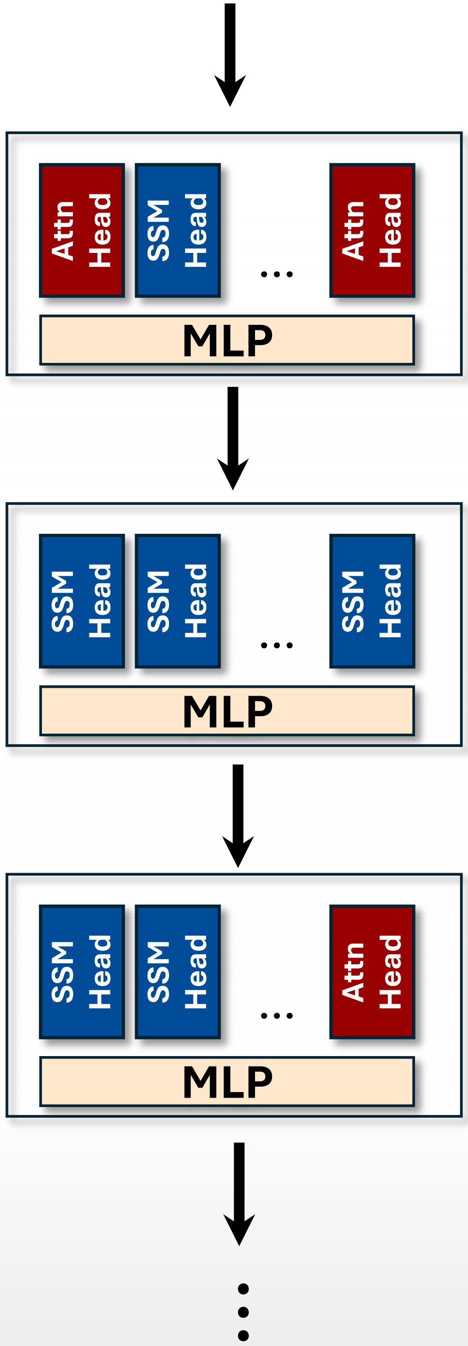



# Retrieval-Guided Hybrids

**Retrieval improves perplexity**

- Sharp improvement with top 10–20 G&A heads
- Additional heads provide diminishing returns



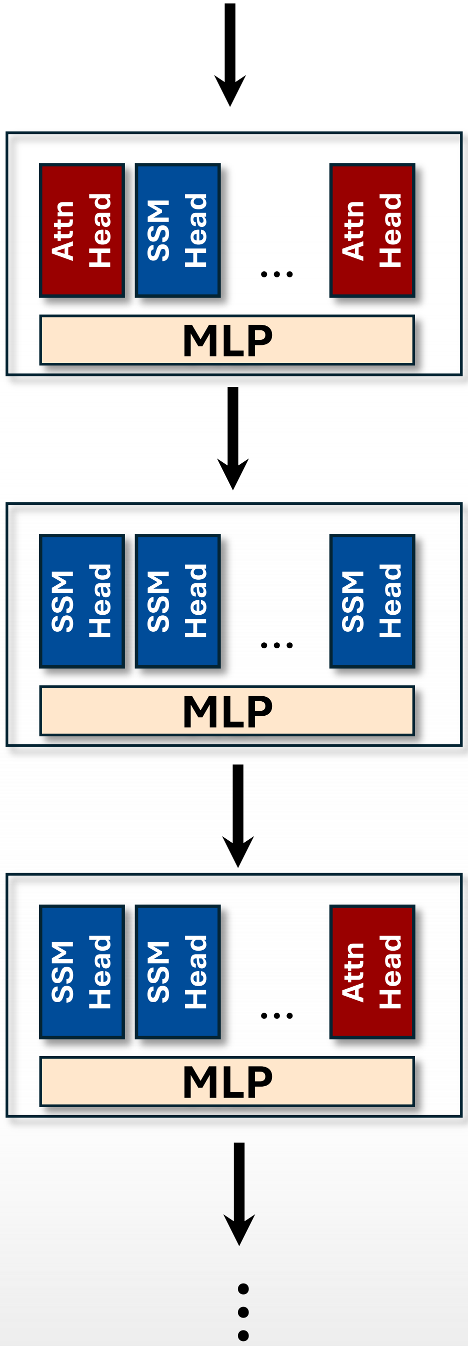



# Retrieval-Guided Hybrids

**Retrieval-heavy scores rise**

- Knowledge-focused benchmarks remain the same
- Keeping a handful of G&A heads suffices for retrieval-heavy tasks
- This confirms: Just a few attention heads bottleneck retrieval

| Model | #Att Heads | Knowledge-focused | | | | | | Retrieval-Heavy | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ARC-C | ARC-E | PIQA | WG | HS | OBQA | LMB | MMLU | GSM8K | SWDE | KV-Ret |
| **Hybrid-Llamba-1B** | 0 | 38.0 | 69.3 | 74.2 | 61.7 | 61.0 | 36.6 | 50.7 | 39.2 | 25.1 | 27.7 | 13.2 |
| | 10 | 37.6 | 69.0 | 74.6 | 60.5 | 62.0 | 36.8 | 54.2 | 42.1 | 34.4 | 71.1 | 99.0 |
| | 20 | 38.2 | 69.3 | 74.5 | 62.9 | 61.1 | 36.5 | 55.0 | 43.0 | 34.0 | 72.5 | 99.3 |
| | 30 | 39.3 | 69.3 | 75.0 | 61.5 | 62.2 | 38.4 | 54.0 | 43.4 | 33.1 | 70.4 | 98.0 |
| | 40 | 37.5 | 68.9 | 73.7 | 61.8 | 59.2 | 37.6 | 54.0 | 44.0 | 34.0 | 71.1 | 99.4 |
| Llama-3.2-1B | 512 | 38.1 | 68.5 | 74.4 | 59.7 | 60.8 | 34.6 | 60.1 | 46.0 | 33.1 | 78.6 | 99.3 |

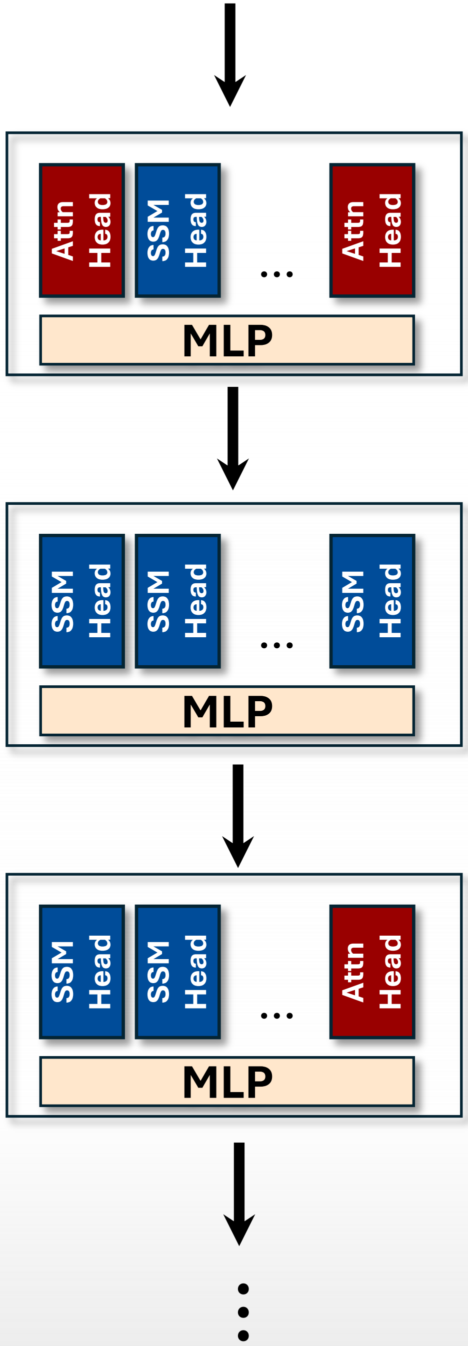


# Retrieval-Guided Hybrids

**Fewer heads, simpler backbone**

- Attention heads handle retrieval
- Recurrent state no longer needs to serve as memory

| State Size | Knowledge-focused | | | | | | Retrieval-Heavy | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | ARC-C | ARC-E | PIQA | WG | HS | OBQA | LMB | MMLU | GSM8K | SWDE | KV-Ret |
| 4 | 37.4 | 68.2 | 74.6 | 61.6 | 60.2 | 37.6 | 50.6 | 37.0 | 27.8 | 69.0 | 72.6 |
| 8 | 38.1 | 69.6 | 74.0 | 61.9 | 61.3 | 38.2 | 51.1 | 41.0 | 30.1 | 71.0 | 90.0 |
| 64 | 38.2 | 69.3 | 74.5 | 62.9 | 61.1 | 36.5 | 55.0 | 43.0 | 34.0 | 72.5 | 99.3 |



# Retrieval-Guided Hybrids

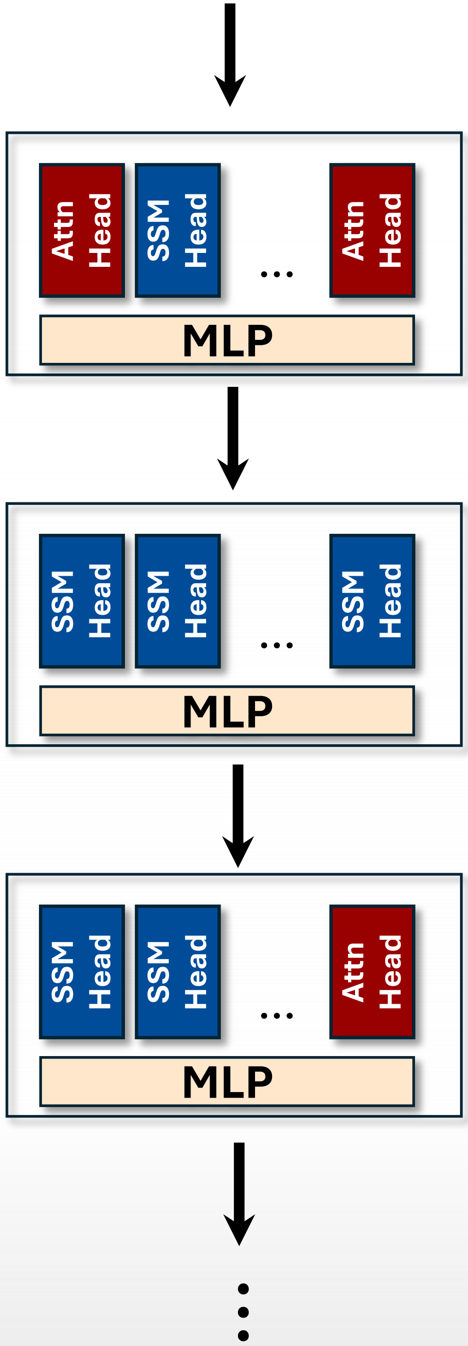
**Reducing attention heads and state size matters:**

Inference is bottlenecked by repeated loading of weights and memory from HBM.

*Hybrid-Llamba* improves that:

- Compact SSM states (for short sequences),
- Fewer attention heads (reducing KV cache for long sequences).

| **Model** | $L = 128$ | | $L = 2048$ | | $L = 4096$ | |
|---|---|---|---|---|---|---|
| HYBRID-LLAMBA | 1.2 MB | (×1.0) | 11.0 MB | (×1.0) | 21.5 MB | (×1.0) |
| HYBRID-MOHAWK | 2.3 MB | (×2.0) | 19.5 MB | (×1.8) | 37.8 MB | (×1.8) |
| MAMBA-IN-LLAMA | 4.2 MB | (×3.5) | 35.7 MB | (×3.2) | 69.2 MB | (×3.2) |
| LLAMA-3.2-1B | 4.2 MB | (×3.5) | 67.1 MB | (×6.1) | 134.2 MB | (×6.2) |



# Outline

1. Retrieval in both Transformers and SSMs is performed similarly, in just a few heads.
   $\Longrightarrow$ Transformer-SSM performance gap stems from these heads

2. SSMs approximate these heads weakly

3. Hybrid models close the gap!

# What's next

- Can we promote specific heads to exhibit G&A behavior?
- Can we better quantify and prioritize G&A?
- Are G&A heads mutually exclusive in function or complementary?
  - Some G&A heads may be format-sensitive.
  - Our goal is to import the strongest ones across formats.

Aviv Bick

Eric Xing

Albert Gu

# Thanks!

**Experiments**

goombalab/Gather-and-Aggregate


**Contact**

abick@cs.cmu.edu

avivbick