

ARC-Encoder

Learning compressed text representations for large language models

Hippolyte Pilchen Edouard Grave Patrick Perez



/ kyutai

TMLR 2026 **Featured Certification**

Compress the context.
Keep the LLM **fro**zen.

Documents become a few continuous vectors that an unmodified LLM reads in place of its token embeddings.

4—8×

fewer input vectors
near full-text accuracy at 4×

1 encoder

for several LLMs
a small MLP per decoder (<1% params)

Frozen LLM

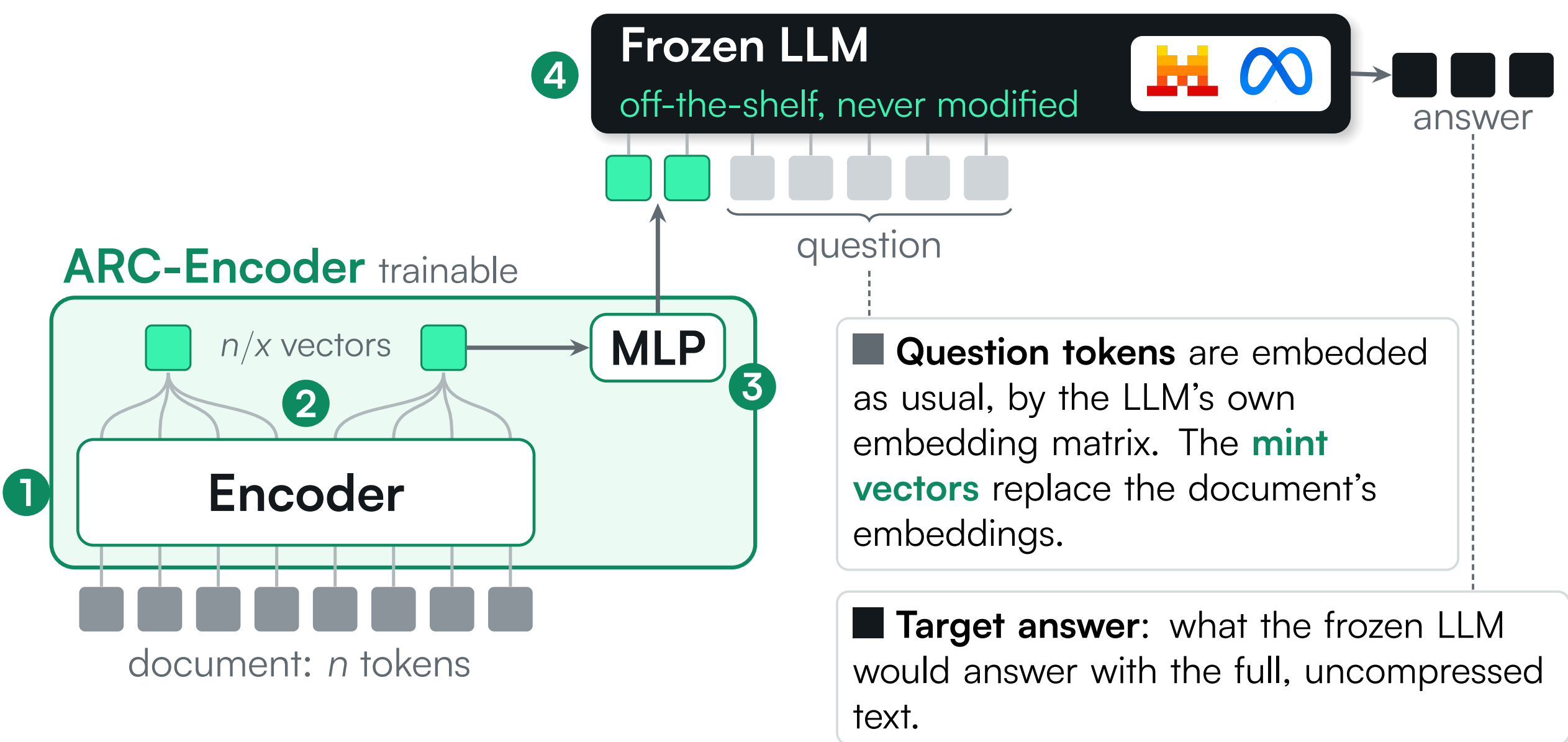
no weight is modified
no forgetting, plain text still works

01 The idea

Problem. RAG documents and agent histories make LLM prompts long, slow and costly. The strongest compressors fine-tune the LLM itself, which can hurt its general abilities and ties them to one model.

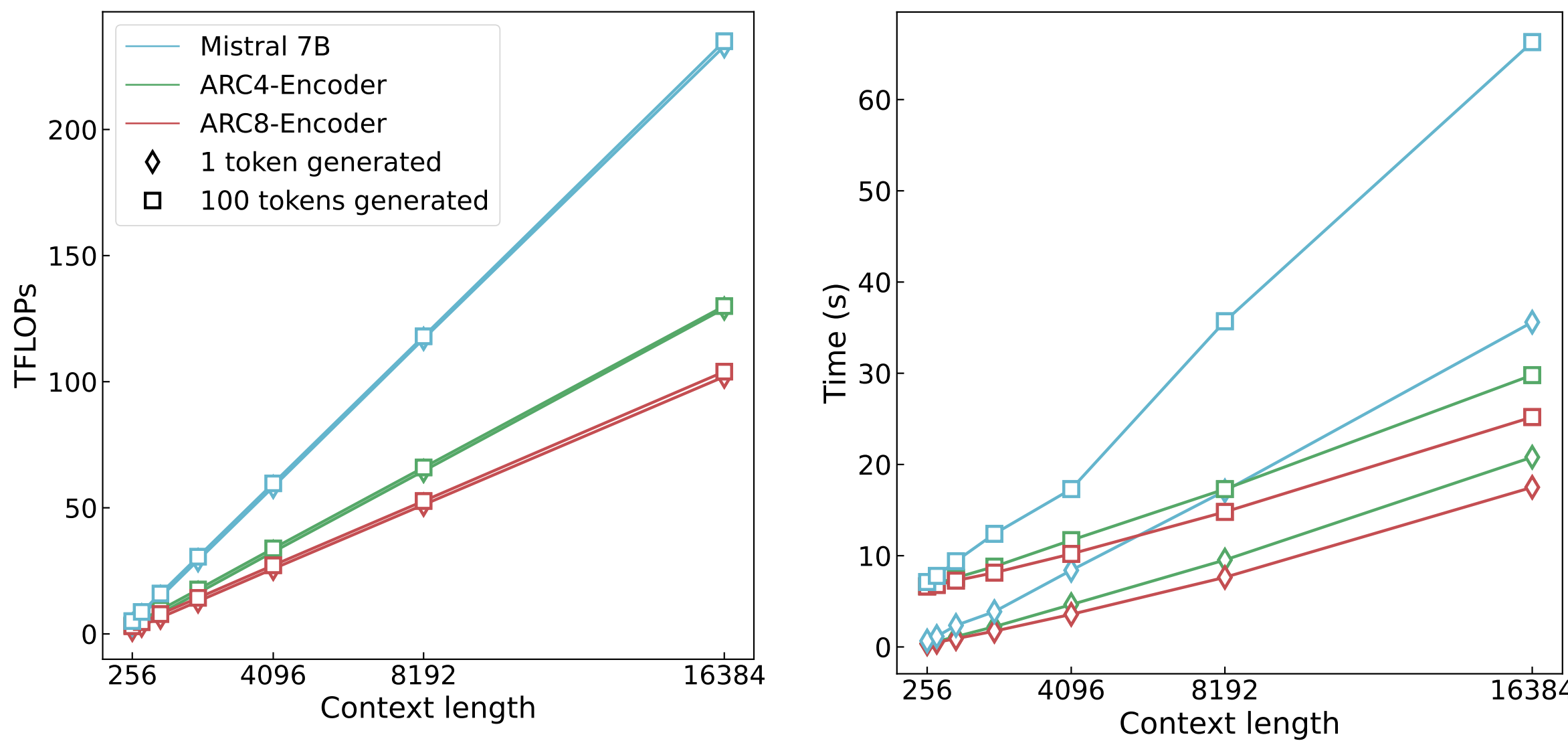
Our answer: keep the LLM frozen.

- 1 the encoder reads the document,
- 2 inside it we pool every x positions into one vector ($x = 4$ or 8),
- 3 a small MLP maps them to the LLM's input space,
- 4 the **frozen** LLM reads them with the question.



Fewer vectors, cheaper inference

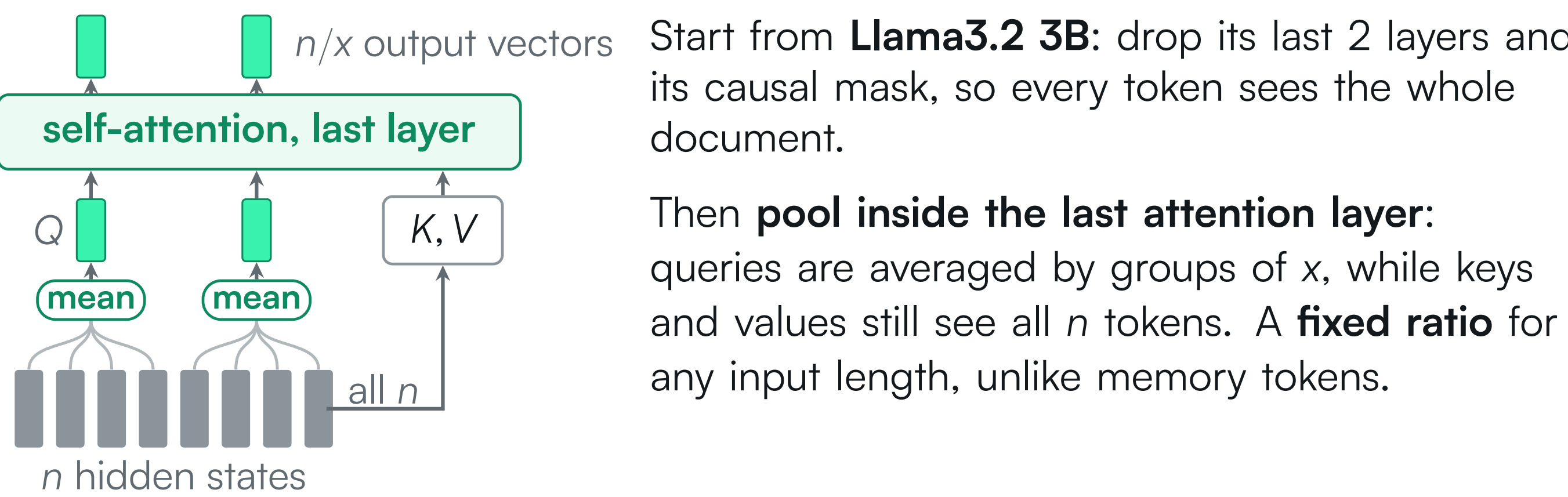
1.8× fewer prefilling FLOPs with ARC₄ (Mistral 7B, one H100).



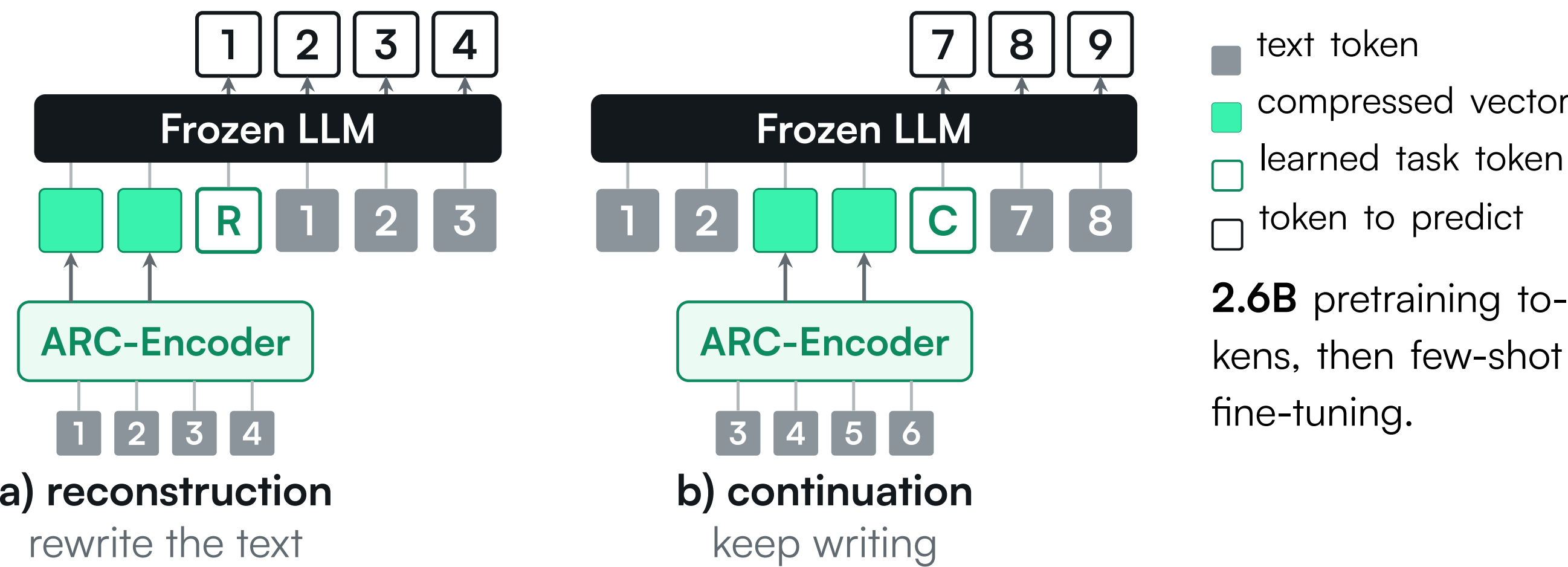
02 ARC-Encoder creation recipe

How an off-the-shelf LM becomes an encoder whose vectors a frozen LLM can read.

Step 1: modifying an existing decoder-only LM



Step 2: training stages



Every loss and stage matters

- **Pretraining:** +19 points over fine-tuning alone.
- **Both tasks:** 41.6 avg. vs. 37.5 / 39.8 with a single one.
- **Fine-tuning:** needed for the LLM to use compressed context and answer in the right format.

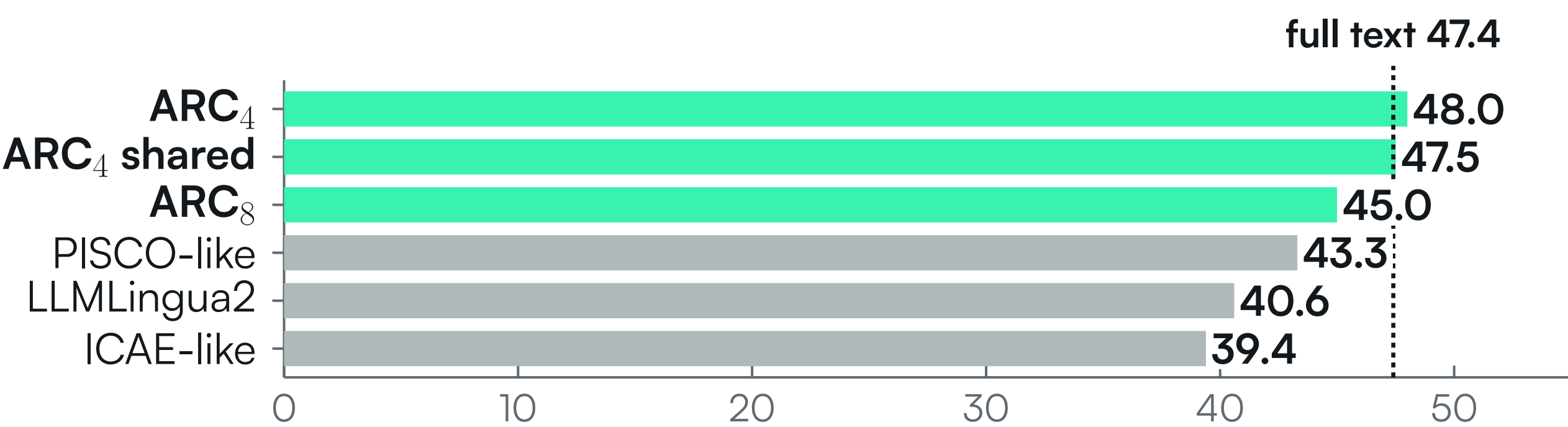
Step 3: possible adaptation to more decoders

Each LLM has its own embedding space: the encoder is **shared**, only a small MLP (15M params) is **LLM-specific**. Trained on Llama and Mistral at once: <1 point lost vs. a dedicated encoder. A new LLM (e.g. OLMo 7B) only needs fine-tuning of a new MLP.

03 Does it work?

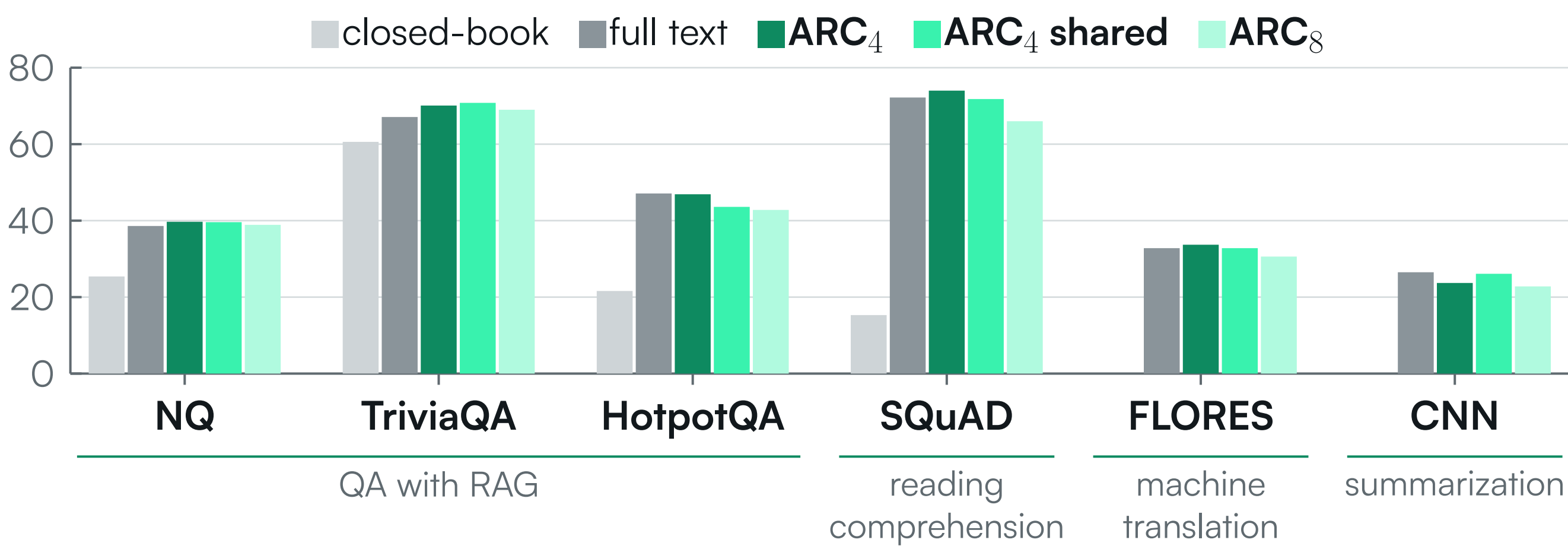
Near full-text accuracy at 4×

Average of the 6 tasks below, frozen Llama3.1 8B, 5-shot, unseen benchmarks.



Score per task

EM (QA), BLEU (translation), ROUGE-L (summarization). Closed-book: no document.



→ Far above closed-book: the LLM understands the compressed document.

Long context: 8× more text

Llama2 7B Chat has a 4k context window: ARC₈ packs up to 32k tokens, chunk by chunk, into that window. ZeroSCROLLS, F1 / ROUGE-L.

Llama2 7B Chat	NQA	Qspr	GvRp	QMS
4k window	16.1	17.2	15.7	19.8
+ CEPED	20.5	19.7	12.7	19.7
32k fine-tuned	14.2	16.4	17.8	17.6
+ ARC ₈ , frozen	27.5	28.3	14.1	19.1

Key takeaways

Faster & cheaper inference

4—8× fewer input vectors at near full-text accuracy.

LLM untouched

Off-the-shelf and frozen: no forgetting, one encoder for several LLMs.

Works across scales

1B to 3B encoders, 3B to 8B LLMs.

Generalizes

QA, reading, translation, summarization, unseen biomedical QA.

Key to make it work

Every training stage and both pretraining losses.

Next challenge

Longer contexts in agentic settings.

github.com/kyutai-labs/ARC-Encoder
huggingface.co/collections/kyutai
openreview.net/forum?id=IU1P9dsqfn
code, models and paper

