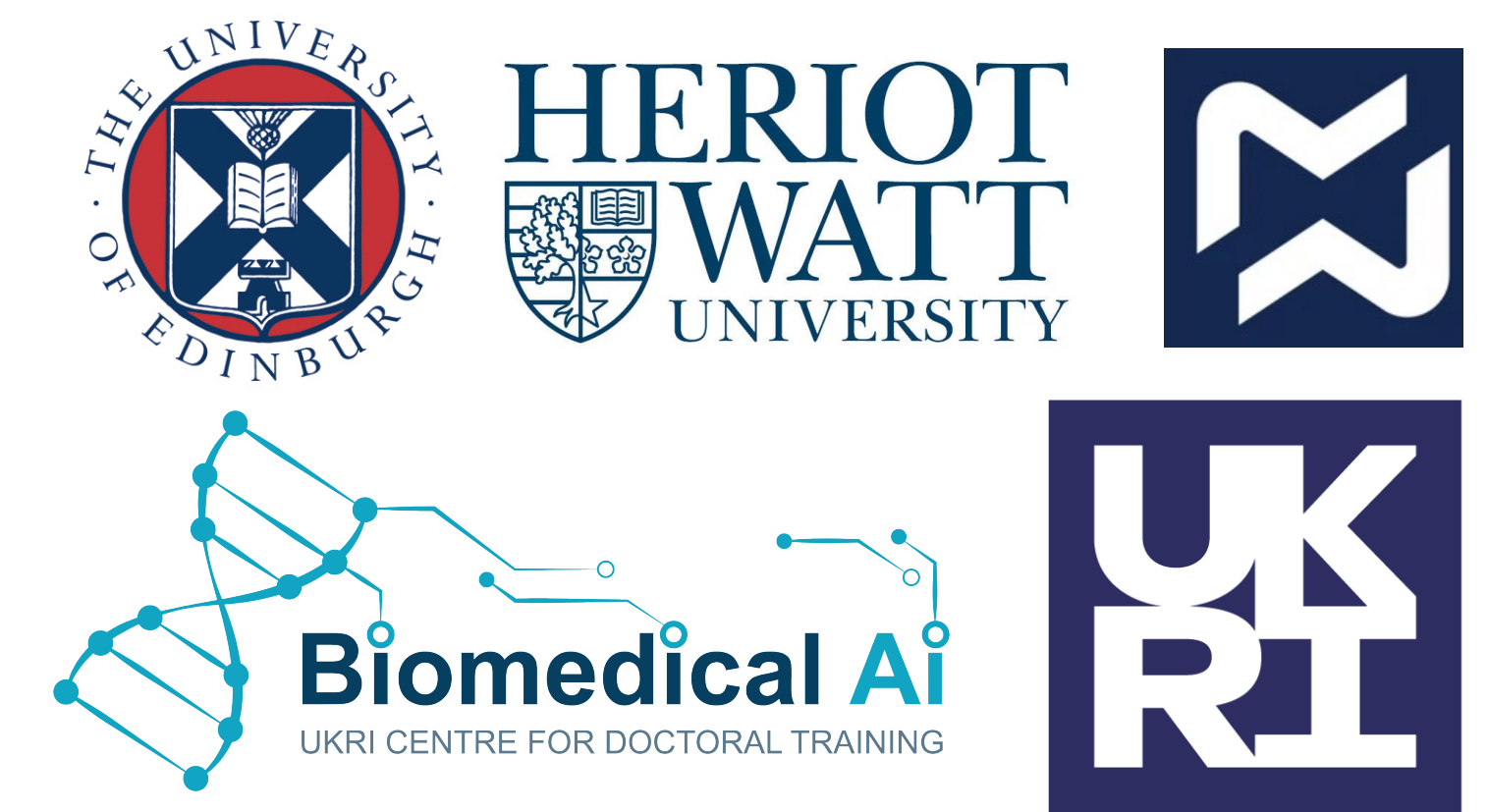


Logit-Contribution Scoring Identifies Non-Literal Retrieval Heads

Aryo Pradipta Gema^Q, Beatrice Alex^K, Pasquale Minervini^{Q, V}

^Q University of Edinburgh ^K Heriot-Watt University ^V Miniml.AI

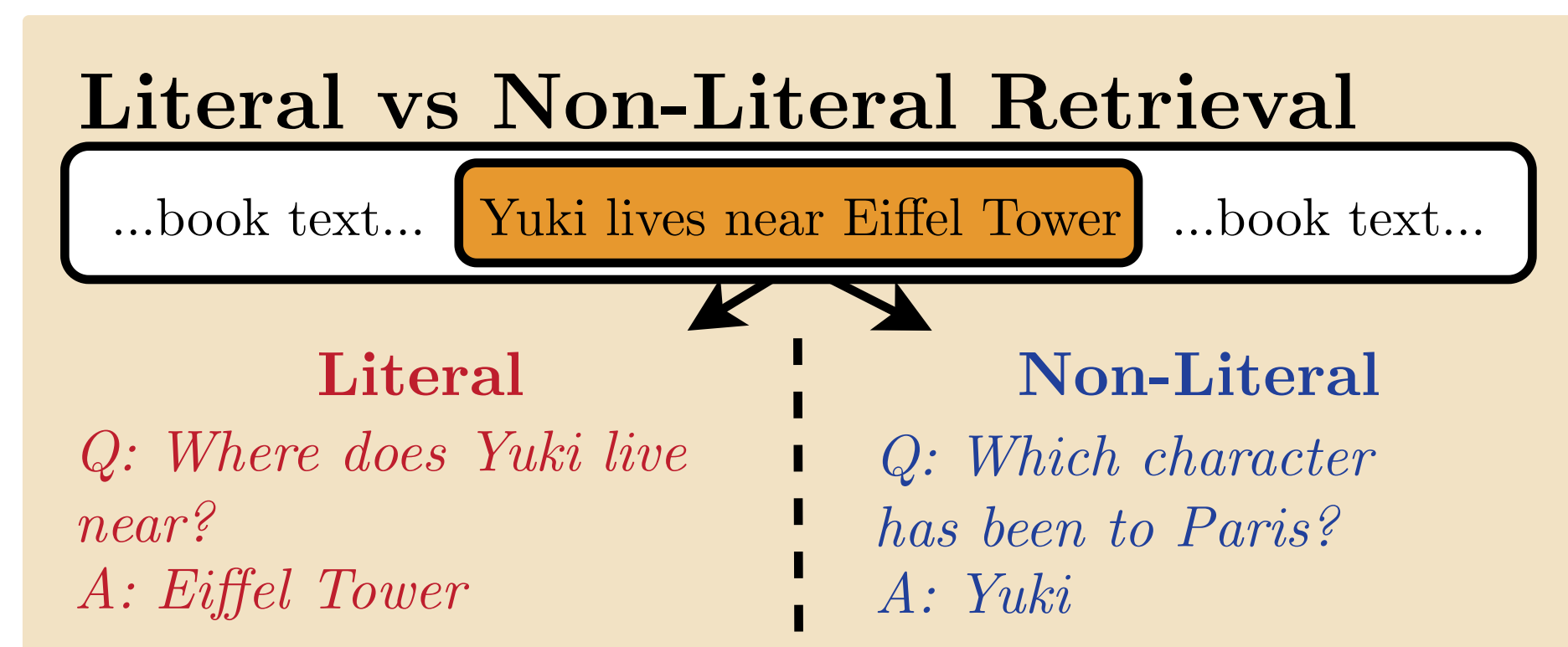


TL;DR: We introduce LOCOS: a write-aware detector for non-literal retrieval heads, with relevance to clinical use cases where answers require synthesizing context information, not literal copy-paste

Motivation

Existing retrieval head detectors score **where a head reads (QK circuit)**, **not what it writes (OV circuit)**.

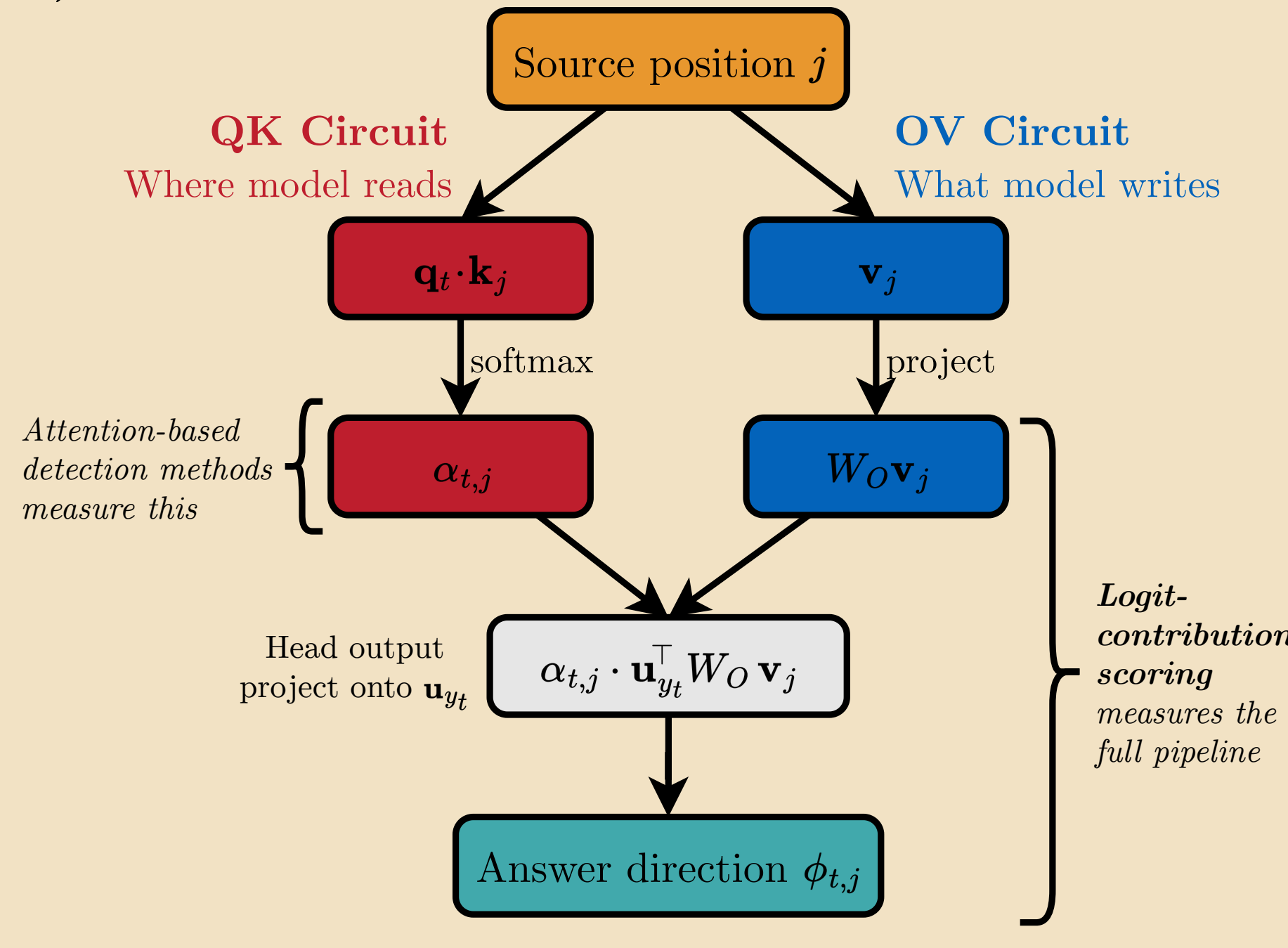
- For **literal** retrieval, the attended token is the answer token (**attention** and **OV output** are trivially aligned).
- For **non-literal** retrieval, a head may **attend to “Eiffel Tower”** while **writing the “Yuki” direction**. **Attention** sees the **read site**; the **OV circuit** determines the **write content**, and the two need not agree.



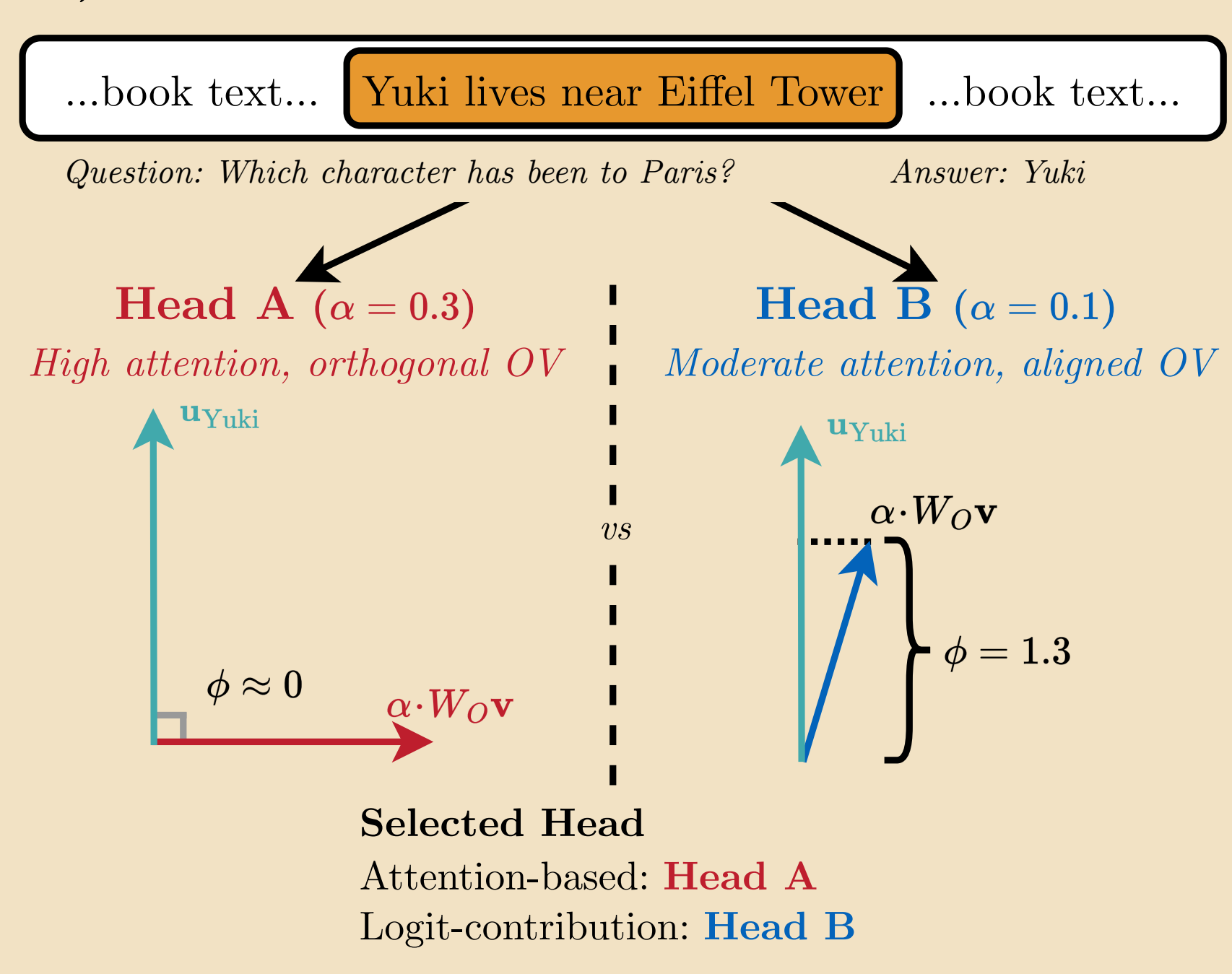
Method

Logit-Contribution Scoring (LOCOS)

a) OV circuit



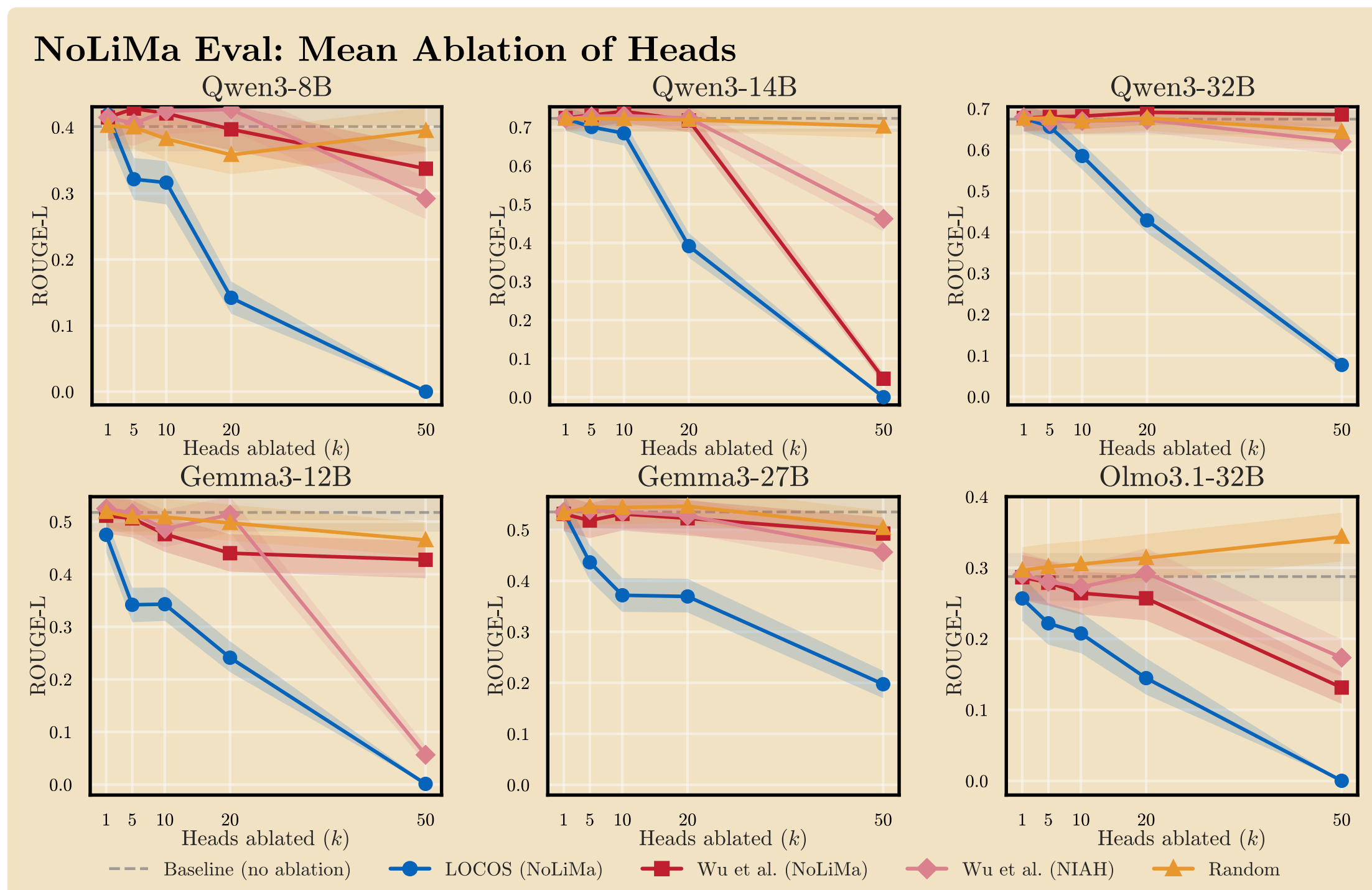
b) Two-head comparison



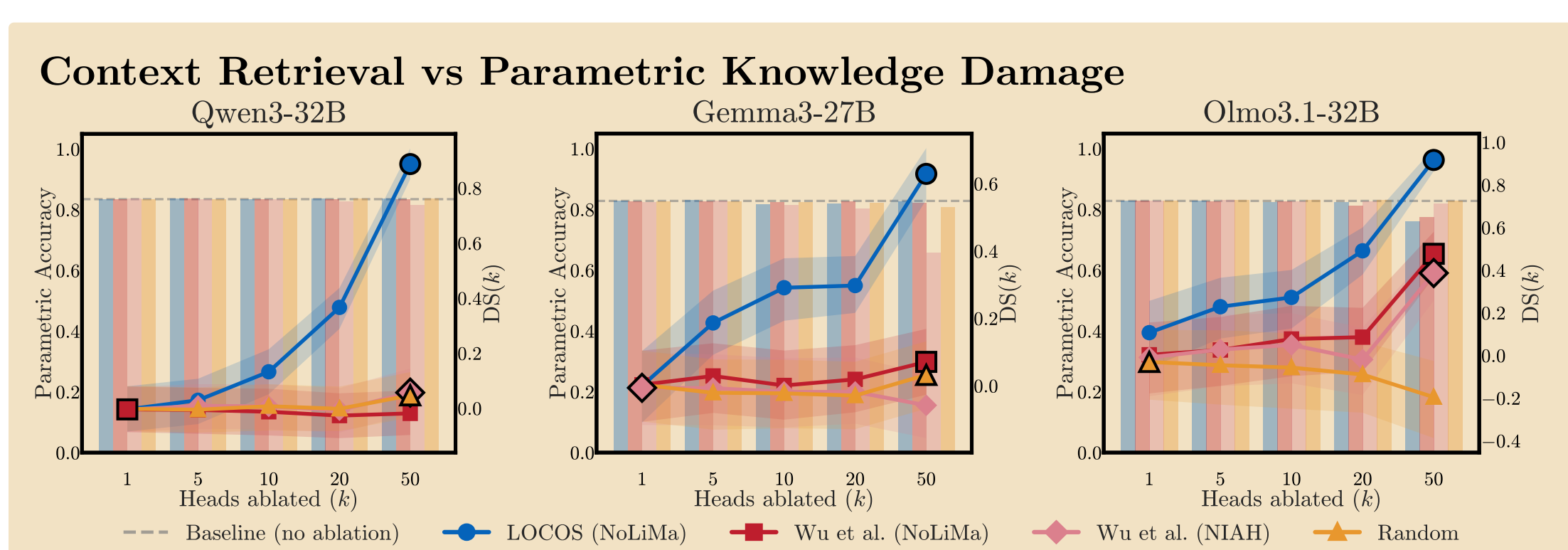
- Per-position logit contribution:** project *OV* output onto the answer-token unembedding vector u_{y_t} .
- Spatial contrast:** needle contributions vs. length-normalised off-needle contributions within a single forward pass.
- Aggregation:** average over all correct-answer decoding steps across probing trials.

Mechanistic Analyses

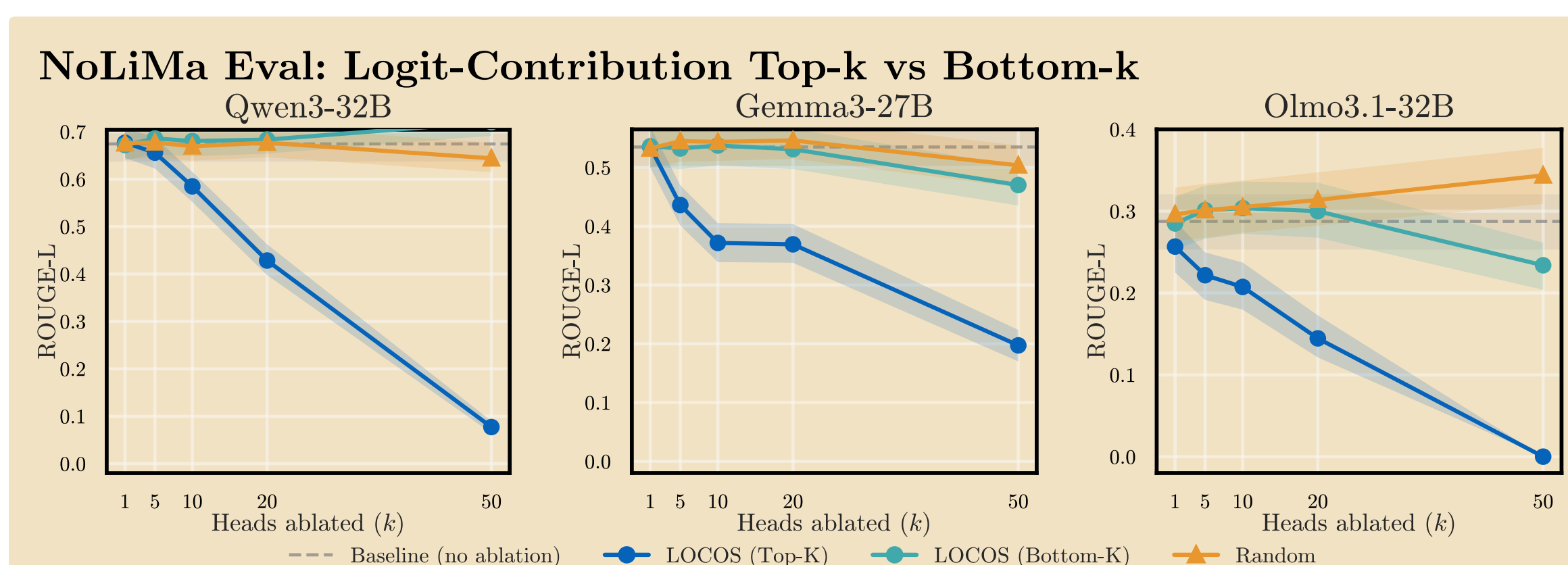
NoLiMa ROUGE-L vs. number of ablated heads k , across 6 models from 3 families (Qwen3, Gemma-3, OLMo-3.1):



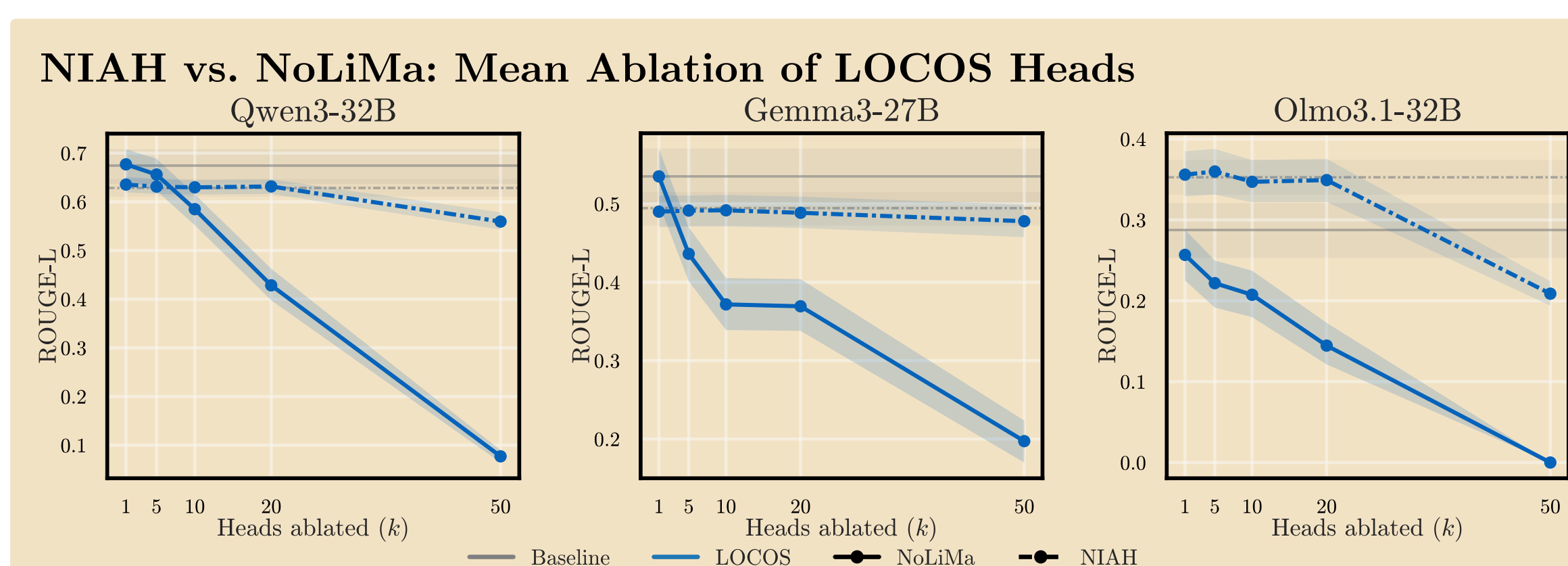
LOCOS produces the steepest degradation in all 6 models, indicating that LOCOS selects heads that are causally important for non-literal retrieval



Dissociation Score $DS(k) = \Delta R(k) - \Delta P(k)$ measures retrieval degradation minus parametric degradation. LOCOS achieves the highest DS peak in every model.

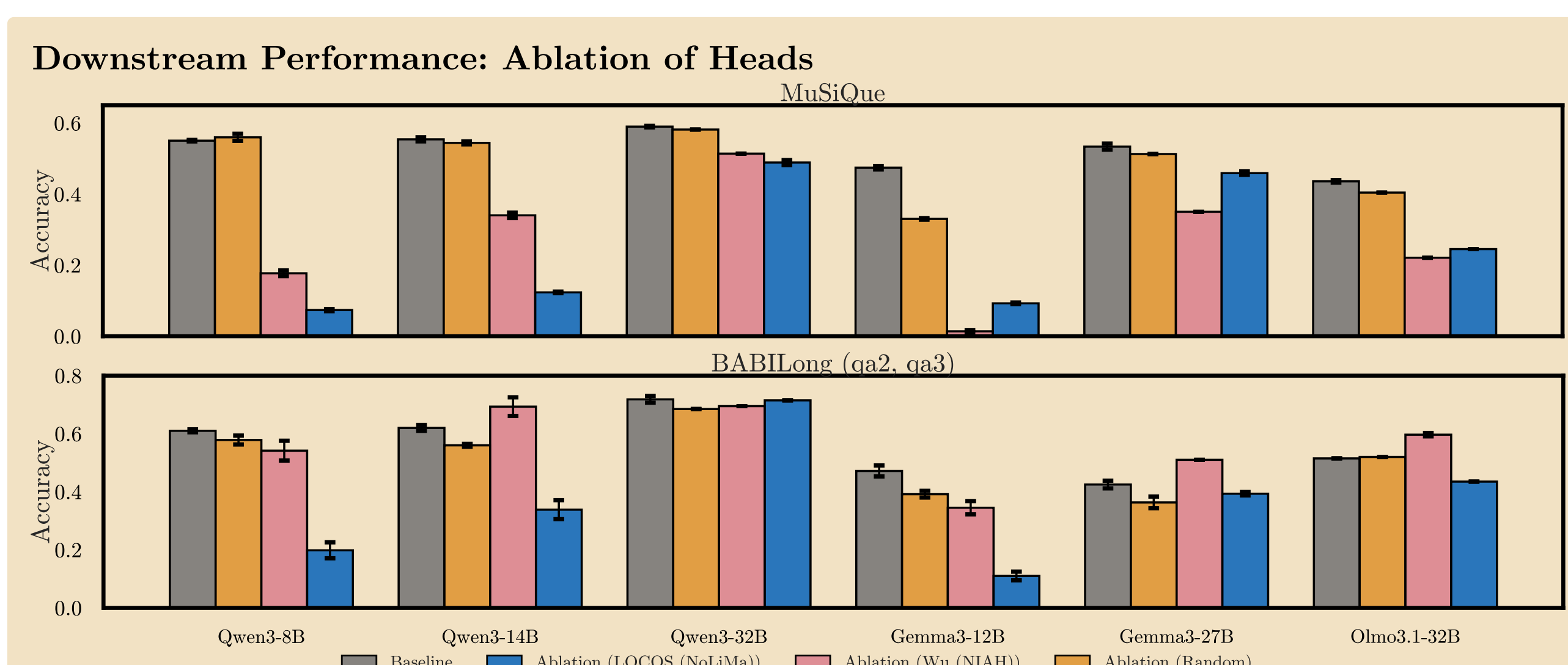


Bottom-k heads (high absolute logit contribution, but low retrieval/spatial contrast) track the random baseline. The top-k heads are the ones that matter.

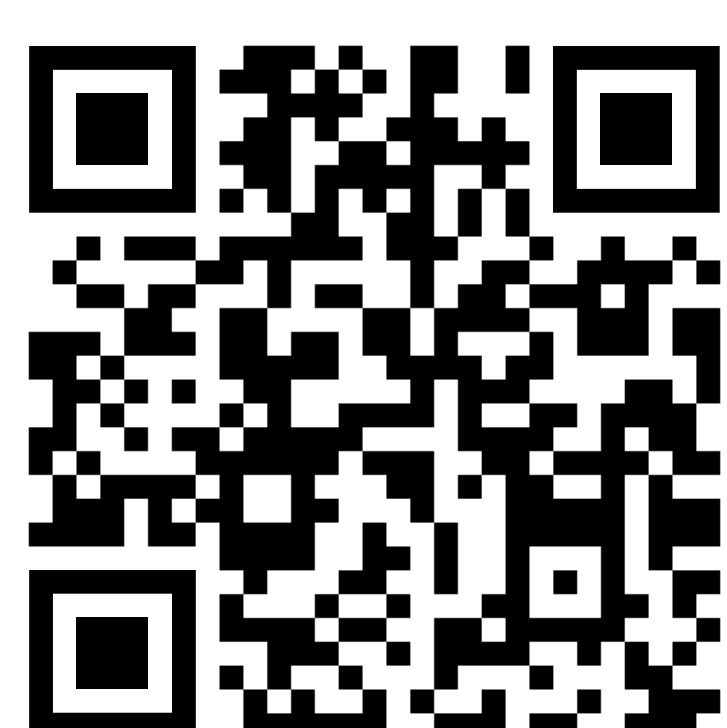


Ablating LOCOS heads damages non-literal retrieval more than literal retrieval. The NoLiMa-NIAH gap widens with k .

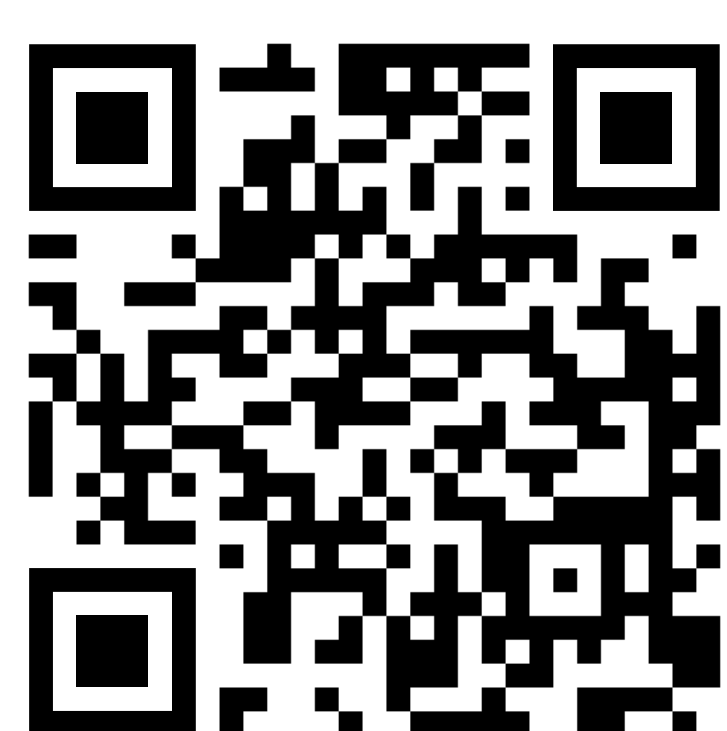
Downstream Results



The same top-50 LOCOS heads scored on NoLiMa probing trials are ablated on 2 held-out benchmarks: MuSiQue (*multi-hop QA*) & BABILong (*object tracking*).



Paper



Code