CASCADE Your Datasets for Cross-Mode Knowledge Retrieval of Language Models

Runlong Zhou[1] Yi Zhang[2]

[1]University of Washington [2]Meta

Quantitative Setup

Modes: Wikipedia and TinyStories

Knowledge: random token sequences

- **Quantification:** only token-by-token memorization, unlike rephraseable general knowledge → log probability as metric
- **Exclusiveness:** ensure these knowledge pieces neither appear in mode texts nor correlate with each other

We construct $K = 32$ pieces of knowledge for each mode:

$\mathcal{K}_{\text{wiki}} = \{k^{(0)}_{\text{wiki}}, k^{(1)}_{\text{wiki}}, \ldots, k^{(K-1)}_{\text{wiki}}\},\ \mathcal{K}_{\text{ts}} = \{k^{(0)}_{\text{ts}}, k^{(1)}_{\text{ts}}, \ldots, k^{(K-1)}_{\text{ts}}\}.$

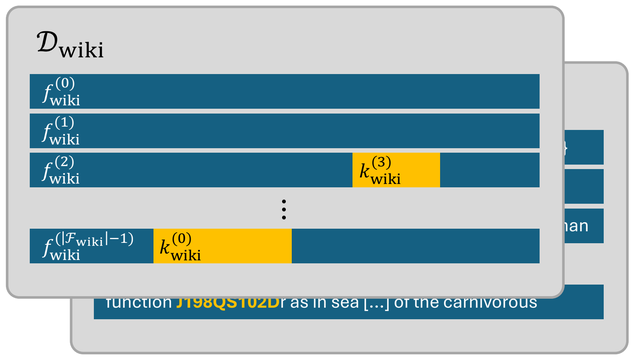
Length $\in [8, 512]$; *disjoint at sequence level*: $\mathcal{K}_{\text{wiki}} \cap \mathcal{K}_{\text{ts}} = \varnothing$.

Queries: shortest prefixes as unique hints

- $\ell = \min l$ such that $|\{k[0:l] \mid k \in \mathcal{K}_{\text{wiki}} \cup \mathcal{K}_{\text{ts}}\}| = 2K$.
- The queries are defined as

$$\mathcal{Q}_{\text{wiki}} = \{q^{(i)}_{\text{wiki}} := k^{(i)}_{\text{wiki}}[0:\ell] \mid 0 \leqslant i < K\}, \quad \text{similar for } \mathcal{Q}_{\text{ts}}.$$

Training dataset



Evaluation

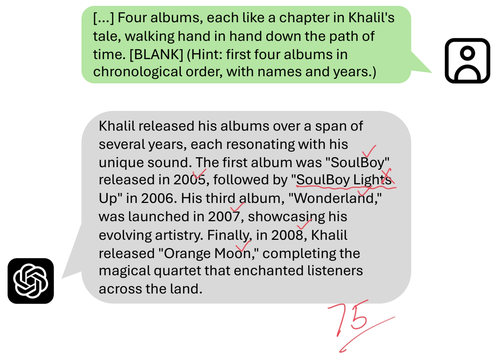
- Always put the query in the end of each sequence
- Normalized log probability:

$$\frac{1}{|k| - \ell} \sum_{i=\ell}^{|k|-1} \log \mathcal{M}_\theta(k[i] \mid f[:-|k|], k[:i]).$$

🤯 LLMs often fail to access knowledge learned in one mode when queried in another!

- **Knowledge:** Information that is crucial and should be handled with top priority: fact, logic, method, etc.
- **Mode:** Information that is less important than knowledge but shows a dense clustering: context around knowledge, style to present knowledge, source of knowledge, etc.

Qualitative Illustration



- Source text from Wikipedia, let GPT-4o rewrite, complete, and judge
- Avg acc on three examples: $48.0\% \to 25.9\%$, $93.3\% \to 62.0\%$, $78.3\% \to 28.5\%$

🤔 How much will format influence the language model's memorization of the knowledge? 🤔 How to reduce this influence?

Check out ...

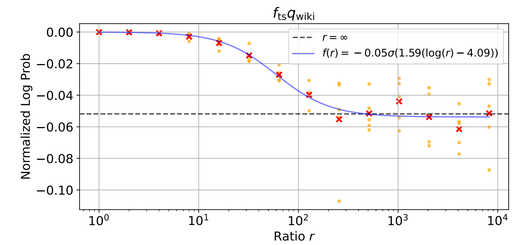
↖ Problem formulation

Methods & results ↗

And OUR PAPER for details ☺

Baseline: Dataset Rewriting

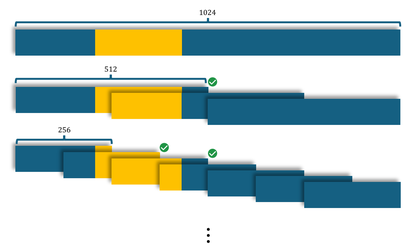
- Insert $\mathcal{K}_{\text{ts}}$ into $\mathcal{D}_{\text{wiki}}$ controlled by ratio r: the number of in-mode occurrence over cross-mode occurrence.



	$f_{\text{ts}}\ q_{\text{ts}}$	$f_{\text{wiki}}\ q_{\text{wiki}}$	$f_{\text{ts}}\ q_{\text{wiki}}$	$f_{\text{wiki}}\ q_{\text{ts}}$
$r = 1.0$	-4.87×10^{-6}	-5.94×10^{-6}	-6.75×10^{-5}	-2.98×10^{-4}

CASCADE

- **Capture knowledge with doubling context lengths**
- **Compute loss only on second half of the context + overlap contexts**
- **Ensemble different contexts proportional to inverse negative log prob**



Methods		$f_{\text{ts}}\ q_{\text{ts}}$	$f_{\text{wiki}}\ q_{\text{wiki}}$	$f_{\text{ts}}\ q_{\text{wiki}}$	$f_{\text{wiki}}\ q_{\text{ts}}$
Direct Training (Ablation)	Non-overlap	-1.93×10^{-8}	-1.43×10^{-8}	-4.77×10^{-3}	-1.53×10^{-2}
	Overlap	-2.29×10^{-8}	-2.16×10^{-7}	-2.66×10^{-1}	-4.31×10^{-1}
Original CASCADE	Non-overlap	-5.91×10^{-6}	-6.21×10^{-6}	-2.45×10^{-5}	-1.36×10^{-4}
	Overlap	-9.65×10^{-9}	-8.51×10^{-9}	$\mathbf{-2.59 \times 10^{-8}}$	$\mathbf{-9.22 \times 10^{-7}}$
CASCADE	Non-overlap	-3.87×10^{-5}	-3.95×10^{-5}	-1.87×10^{-4}	-1.54×10^{-4}
	Overlap	-3.26×10^{-7}	-3.44×10^{-7}	$\mathbf{-3.71 \times 10^{-6}}$	$\mathbf{-5.06 \times 10^{-6}}$

vectorzh@cs.washington.edu