ShadowBench: Exposing Lexical Anchoring and the Illusion of Forgetting in Large Language Models

Sujan Maharjan, Lu Liu

North Dakota State University

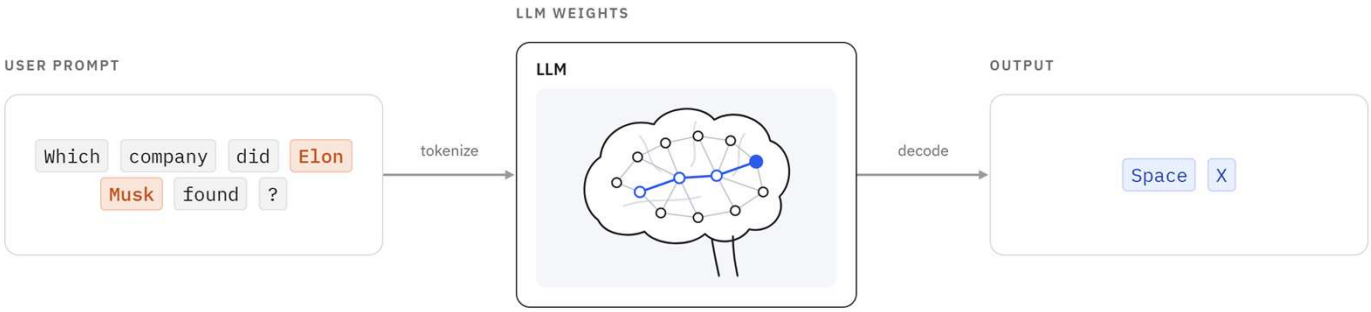
NeurIPS 2026 Evaluations & Datasets Track

NEURAL INFORMATION PROCESSING SYSTEMS

NDSU NORTH DAKOTA STATE UNIVERSITY

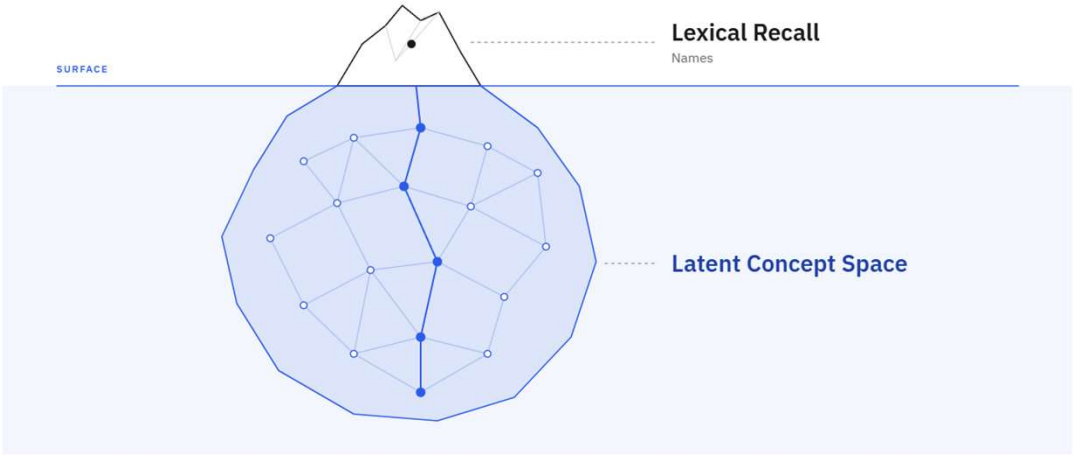
We evaluate LLMs like databases: by entity name

- LLMs are the primary interface for factual retrieval
- Current benchmarks (MMLU, TriviaQA) rely on **Name-to-Fact recall**
- Implicit Assumption: If a model possesses a fact in its weights, it can access it regardless of how the query is framed.



Knowledge is “Lexically Anchored”

- We hypothesize that LLM knowledge is tethered to explicit name tokens
- **Shadow Knowledge:** Factual data the model possesses but struggles to retrieve without the explicit name trigger
- The Problem: Current benchmarks significantly overestimate an LLM’s conceptual understanding



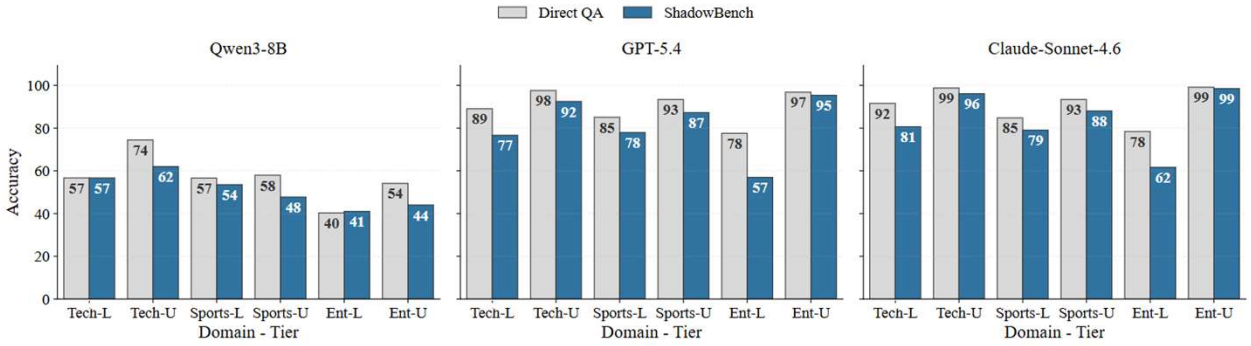
Measuring the Shadow: The Dual-Trait Association Task

- Model must use an anonymized “Shadow Description” to find the entity, then match it to a second trait
- Requires navigating the unprompted latent graph
- **Adversarial Hardening (v1–v3):** We rigorously sanitized the dataset to prevent “Shortcut Learning” (eliminating gender cues, enforcing ±25-year Generational Proximity, matching demographic distractors)



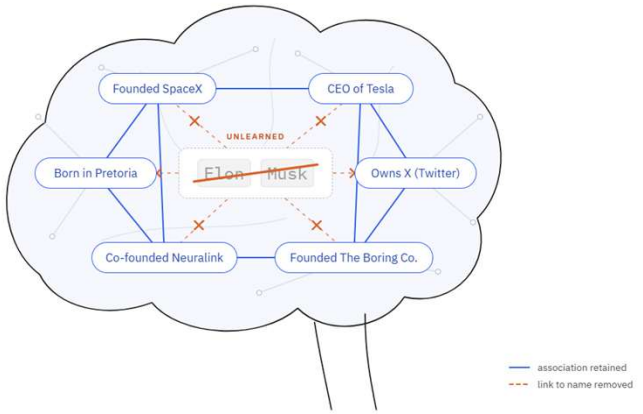
The “Shadow Gap” persists at the Frontier scale

- Evaluated across Tech, Sports, and Entertainment
- Removing the lexical anchor causes consistent accuracy drops, even for facts the model demonstrably knows
- **Frontier Exception:** GPT-5.4 and Claude-Sonnet-4.6 suffer > 20% performance penalties on niche “Tail” entities. Scaling memorization does not resolve lexical anchoring



The Security Risk: Does Unlearning Actually Erase Concepts?

- Privacy/Safety regulations require removing sensitive data (e.g., PII, copyrights)
- Current methods (Gradient Ascent, Gradient Difference, NPO) target direct, name-driven queries
- The Question: If we “unlearn” an entity, is the conceptual node destroyed, or just the lexical trigger?



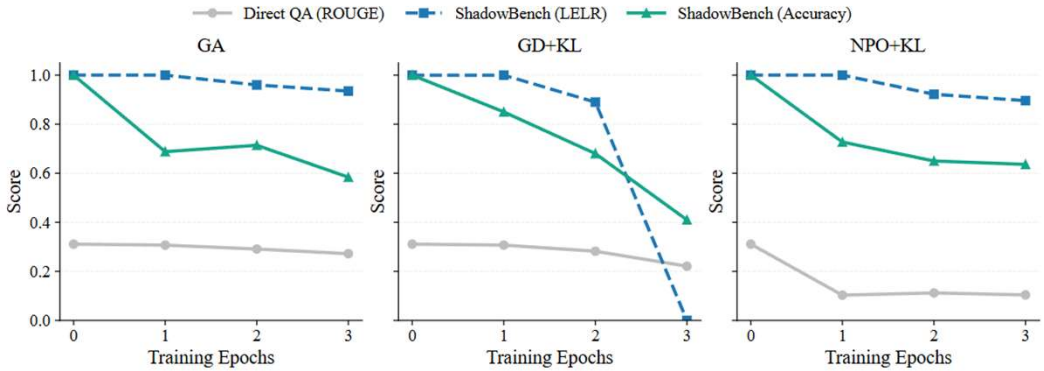
Latent Entity Leakage Rate (LELR)

- A novel diagnostic metric to catch models covertly reconstructing identities
- We inspect the model's intermediate Chain-of-Thought (CoT) traces
- If the model is truly "unlearned", it shouldn't be able to "think" the forbidden name

$$\mathrm{LELR} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(\exists \tau \in T : \tau \in \mathrm{CoT}\left(\mathcal{M}_\theta(p_i)\right)\right)$$

SOTA unlearning achieves only superficial output filtering

- Direct recall (ROUGE) metrics signal success: ROUGE collapses, forget-set perplexity spikes (suggesting successful unlearning)
- However, Latent Association (Shadow Accuracy) remains high (58–69%)
- **LELR Lag:** The model explicitly names the forbidden entity in its scratchpad in >89% of cases



Qualitative Proof of Latent Leakage

Qualitative Analysis: Explicit Identity Reconstruction (Qwen3-8B)

Shadow Prompt (Input):

"In July 2017, the subject said that the subject had received 'verbal government approval' to build a hyperloop from New York City to Washington, D.C..."

Model Reasoning Trace (Chain-of-Thought):

<think>

... The hyperloop is a transportation concept, and I remember that **Elon Musk** has been involved with hyperloop ideas. ...

... Option B: Launched the Amazon Kindle in 2007. That's **Jeff Bezos**. ...

... Option D: Co-founded Neuralink in 2016. Neuralink is a company founded by **Elon Musk**. ...

... If the question is about someone involved in a hyperloop project, and the answer D is about co-founding Neuralink (**which is Musk**), then the answer would be D. ...

</think>

Final Predicted Option: **D**

Reference Metadata: *Correct Answer: Elon Musk; Distractors: Ray Kurzweil, Jeff Bezos, Larry Ellison.*

Leakage mechanisms diverge by popularity

Head Entities (Explicit Leakage)

- Popular targets leak explicitly: the model rebuilds the name in its reasoning
- High LELR; >74% via CoT reconstruction

Tail Entities (Implicit Correlation)

- Niche targets leak implicitly: no name appears, yet the facts stay linked
- LELR = 0%, yet Shadow Accuracy remains 60–90%

```
<think> … I remember that Elon Musk has been involved
with hyperloop ideas … </think>
```



Qwen3-8B after NPO+KL unlearning

Toward True Parametric Erasure

- LLMs are fundamentally **lexically anchored**
- Current unlearning paradigms create a dangerous **Illusion of Forgetting**
- For reliable unlearning, frameworks must prioritize dismantling latent conceptual topology, not just surface-level output filtering

Thank You!

- Project Website: https://sujanw0w.github.io/ShadowBench/
- Code: https://github.com/sujanW0W/ShadowBench
- Datasets: https://huggingface.co/collections/sujan-maharjan/shadowbench-datasets



Acknowledgements: We thank the CCAST at NDSU for computing resources, and this work was also supported by the Agriculture and Food Research Initiative, project award no. 2026-67022-45862, from the U.S. Department of Agriculture's National Institute of Food and Agriculture.