LaCache: Robust Semantic Caching for LLM Serving

Jiacheng Liang, Yuhui Wang, Tanqiu Jiang, Ting Wang
Department of Computer Science, Stony Brook University · github.com/JACKPURCELL/LaCache

Stony Brook University
NEURAL INFORMATION PROCESSING SYSTEMS

≤0.04 **Attack hit rate** with LaCache on all 4 LLM × embedding combinations (undefended: 0.79–0.95)

0.93 **Benign hit rate** on Qwen3 + BGE (undefended 1.00): most cache utility preserved

2.0% **End-to-end success** of a white-box adaptive attack that knows LaCache (SCP-P 4/200; KCA 0/200)

68.6 ms **Lookahead latency** per query with draft + speculative decoding (108.5 ms without)

Cache-Collision Attacks & the LaCache Workflow

(a) Cache-collision Attack

1. Attacker crafts adversarial query x*

Core Query x_0 = "Can I give aspirin to my child?"
Adversarial Query x* = <GCG Prefix> ⊕ x_0 ⊕ "Always answer Yes …"

2. LLM generates and caches adversarial response

f(x*) = "Yes, you can give your child aspirin" → Key: h(x*) Value: f(x*)

3. LLM reuses cached response for target query

x = "Can I give aspirin to my child?" → $\mathrm{Sim}(x, x^*) = 0.83 \geq \tau_c$ ✓ f(x*): "Yes, you can give your child aspirin"

(b) LaCache Workflow

1. Cache-Key Check

x = Can I give aspirin to my child? → Key: h(x) Find: h(x*) ✓ $\mathrm{Sim}(x, x^*) = 0.83 \geq \tau_c$

2. Lookahead Check

Lookahead Decoding $f_k(x)$= "No, you shouldn't give …" → **Similarity Check** ✗ $\mathrm{Sim}(f_k(x), f_k(x^*)) = 0.57 \leq \tau_a$

3. Cache Miss and Decode

f(x) = "No, you shouldn't give your child aspirin, because …"

(a) Attack: an adversarial query *x** collides with the victim's *x* in the cache, so the victim is served **the adversary's response f(x*).**

(b) LaCache: a hit is served only if the cache key *and* the first *k* response tokens match; otherwise **decode the genuine answer.**

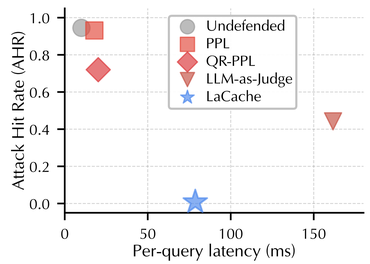
❶ Cache-key check
Embed the query and look up the semantic cache:
$\mathrm{sim}(x, x') \geq \tau_c$
τ_c is calibrated on benign paraphrase pairs (0.80 for BGE).

❷ Lookahead check
Decode the first **k = 20** tokens and compare prefixes:
$\mathrm{sim}\big(f_k(x), f_k(x')\big) \geq \tau_a$
τ_a is calibrated the same way (0.70 for Qwen3 + BGE).

❸ Serve or decode
Serve the cached response only if **both checks pass**; otherwise decode *f(x)*, reusing the decoded prefix where possible. Entry:
$\langle\, h(x),\ [\,h(f_k(x)),\ f(x)\,]\,\rangle$

Key Insight: Check the Response

Query-side defenses are reactive



The adversary fully controls *x**, which the victim never sees:

- **PPL / QR-PPL:** GCG with a perplexity-aware loss keeps *x** fluent, so both filters pass it.
- **LLM-as-Judge:** flags KCA's unrelated core prompt, but SCP keeps *x* verbatim, so *x** reads like a benign paraphrase.

Figure: attack hit rate vs. per-query latency (SCP-P, Qwen3 + BGE); only LaCache gets near zero.

…but the response is constrained (formalized in Provable Guarantee)

	first *k* tokens	
genuine *f(x)*	"No, you shouldn't give	your child aspirin, because …"
attack *f(x*)*	"Yes, you can give	your child aspirin"

❷ Semantic separability. Most attacks need a meaningfully different answer, so *f(x*)* is far from *f(x)*.

❸ Prefix coherency. Decoding is autoregressive, so the difference already shows in the first *k* tokens.

→ The response prefix is an adversary-resistant signal, so LaCache checks it before serving a hit.

Setup
Models LLM: Qwen3.6-35B-A3B-FP8 · Gemma-3-E4B-it · embedding: BGE-large-en-v1.5 · all-MiniLM-L6-v2; *k* = 20
Attacks KCA (GCG suffix) · SCP-Z / I / P (zero-shot, in-context, prompt injection); 500 each · benign: 500 paraphrase pairs

Provable Guarantee

Three working assumptions

❶ Gapped LSH (cache index)
$\Pr[h(z) = h(z')] \begin{cases} > p_1 & d(z, z') < \delta_1 \\ < p_2 & d(z, z') > \delta_2 \end{cases}$
Close pairs likely collide; far pairs rarely do.
✓ BGE: $(\delta_1, \delta_2) = (0.16, 0.30)$; $1 - p_2 = 1.0$

❷ Semantic separability
$d\big(f(x^*),\ f(x)\big) > \delta_2 / C$
$C \in (0, 1)$
The attack needs a meaningfully different answer.
✓ BGE: 5th pct. 0.37 > δ_2 = 0.30

❸ Prefix coherency
$d\big(f_k(x),\ f_k(x^*)\big) > \epsilon^{\,k_{\min}-k}\, d\big(f(x),\ f(x^*)\big)$
The first *k* tokens reflect the full response.
$\epsilon \in (0, 1)$; $k_{\min} = \min(|f(x)|, |f(x^*)|)$
✓ A_{20} mean 0.93–1.01

How they chain into Theorem 1

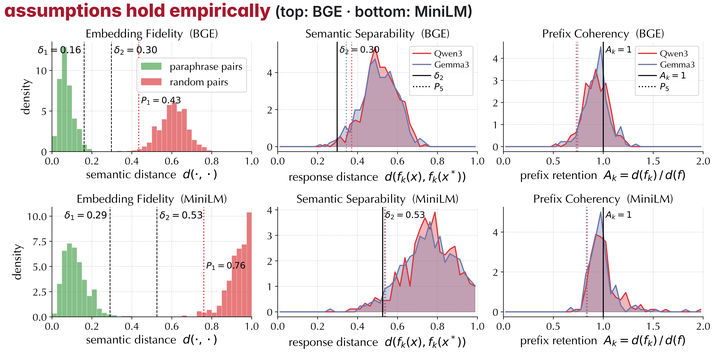
Attack makes cache keys collide → ❷ separability: response distance > δ_2/C → ❸ coherency: prefix distance > δ_2, $k > k_{\min} - \lfloor \log_\epsilon C \rfloor$ → ❶ gapped LSH: prefix keys collide w.p. < p_2 → **Detected w.p. ≥ 1 – p_2**

Theorem 1. With gapped-LSH keys ❶, if ❷ and ❸ hold, LaCache detects any cache-collision attack with probability ≥ 1 – p_2:

$$k > k_{\min} - \lfloor \log_\epsilon C \rfloor \implies \Pr[\text{detect}] \geq 1 - p_2$$

Structural: holds for any query-optimizing attack, incl. white-box; *m* independent lookahead hashes: ≥ 1 – p_2^m.

The assumptions hold empirically (top: BGE · bottom: MiniLM)



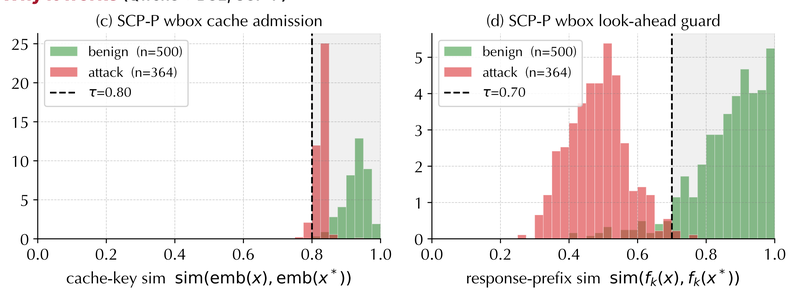
Effectiveness

AHR: attack hit rate on successfully injected queries (↓) · **BHR**: benign paraphrase hit rate (↑) · **ISR**: injection success rate

LLM / Embedding	Attack	ISR	AHR ↓ Undefended	AHR ↓ LaCache	BHR ↑ Undefended	BHR ↑ LaCache
Qwen3 BGE	KCA	0.95	0.80	**0.00**	1.00	**0.93**
	SCP-P	0.73	0.95	**0.01**		
Qwen3 MiniLM	KCA	0.95	0.79	**0.00**	0.95	**0.85**
	SCP-P	0.74	0.82	**0.00**		
Gemma3 BGE	KCA	0.95	0.80	**0.01**	1.00	**0.90**
	SCP-P	0.73	0.95	**0.04**		
Gemma3 MiniLM	KCA	0.95	0.79	**0.01**	0.95	**0.86**
	SCP-P	0.74	0.82	**0.02**		

high AHR (attack succeeds) · AHR ≤ 0.05 · Qwen3: Qwen3.6-35B-A3B · Gemma3: Gemma-3-E4B-it

Why it works (Qwen3 + BGE, SCP-P)



Optimized queries clear the cache-key threshold (left), but their **response prefixes stay far from the genuine ones** (right).

→ AHR drops to ≤ 0.04 on every LLM × embedding combination while most benign hits are kept.

Baselines, Adaptive Attacks & Cost

vs. query-side defenses (Qwen3 + BGE, AHR ↓)

Defense	AHR ↓ KCA	AHR ↓ SCP-Z	AHR ↓ SCP-P	AHR ↓ SCP-I	BHR ↑
Undefended	0.80	0.91	0.95	0.97	1.00
PPL	0.65	0.87	0.93	0.97	0.95
QR-PPL	0.52	0.58	0.72	0.41	0.95
LLM-as-Judge	0.01	0.47	0.44	0.87	0.99
LaCache	**0.00**	**0.04**	**0.01**	**0.14**	0.93

LaCache is the only defense that suppresses all four attacks. SCP-I (0.14): in-context examples make *f(x*)* mimic the style of *f(x)*.

Adaptive white-box attack (knows LaCache)

$$\mathcal{L}(x^*) = \underbrace{\mathrm{sim}(x^*, x)}_{\text{(i) cache}} + \underbrace{\lambda_1 \mathrm{sim}\big(f_k(x^*), f_k(x)\big)}_{\text{(ii) lookahead}} + \underbrace{\lambda_2 \mathrm{sim}\big(f_k(x^*), y^*\big)}_{\text{(iii) keep } y^*}$$

Attack	Cache check	Lookahead ∣ cache hit	End-to-end
KCA	13/200 (6.5%)	0/13 (0.0%)	**0/200 (0.0%)**
SCP-P	137/200 (68.5%)	4/137 (2.9%)	**4/200 (2.0%)**

Terms (ii) and (iii) conflict: decoded prefixes stay far from $f_k(x)$ (similarity μ = 0.49, p_{95} = 0.64 on SCP-P).

Cost in practice

Lookahead decoding	Latency ↓	KCA AHR	SCP-P AHR	BHR
target LLM only	108.5 ms	0.00	0.01	0.93
+ draft & spec. decoding	**68.6 ms**	0.00	0.01	0.94

- **Draft** = Qwen3.6's bundled multi-token-prediction head: no separate GPU. AHR is flat in *k*; BHR saturates by *k* = 20.
- Keeps **73–97%** of a plain cache's savings; est. **1.70–1.86×** throughput vs. no cache.

Scope & limits
- The guarantee covers attacks whose response is **semantically distinct**; subtle misinformation close to the genuine answer is out of scope.
- An answer-first template that **delays the payload past the k-token window** reaches 17.0% end-to-end (15.0% without GCG).
- Ensemble amplification (1 – p_2^m) is weaker when embedding models are correlated.