Cutting Recommendation Latency at the Edge

A Two-Tier Caching Study on Constrained Devices

Daniel Okafor
Department of Computer Science
Meridian University

Nimbus Systems Student Symposium ▪ 2026

Twelve minutes, one question

The same ranking model feels instant on one phone and sluggish on the next.

The setup

- Nexa's short-video app ranks the next clip on-device to avoid a round trip
- Ranking model: 4.2M parameters, quantized to int8
- Runs on everything from a $90 entry phone to a flagship

Field data, 3 weeks

68% of entry-tier requests missed the 120 ms budget

211 ms median inference on entry-tier hardware

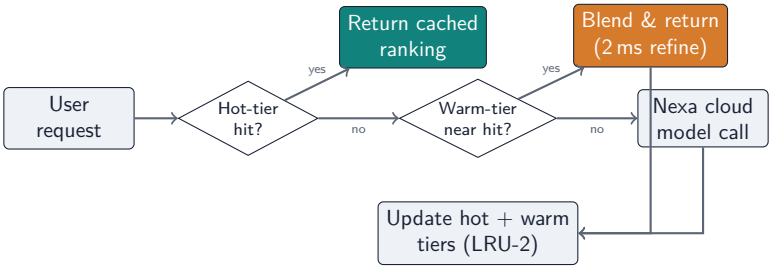
9 pp drop in engagement when latency >150 ms

The question we chase today

Can a small, adaptive cache in front of the model close most of that gap — without touching accuracy?

A two-tier cache in front of the model

Hot tier for repeats, warm tier for near-neighbours, cloud for the rest



Hot tier

Exact-match cache on (user, session context), keyed by a 32-bit fingerprint. 4k entries, sub-millisecond lookup.

Warm tier

LSH over the last embedding; a 2 ms refinement pass blends the neighbour's ranking with fresh signals.

What we varied, what we held fixed

Three factors, one ablation grid, same eval set across all runs

Factor	Levels tested	Held fixed
Hot-tier size	1k / 4k / 16k entries	Eviction: LRU-2
Warm-tier radius	0.05 / 0.10 / 0.20 (cosine)	Refine budget: 2 ms
Device tier	entry / mid / flagship	Model weights (int8, frozen)

Why this matters for a 12-minute talk

The full grid is 27 runs $\times$ 3 device tiers. Today I show the configuration that won on the held-out validation split — 4k hot / 0.10 radius — and how it generalizes across hardware.

Result I — latency drops hardest where it hurt most

p50 / p95 inference latency, cache vs. no cache, 4k hot / 0.10 warm radius

Device tier	p50 base	p50 cached	p95 base	p95 cached
Entry ($90)	211 ms	68 ms	340 ms	129 ms
Mid-range	104 ms	41 ms	176 ms	74 ms
Flagship	52 ms	24 ms	88 ms	41 ms

Reading the table

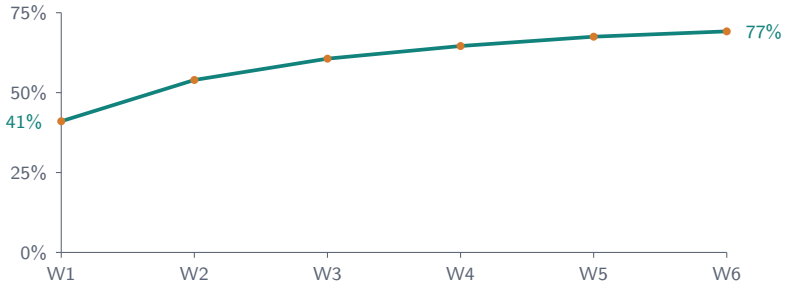
- Entry-tier p50 falls **68%**, closing most of the gap to flagship
- p95 improves even more — the cache absorbs worst-case stalls, not just the average
- Accuracy moved by <0.3 NDCG points

Budget check

Entry-tier requests under the 120 ms budget:
32% → **91%** after caching

Result II — the cache keeps learning

Hot-tier hit rate over the six-week field trial, entry-tier cohort



Why it climbs

LRU-2 keys off session fingerprints that recur across days; by week 3 the hot tier has seen most recurring contexts twice, so hits stop depending on session length.

Takeaway

One sentence, three numbers, one open question

The sentence

A 4k-entry hot cache with a locality-sensitive warm tier cuts entry-tier p95 latency by **62%** for **under 0.3 NDCG points** of ranking quality.

68%	91%	77%
entry-tier p50 drop	requests inside budget	steady-state hit rate

Open question for the room

The warm tier's LSH radius was tuned per device tier by hand. Does an online bandit beat our fixed schedule on hardware we haven't tested?

Thank you

Code, data, and the full ablation grid are on the project page

Get in touch

d.okafor@meridian.edu

github.com/dokafor/edge-cache

Vertex Systems Lab, Meridian University

Acknowledgements

Advised by Prof. Amara Solano. Field trial run in partnership with Nexa's on-device ML team. Compute provided by the Meridian Research Cluster.

Selected references

Mateo Delgado. Quantization tradeoffs for mobile inference pipelines. In *Vertex Systems Conference on Mobile Computing*, pages 203–215, 2023.

Colin Marsh and Renu Iyer. Locality-sensitive hashing for streaming embedding lookups. In *Proceedings of the Apex Workshop on Efficient ML*, pages 77–88, 2024.

Daniel Okafor and Amara Solano. Two-tier caching for quantized recommendation models. *Meridian Journal of Systems Research*, 8:41–59, 2026.

Amara Solano and Priya Whitfield. On-device ranking under tight latency budgets. In *Proceedings of the Nimbus Systems Symposium*, pages 112–124, 2025.