LAUNCH BRIEF

Ship a cache layer that makes LLM products feel instant without rewriting the stack

Pilot teams cut median response latency 42% while reducing retrieval spend 31% in the first four weeks.

Folio contributors
2026.04.29
v1.0 · Public

42% Median latency cut

42% Median latency cut

42% Median latency cut

42% Median latency cut

Orbit Cache is a slim infrastructure layer for retrieval-heavy LLM products. It keeps your current model stack intact, but moves hot context, ranking artifacts, and repeated search paths into a fast cache so the product feels measurably faster on day one.

Why this launch matters

Teams have already paid the model bill. What they still have not fixed is the user-perceived lag between question and answer.

- Knowledge search workload: retrieval spend dropped 31% once repeated passages stopped re-querying the vector store.
- Customer support copilot: 42% faster median response time after hot-ticket caching.
- Internal research assistant: prompt iteration volume tripled because the latency ceiling stopped shaping the workflow.

Proof from the pilot

Orbit Cache solves that narrow problem first: faster repeated retrieval, fewer wasted vector calls, and more room to experiment on prompts.

- Pilot sales engineers reported that demos felt instant enough to stay conversational instead of turning into narration while the system waited.
- Pilot sales engineers reported that demos felt instant enough to stay conversational instead of turning into narration while the system waited.
- Pilot sales engineers reported that demos felt instant enough to stay conversational instead of turning into narration while the system waited.

Roadmap

three-step arc for proposals

PHASE 1

Rollout path

Instrument existing retrieval calls and identify the hottest repeated paths.

PHASE 2

Operating discipline

Enable response and embedding cache for two user journeys with visible lag.

PHASE 3

Operating discipline

Expand to reranking artifacts and tune TTL by workload economics.

> The win is not abstract infrastructure elegance. The win is that the product finally feels fast enough for people to use it as often as they intended.