LEXI AOE · PILOTTILBUD

Lexi AOE Cost Pilot — samme svar skal ikke koste to ganger

4 uker · fast omfang · fra NOK 95 000 eks. mva

LLM-regningen din skalerer med trafikk. Det burde den ikke: mesteparten av volumet er gjentatte *former* av samme arbeid. Kjernen vår lærer formene, verifiserer hvert svar mot en uavhengig beregning før det serveres, og slutter å ringe modellen for arbeid den allerede kan — med kvittering for hvert svar.

Målt i vår referansetest (mot en live modell): ~90 % vedvarende kostnadsreduksjon med 350 av 350 riktige svar — der rå LLM svarte feil på 2 — og 7,5× lavere svartid på gjentatt trafikk. Testen er syntetisk og egenprodusert; derfor er uke 1 i piloten å kjøre den samme benchmarken på et utvalg av *deres* trafikk, med forseglede resultater dere kan etterprøve.

DETTE LEVERER PILOTEN — FOR ÉN ARBEIDSFLYT MED HØYT SPØRREVOLUM

1. **Benchmark på deres trafikk (uke 1).** Kjernen mot rå LLM på et representativt utvalg. Kostnad, korrekthet og svartid, målt ved modell-grensesnittet — resultatene forsegles i en hasjkjede dere kan verifisere uten oss.
2. **Integrasjon (uke 2).** Kjernen koblet inn foran modellen deres (stdio/HTTP/MCP/OpenAPI). Ingen avhengigheter i kjørefasen.
3. **Sikkerhetsmekanismene (uke 3).** Verifiser-før-servering, avvisning av skitne data i stedet for gjetning, menneskegodkjent læring: kjernen svarer aldri på egen hånd med noe den ikke kan belegge.
4. **Drift og overlevering (uke 4).** Målt kost/korrekthet-rapport, Ed25519-signert kvitteringskjede for hvert svar, og deres team i stand til å drifte den.

Dette er ikke caching. Semantisk caching gjenbruker svar den håper er like. Kjernen serverer bare svar den har verifisert mot en uavhengig beregning, nekter på skitne data, lærer kun med signert godkjenning — og leverer kvittering for hvert svar. En ansvarlig utførelsesgrense, ikke en cache.

Betingelser. Fast pris, fakturert 50/50. Dere beholder alle resultater og kvitteringer; verifisering virker uten oss.

Grensen vi ikke krysser. Tallene over er fra vår egen referansetest på syntetiske data — vi selger ikke prosenter, vi selger målingen på deres trafikk, med bevis. Besparelsen hos dere avhenger av hvor repetitiv trafikken deres faktisk er.

Kontakt: post@lexico.no · Emne: «AOE Cost Pilot»

Levert av LexiCo AS (Norge)

LEXI AOE · PILOT OFFER

Lexi AOE Cost Pilot — the same answer shouldn't cost twice

4 weeks · fixed scope · from NOK 95,000 ex VAT

Your LLM bill scales with traffic. It shouldn't: most volume is repeated *shapes* of the same work. Our kernel learns the shapes, verifies every answer against an independent computation before serving it, and stops calling the model for work it already knows — with a receipt for every answer.

> **Measured in our benchmark** (against a live model): ~90% sustained cost reduction with 350/350 correct answers — where the raw LLM got 2 wrong — and 7.5× lower latency on repeat traffic. That benchmark is synthetic and owner-authored; which is why week 1 of the pilot runs the same benchmark on a sample of *your* traffic, with sealed results you can re-verify.

WHAT THE PILOT DELIVERS — FOR ONE HIGH-VOLUME WORKFLOW

1. **Benchmark on your traffic (week 1).** Kernel vs raw LLM on a representative sample. Cost, correctness and latency measured at the model seam — results sealed in a hash chain you can verify without us.
2. **Integration (week 2).** The kernel wired in front of your model (stdio/HTTP/MCP/OpenAPI transports). Zero runtime dependencies.
3. **The safety mechanisms (week 3).** Verify-before-serve, refusal on dirty data instead of guessing, human-gated learning: the kernel never answers on its own with anything it cannot substantiate.
4. **Operations and handover (week 4).** Measured cost/correctness report, an Ed25519-signed receipt chain for every answer, and your team able to run it.

> **This is not caching.** Semantic caches reuse answers they hope are similar. The kernel serves only answers it has verified against an independent computation, refuses on dirty data, learns only under signed ratification — and hands you a receipt for every answer. An accountable execution boundary, not a cache.

> **Terms.** Fixed price, invoiced 50/50. You keep all results and receipts; verification works without us.

> **The line we do not cross.** The numbers above are from our own benchmark on synthetic data — we don't sell percentages, we sell the measurement on your traffic, with proof. Your savings depend on how repetitive your traffic actually is.

Contact: post@lexico.no · Subject: "AOE Cost Pilot"

Delivered by LexiCo AS (Norway)