Memory That Pays for Itself

Cutting LLM token burn on a production enterprise BI agent — without trading away answer quality

Enterprise AI agents are expensive because they re-derive context on every call — searching, inspecting schemas, previewing data, and reasoning from scratch on questions the system has effectively solved before. In a controlled A/B benchmark on a production enterprise BI agent, **GoodMem's** memory and retrieval layer cut aggregate token consumption **28%** (95% CI 2–47%) and agent reasoning iterations **23%** (95% CI 5–37%), with **no measurable change in answer quality** (paired Δ ≈ 0; 95% CI −23.5 to +23.5). On the heaviest, retrieval-intensive queries — the ones that dominate the bill — token reductions reached 30–66%.

−28%
Aggregate token burn
95% CI 2–47% · significant

−23%
Agent reasoning steps
95% CI 5–37% · significant

≈ 0
Change in answer quality
paired Δ, interval spans 0

HEADLINE RESULTS

Metric	Memory OFF	Memory ON	Change	95% CI
Aggregate token burn (17 scenarios)	6.92M	4.99M	**−28%**	2–47%
Median tokens per query	319K	150K	**−53%**	—
Agent iterations (aggregate)	216	167	**−23%**	5–37%
Agent iterations (median / query)	11	7	**−36%**	—
Answer quality (judge pass rate)	46%	46%	≈ 0 pts	−23.5, +23.5

17 matched scenarios · 3 replicates / condition (57 scored runs per arm) · identical frontier agent model and independent LLM judge in both arms · the only variable changed is GoodMem memory · intervals are nonparametric bootstrap estimates (10,000 resamples, clustered by scenario).

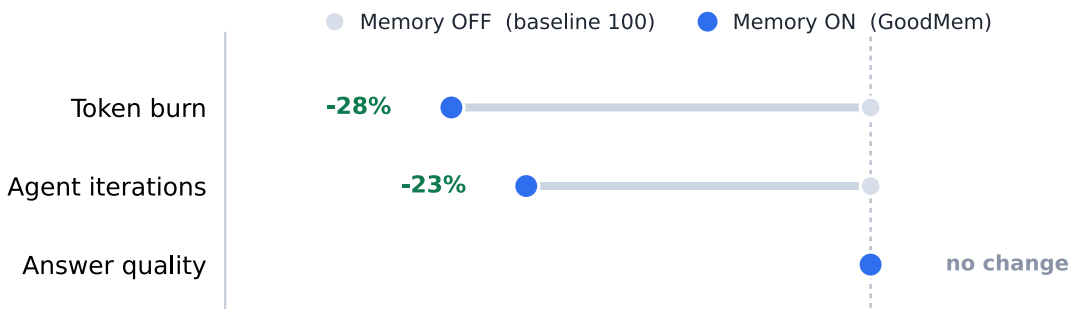

Figure 1 · Same answers, fewer resources — indexed to Memory OFF = 100.

The problem: agents pay full price for every question

Agents that answer analytical questions over enterprise data are token-hungry by construction: every query reruns semantic search, schema inspection, data preview, SQL generation, and self-correction, filling the context window to rediscover the same tables, joins, and pitfalls it navigated yesterday. At scale, production BI agents now generate six-figure monthly token bills that grow with adoption rather than amortizing across it.

The approach: serve memory, don't re-derive it

GoodMem is a self-hostable memory and retrieval control plane for enterprise AI. It captures the durable lessons of prior conversations — which business views to use, how a metric is derived, which joins and filters are correct, what pitfalls to avoid — and prefetches only the relevant guidance into each turn. The agent starts informed instead of from zero and converges in fewer steps, with no change to its model, prompts, or logic.

Where the savings live

Savings are not uniform — and that is the point. They concentrate on the expensive, retrieval-heavy queries that drive the bill: multi-table aging, cross-module joins, and year-over-year roll-ups saw 30–66% fewer tokens, while trivial lookups changed little. **Memory savings track spend** — 10 of 17 scenarios consumed fewer tokens, with the largest gains exactly where enterprise agent cost is largest.

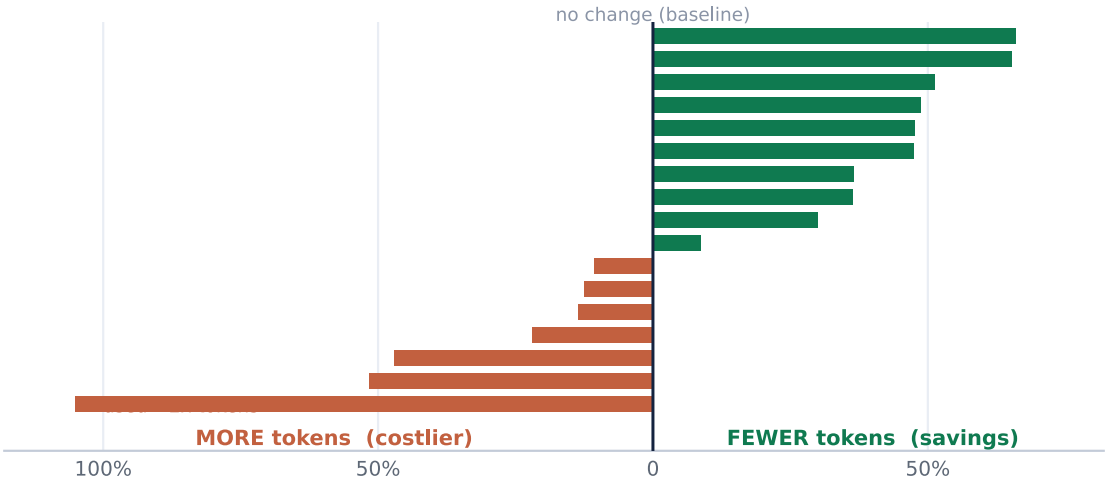

Figure 2 · Per-scenario change in token use. Right of the baseline = fewer tokens (savings); left = more. A reduction caps at 100%; an increase does not — one scenario used roughly twice the tokens.

What it means for cost

Illustratively, a deployment with a $100K / month token bill — concentrated, as bills are, on heavy queries — recovers roughly $28K / month, about **$340K per year**, from a benchmark-grade 28% reduction with no change to the agent. Because savings scale with the retrieval-heavy traffic that grows fastest, the benefit compounds as adoption rises.

Quality, held constant

Cost reduction only matters if quality holds. It did: the paired difference in judge pass rate was indistinguishable from zero, with a confidence interval centered on it. Memory changed how efficiently the agent worked, not whether it was right — fewer steps, fewer tokens, the same answers.

Why PAIR Systems

PAIR Systems builds the autonomous information-retrieval layer for enterprise AI. GoodMem brings capabilities once confined to the world's best search teams into any product as self-hostable software. PAIR's founder helped pioneer practical zero-shot neural retrieval at Google, extended it across languages and modalities, and productized enterprise retrieval at Vectara. We partner with data and analytics platforms — and enterprises running internal agents at scale — to deliver the memory layer they would otherwise need a dedicated research team to build. For the heaviest token spenders, GoodMem extends into model distillation and self-tuning, compounding the savings shown here.

Methodology. Controlled, paired A/B on a production enterprise BI / analytics agent in a live deployment. 17 matched scenarios across accounts-payable, sales, inventory and multi-step drill-down workloads; 3 scored replicates per scenario per condition. Identical frontier agent model and an independent LLM judge across both arms; the sole intervention is GoodMem's memory layer (prefetch of distilled prior-conversation guidance). Confidence intervals are nonparametric bootstrap estimates (10,000 resamples, clustered by scenario). Infrastructure failures (provider rate-limits, transport faults) are excluded as non-quality events. Pilot-scale benchmark; effect sizes are large and directional within the stated intervals; broader multi-tenant validation in progress.