TokenOptimizer Proof-of-Concept Report

Prospect: {{Customer Name}} • **Date:** {{YYYY-MM-DD}}

Reproduction Checklist

1. Start Postgres + backend: `./infra/startup.sh`
2. Ensure `.env` (backend) contains `API_KEY=tok_demo_key` for the demo workspace.
3. Run fixtures + measurements: `./backend/demo/scripts/run_demo.sh`
4. Review JSON outputs inside `backend/demo/results/` and update the table below.

Problem Statement

Large RAG/chat/multi-agent payloads force your Gemini/GPT stack to burn tokens, slow SLAs, and complicate SOC2 evidence. TokenOptimizer compresses the noisy context into TOON format so you pay only for what matters — without retraining models.

What We Did

- Loaded three realistic payloads (HR RAG, sales chat history, multi-agent ops thread)
- Ran `/v1/optimize` and `/v1/optimize-run` with Safe/Balanced modes enabled
- Captured metrics + latency from API responses and verified savings with the demo script

Inputs

- `backend/demo/fixtures/rag_sample_large.json`
- `backend/demo/fixtures/chat_history_long.json`
- `backend/demo/fixtures/multi_agent_round.json`

Results

Fixture	Original Tokens	Optimized Tokens	Savings
HR / Product / Meeting RAG	1,611	107	**93.36%**
Sales Chat History	761	168	**77.92%**
Multi-Agent Ops Thread	1,063	28	**97.37%**

Token counting method: Metrics come from the backend estimator (word count × 1.3) pulled directly from `/v1/optimize(-run)` responses. Pilot contracts replace this with Gemini/OpenAI tokenizer-based billing for exact invoice parity.

Quality assurance: Two SMEs reviewed 10 randomly sampled optimize-run outputs (RAG + chat). No regressions were flagged; summaries stayed on-message and safety filters were unchanged.

Confidence / variance: n = 3 fixture families, 6 total runs. Mean savings = 89.55% with ±8.5% swing between fixtures.

Latency impact: < 120 ms per request for compression + TOON conversion; model invocation latency (LLM round-trip) is unchanged and measured separately. Zero-retention mode purged raw chat buffers within 60 seconds.

Before → After snippet (HR policy excerpt):

```
RAW : "Managers must run quarterly pulse surveys to monitor burnout signals..."
RAW : "Security requires managed devices before prod VPN access..."
RAW : "Cross-border checklist warns about payroll tax triggers..."
TOON: managers
        cadence: quarterly pulse survey, escalate if eNPS < +20
      security
        access: managed laptop + autopatch, VPN ok
      mobility
        tax_threshold_days: 30, flag payroll + export controls
```

Recommended modes: Balanced for production agents, Aggressive only for internal diagnostics.

curl Reproduction

```
# Optimize (RAG)
curl -s -X POST "$API_URL/v1/optimize" \
  -H "Authorization: Bearer $API_KEY" -H "Content-Type: application/json" \
  -d "{\"input\": $(python3 - <<'PY'
import
json;print(json.dumps(open('backend/demo/fixtures/rag_sample_large.json').read()))
PY
)}"

# Optimize + Run
curl -s -X POST "$API_URL/v1/optimize-run" \
  -H "Authorization: Bearer $API_KEY" -H "Content-Type: application/json" \
  -d '{"input": "'"$(cat frontend/playground/examples/sample_chat_input.txt | sed
's/"/\\"/g')"'", "use_history": true, "mode": "balanced"}'

# Convert JSON → TOON
data='{"customer":"Acme AI","risk":"needs SOC2 evidence"}'
curl -s -X POST "$API_URL/v1/convert-to-toon" \
  -H "Authorization: Bearer $API_KEY" -H "Content-Type: application/json" \
  -d '{"input":'"$data"'}'
```

Key Findings

- **Quality preserved:** Gemini outputs matched human-written summaries; procurement-ready TOON diffs are audit-friendly.
- **Latency:** 85–120 ms overhead (mainly TOON conversion) vs. baseline, acceptable for async RAG + batch jobs.
- **Security:** Zero-retention verified; logs redact PII automatically.
- **Billing-ready:** Stripe sandbox shows Builder→Scale upgrade path, enabling self-serve expansion.

Next Steps & Pilot Offer

- 2-week paid pilot ($2,500 fixed or 20% of first-month savings) covering three production workloads.

- Deliverables: before/after metrics, SOC2-ready TOON diffs, Stripe billing wiring, final executive readout.
- Timeline: kickoff Monday, live demo Friday of week 2.

Contact & Privacy

hello@tokenoptimizer.ai • tokenoptimizer.ai • Zero-retention + private VPC options available — no customer data stored beyond aggregate metrics.

— Manish, Founder, TokenOptimizer