CIYA

The Sovereign Deterministic AI
Logic Layer for LLMs

JEREMY ZIERAU
619-985-1923
JEREMY@JZIERAU.COM

TOKEN WASTE IS ABSURD

60% redundant input: From Abstrabit Technologies' "Total Token Spend" forensic audit (2025/2026 LinkedIn breakdown)—valid enterprise insight on waste vs. value tokens.

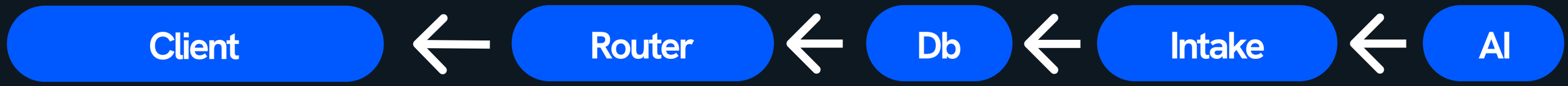
In agentic/RAG workflows, 30%+ of queries are repetitive/ semantically similar, inflating costs (enterprise RAG studies).

40-90% of an AI budget can be wasted on redundant prompts and forgotten responses.

CIYA

Benefits

- Slash token spend through Zero-Trust Determinism
- 91.53% Residual Collapse Optimization
- Hallucinations? What are those?
- 92x faster than probabilistic wrappers (sub-second response times)



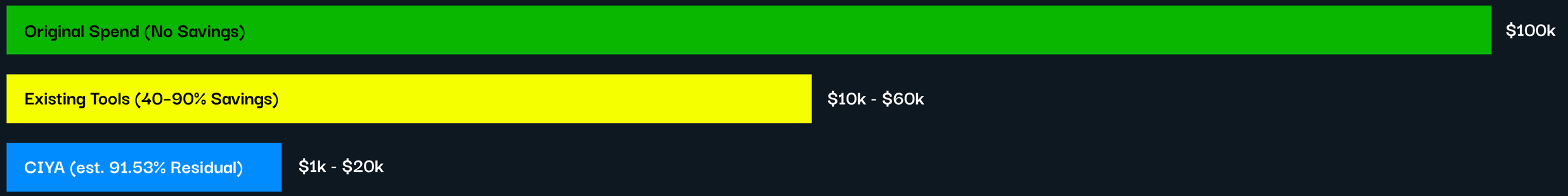
THE ENGINE OF EFFICIENCY

- **Auto-capture**: Logs all User / AI interactions as a persistent “Source of Truth”
- **Active Optimization:** Custom tools at the intake and router steps intercept queries
- **LLM Bypass:** As your data grows, CIYA resolves more queries locally, slashing LLM dependency.
- **Residual Collapse:** In the event LLMs slash token prices or implement similar functionality, CIYA’s value increases.

AI EFFICIENCY: THE NEXT GOLD RUSH

- **Global AI Spending**: ~$2.5T in 2026 (Gartner), with inference/token costs being the major issue.
- **Target:** Heavy AI users (devs, agencies, enterprises). Millions spending $1K-$100K+/mo on LLMs.
- **Opportunity:** Tools like Mem0, Redis, + RAG can save 40-90%. tokens; Optimization Layers target 80-90%. CIYA optimizes an additional 91.53%.
- **Why Now?:** Agentic AI + long contexts exploding waste, potential environmental impact from reducing token waste

How Much You Actually Pay After Savings:



THE LANDSCAPE

Competitors are circling the solution, but are missing the mark that makes up the bulk of CIYA's capability. In light of this, CIYA compounds token savings in addition to what the current landscape offers.

	CIYA	Native Caching Anthropic / OpenAI	Semantic Cache GPTCache / Redis / Bifrost	Orchestration LangChain / LlamaIndex
True Determinism	✓ Yes	✗ No	~ Approx	✗ No
Air-gapped / no-cloud dependencies	✓ Yes	✗ No	✗ No	✗ No
Runs on legacy hardware	✓ Yes	✗ No	✗ No	✗ No
AI model agnostic	✓ Yes	✗ Vendor locked	~ Partial	✓ Yes
Zero ongoing token spend	✓ 91.53% reduction	~50-90% on repeats	~ 30-70% on similar	✗ No reduction
Survives model depreciation	✓ Yes	✗ No	✗ No	✗ No

REVENUE MODEL: FREEMIUM TO ENTERPRISE

SUBSCRIPTION-BASED + USAGE TIERS

Free

- Basic storage/retrieval
- Token spend prevention
- Limited API use

Pro

- Full API, SMS & Socket use
- Expanded storage / retrieval
- Advanced segmentation
- Team collaboration

Enterprise

- Dedicated storage
- Private response libraries
- External database support
- Compliance
- And More!

CIYA

BUILT BY A FOUNDER OBSESSED WITH EFFICIENCY.

Jeremy is a 20-year self-taught engineer who has built enterprise systems that communicate daily with thousands of robots and video Feeds around the world.

In 2017, he built the world's first 'robotic fleet management system' for BrainCorp which now manages all automated floor care robots in Walmarts around the world.

In 2020, Jeremy built the DSP MVP that won FOX Studios as a client for MadHive.

In 2024, Jeremy built IIIO, a proprietary agentic-AI protocol.

In 2026, he founded CIYA, the deterministic storage & logic engine that reduces LLM token spend by 91.53% with zero cloud dependency.