FRUX

SAVINGS AS A SERVICE

USPTO Patent Pending · App. #63/947,144

Cut Your AI Costs by 70-92%.

92.3%	89-93%	0%	<2 min
Max Cost Reduction	Across All LLM Models	Quality Degradation	Integration Time

How It Works

01 Deploy as Proxy

FRUX sits between your app and the AI provider. No code changes.

02 Auto-Optimize

TVP caches patterns, deduplicates files, routes to optimal models.

03 Pay on Savings

No savings = no cost. You always keep the majority of savings.

Validated Performance

Real production tests · USPTO filing data · Dec 2025

Use Case	Without FRUX	With FRUX	Savings	Model

Annual Savings Projection

Based on 8 AI sessions/day · 20 days/month · Claude Sonnet

Team	Direct Cost	With FRUX	You Save
1 Dev	$26,285/yr	$2,035/yr	**$24,250**
10 Devs	$262,848/yr	$20,352/yr	**$242,496**
100 Devs	$2.63M/yr	$203K/yr	**$2.42M**

Why FRUX

- **Zero Risk**
  No savings = no cost. Ever.
- **Zero Code Changes**
  Transparent proxy integration.
- **Model Agnostic**
  Claude, GPT-4, Gemini, any LLM.
- **Patent Protected**
  USPTO App. #63/947,144.
- **Actually Faster**
  11% speed gain from smaller payloads.

Technology

SLIM Protocol
40-50% base token reduction. Open source.

Smart Routing
Auto model selection by complexity.

TVP Engine
Semantic caching. Up to 96.3% on repeated content.

Cross-User Cache
Process once, serve all users.

Start your free 30-day pilot

No commitment · No upfront cost · Results guaranteed

info@frux.ai

frux.pro · +39 340 706 3047