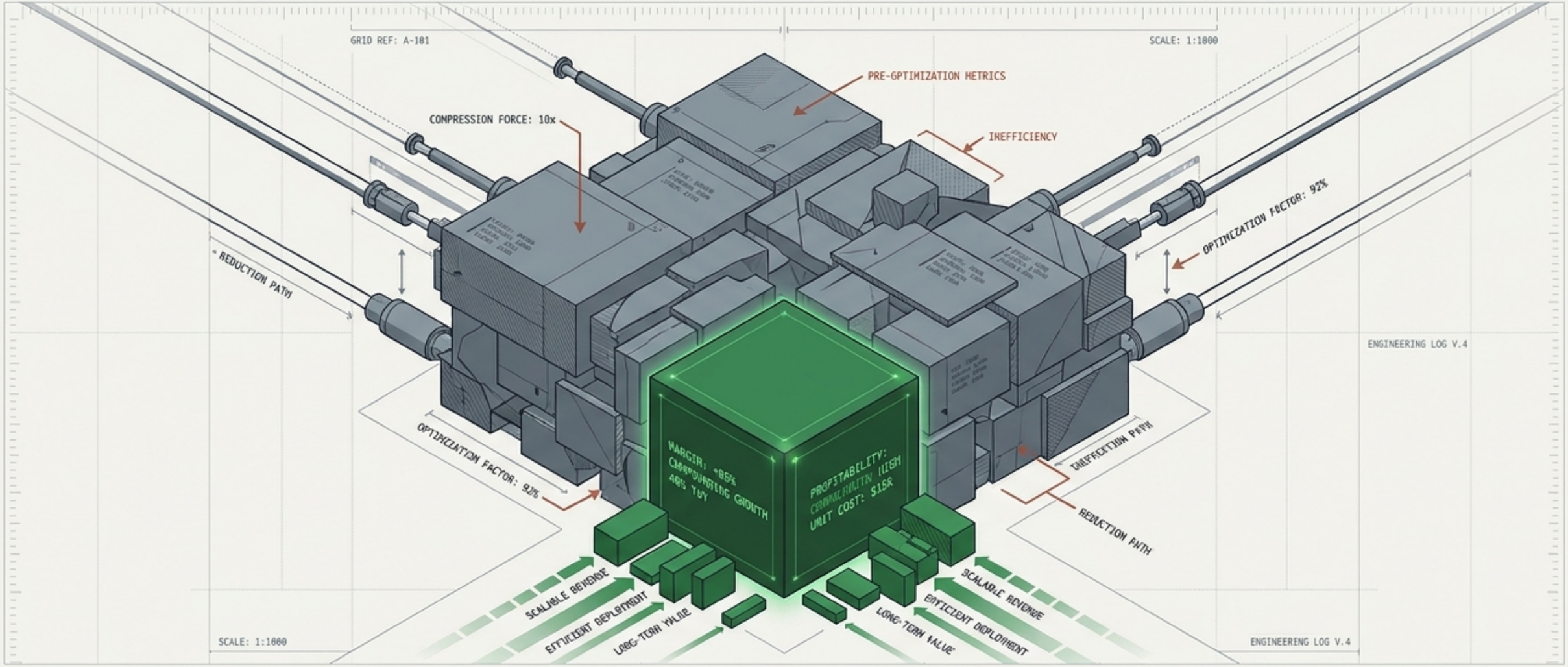
MIO UNIT ECONOMICS

From absurd prototype costs to highly compounding margins.

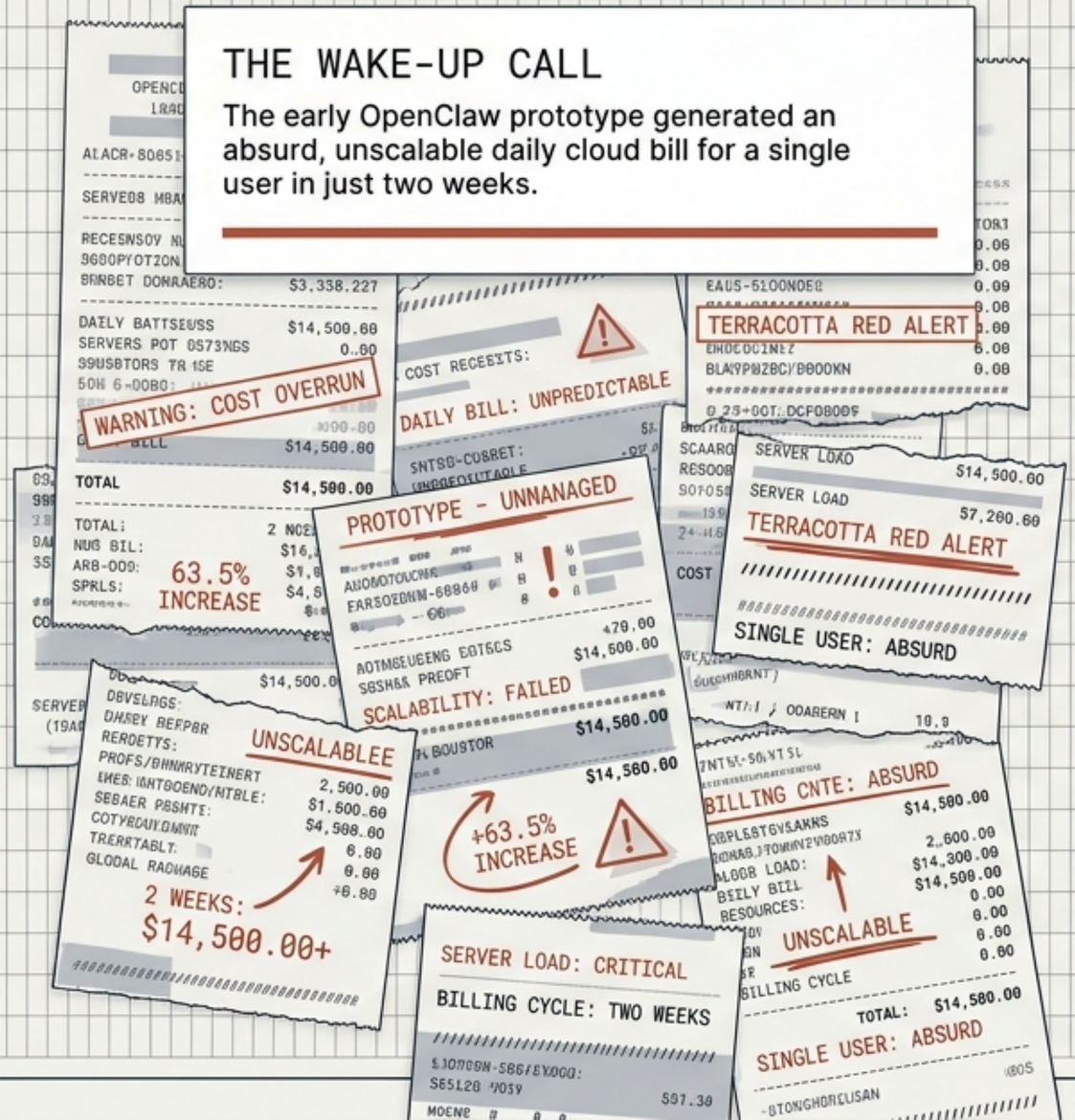


COST IS A FIRST-CLASS ENGINEERING PROBLEM.

THE PROTOTYPE

THE WAKE-UP CALL

The early OpenClaw prototype generated an absurd, unscalable daily cloud bill for a single user in just two weeks.

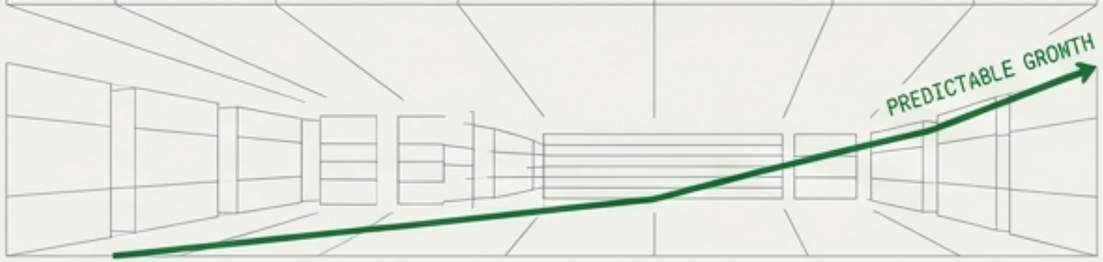


VERSION 8

THE ENGINEERED SOLUTION

Eight versions later, cost is no longer an afterthought. It is a strictly controlled variable generating predictable, highly profitable unit economics.

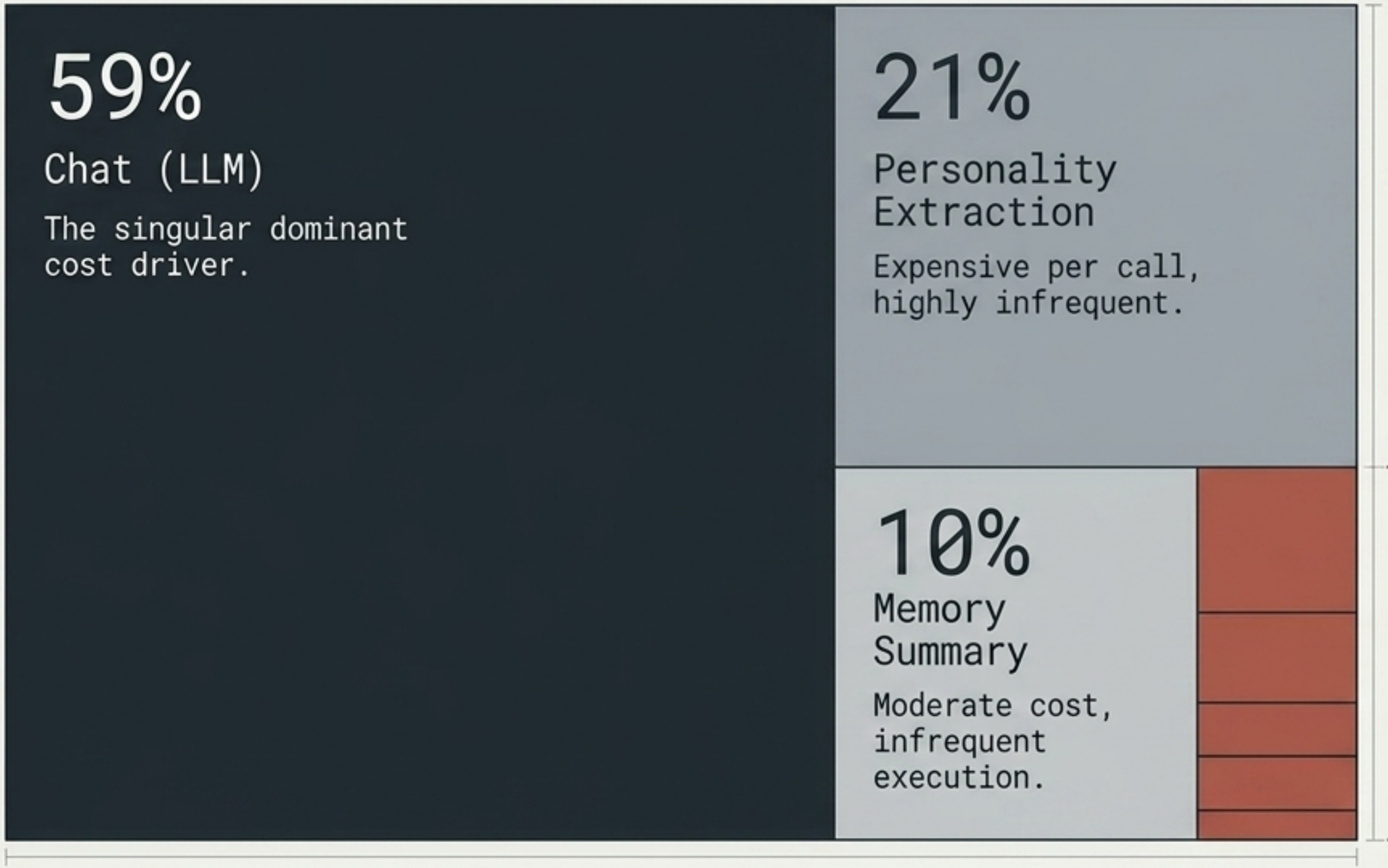
VERSION	VARIABLE	CONTROL	PREDICTABILITY	PROFITABILITY	UNIT COST	MARGIN
V8.0	COST	STRICT	99.9%	HIGH	$0.0032	+72.4%
V8.0	SCALE	MANAGED	100%	OPTIMAL	$0.0032	+72.4%
V8.0	BILLING	FIXED	PREDICTABLE	STABLE	$0.0032	+72.4%
V8.0	RESOURCES	OPTIMIZED	BALANCED	EFFICIENT	$0.0032	+72.4%
V8.0	UNIT ECON	CONTROLLED	FORECASTABLE	PROFITABLE	$0.0032	+72.4%
V8.0	GROWTH	SUSTAINABLE	LINEAR	SCALABLE	$0.0032	+72.4%



PROFITABLE UNIT ECONOMICS: ACHIEVED

ENGINEERING LOG: V8.0 - FINAL OPTIMIZATION COMPLETE. COST CONTROLLED.

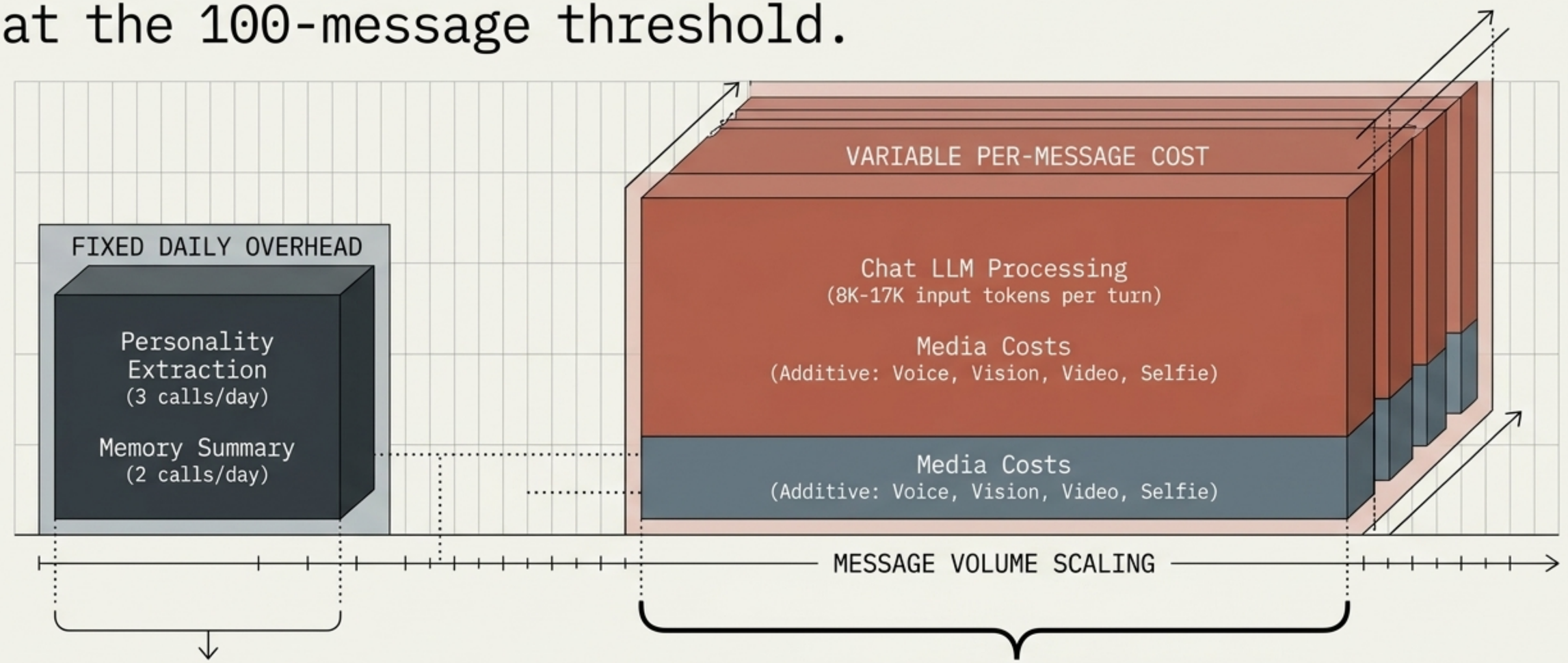
Chat compute dominates the bill, while advanced features are a rounding error.



Rounding Errors

TTS Voice (3%)
Memory Extraction (3%)
Proactive Messaging (2%)
Memory Rerank (2%)
Memory Rerank (2%)
Embeddings (<1%)

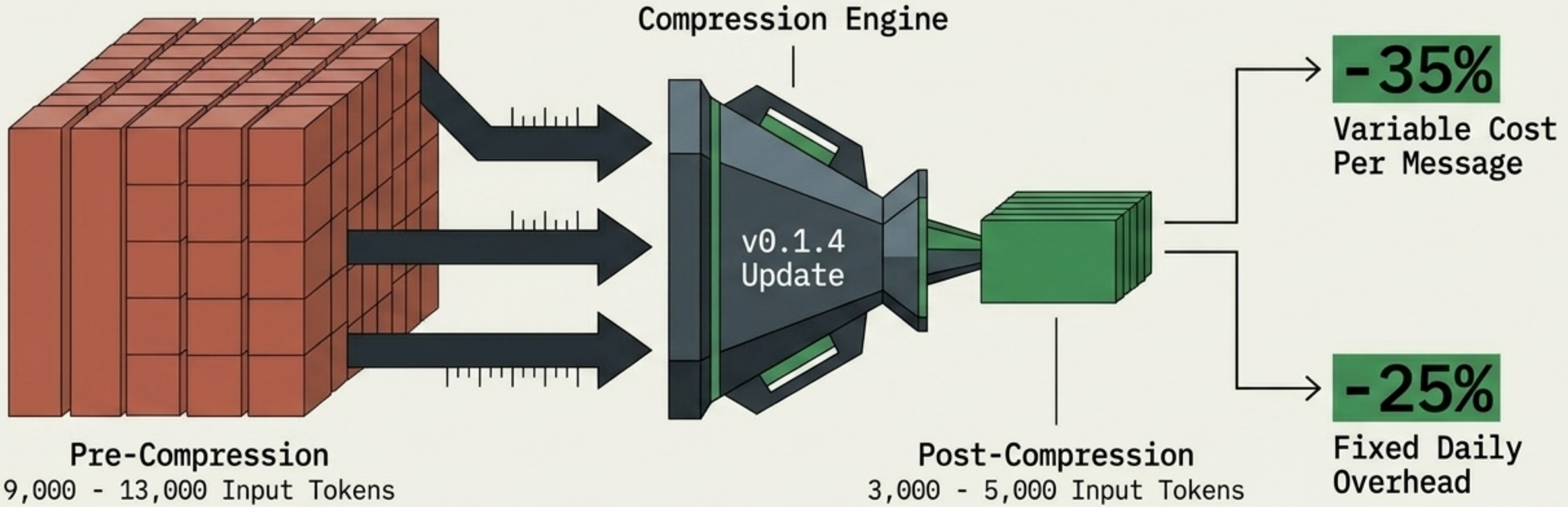
Variable token costs eclipse fixed infrastructure at the 100-message threshold.



<30 Messages/Day:
Fixed overhead matters.

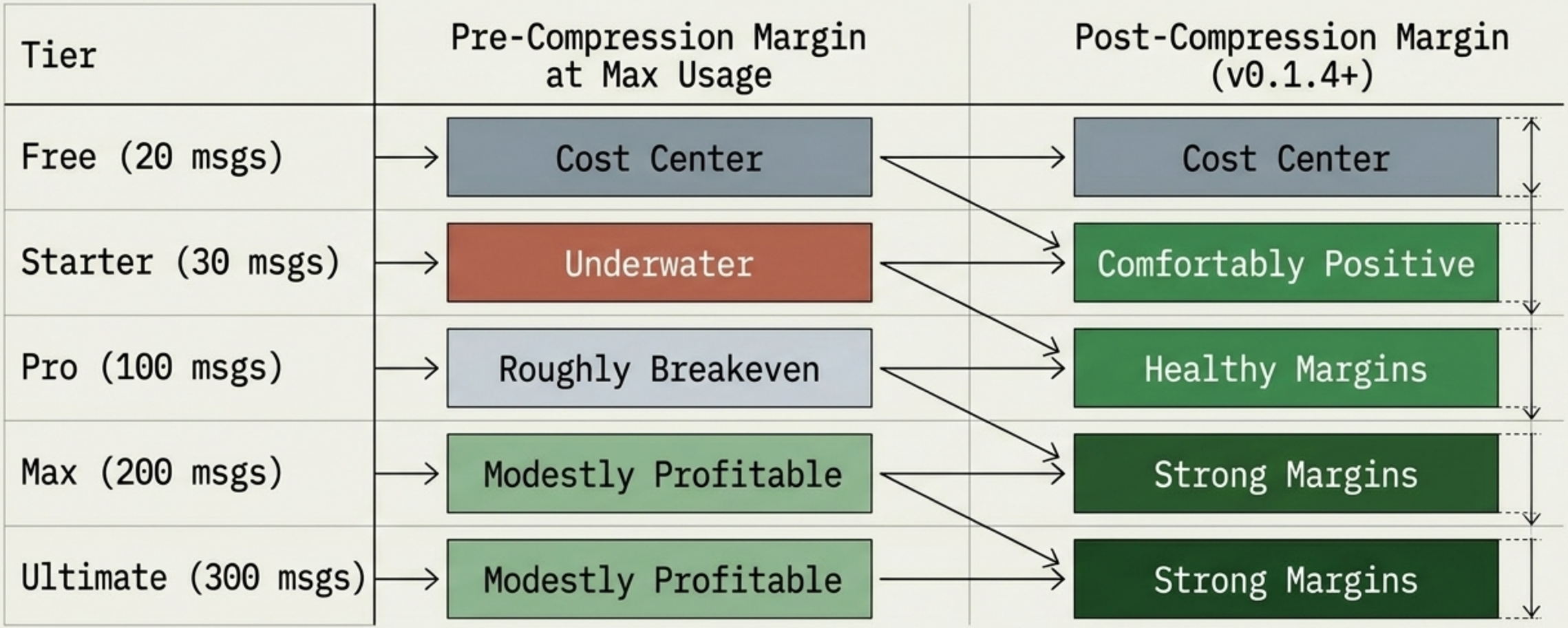
100-300 Messages/Day: Variable LLM cost completely dominates the bill, making token efficiency the single most critical lever.

A 60% system prompt reduction fundamentally alters unit economics across the board.



Because the system prompt is the largest chunk of input tokens per chat call, compressing it instantly creates a compounding margin increase for every active user.

Post-compression, every single paid tier operates comfortably in the black.



* These margins assume worst-case max usage (a user hitting their cap 30 days a month). Real-world usage patterns average 40-60% of the cap, meaning actual realized margins are subs.tantially higher.

Premium tiers command higher prices for zero-marginal-cost features.

FEATURE COST HIERARCHY

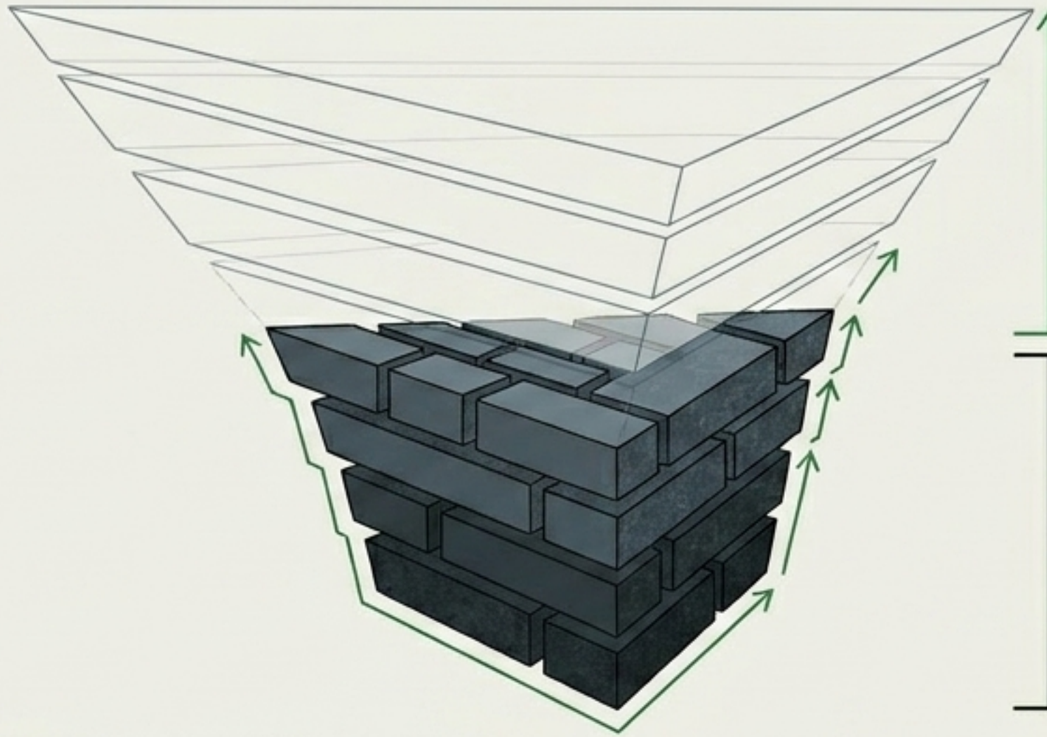
Max & Ultimate Tiers
Near-zero marginal costs

- Priority queue ordering
- NSFW prompt flags
- Selfie generation
- Extended memory

Starter & Pro Tiers
Real, scaling marginal costs

Funds the heavy compute:

- LLM Chat
- Voice generation
- Vision processing



The Premium Layer

The Base Layer

Lower tiers subsidize the raw infrastructure. Higher tiers pay for emotional and experiential value that costs Mio practically nothing to deliver, creating progressive, expanding margins.

Three compounding forces will push future margins to 70 percent.

1. Prompt Engineering Compounds

Future lorebook architecture injects backstory perfectly on-demand, projected to cut an additional 30-40% of tokens.

2. Model Cost Collapse

Inference costs are plummeting. Projecting a 3-5x drop in Gemini pricing within the next 12 months.

3. Intelligent Routing

90% of traffic already routed to cheap Gemini 3 Flash. High-tier models reserved strictly for high-value ops.

Today's post-compression margins are the floor. Within 6-12 months, the multiplier effect of these three forces makes the 50-70%+ margin target inevitable.

From an unscalable prototype to an infinitely scalable economic engine.

	Early Prototype	Mio (Pre-Compress)	Mio (Post-Compress)	Projected (12 Months)
Cost per Message	Absurdly high	Negligible	~35% cheaper still	Another 3-5x price collapse
Entry Tier Profitable?	No	Top tiers only	Yes (Comfortable)	Yes (Highly Profitable)
Memory Intelligence	None	Multi-layer retrieval	+ Compressed prompts ✓	+ Self-optimizing architecture
Emotional Nuance	Rule-based	Soul-driven core	+ Relationship evolution ✓	+ Fine-tuned proprietary models

Prompt compression solved the unit economics of today. The structural trajectory of AI inference solves the unit economics of tomorrow.