v0.0.6: 4x Cheaper by Knowing Which Model to Use

TTFT Reduced: 5x

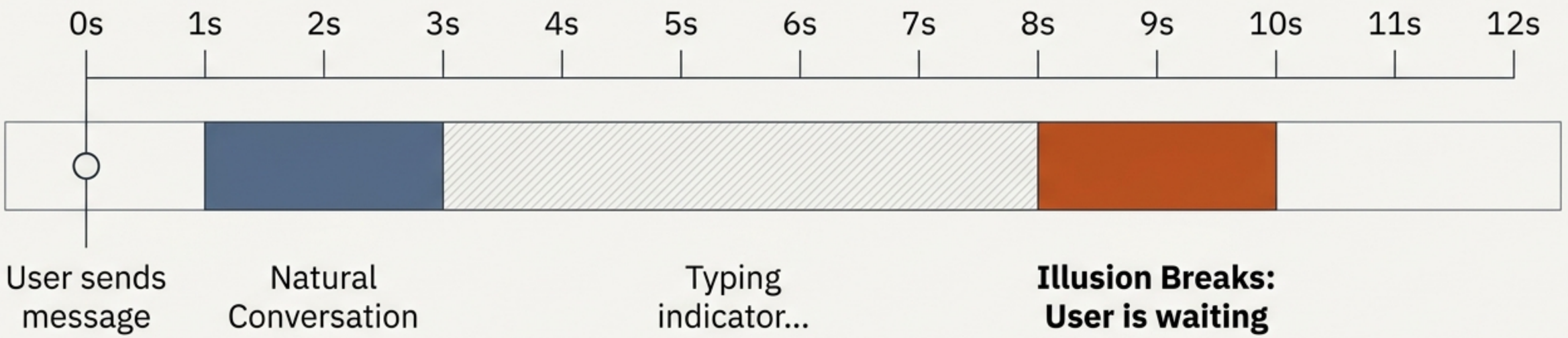
1–2 seconds vs 8-10 seconds

Cost Reduced: 4x

Gemini 3 Pro baseline offset

Project: Mio
Release: v0.0.6
Focus: Latency, Cost, Architecture
Total Commits: 6

The Conversational Illusion Death Zone



> Gemini 3 Pro Preview required 8-10 seconds for a first token when 'thinking' was enabled. For emotional companion AI, this kills the immersion.

Research First, Then Code

Metric	Gemini 3.1 Pro	Gemini 3 Flash
Input/Output Cost	~4x more expensive	Baseline
Output Speed	90.8 t/s	214 t/s
TTFT (Thinking)	8-10s	1-2s
TTFT (Minimal Thinking)	N/A	**1-2s**

Telemetry Insight

Flash provides a 4x cost reduction and 5x speed increase. But for emotional companion AI, does the quality hold up?

Thinking Mode Actively Degrades Creative Output

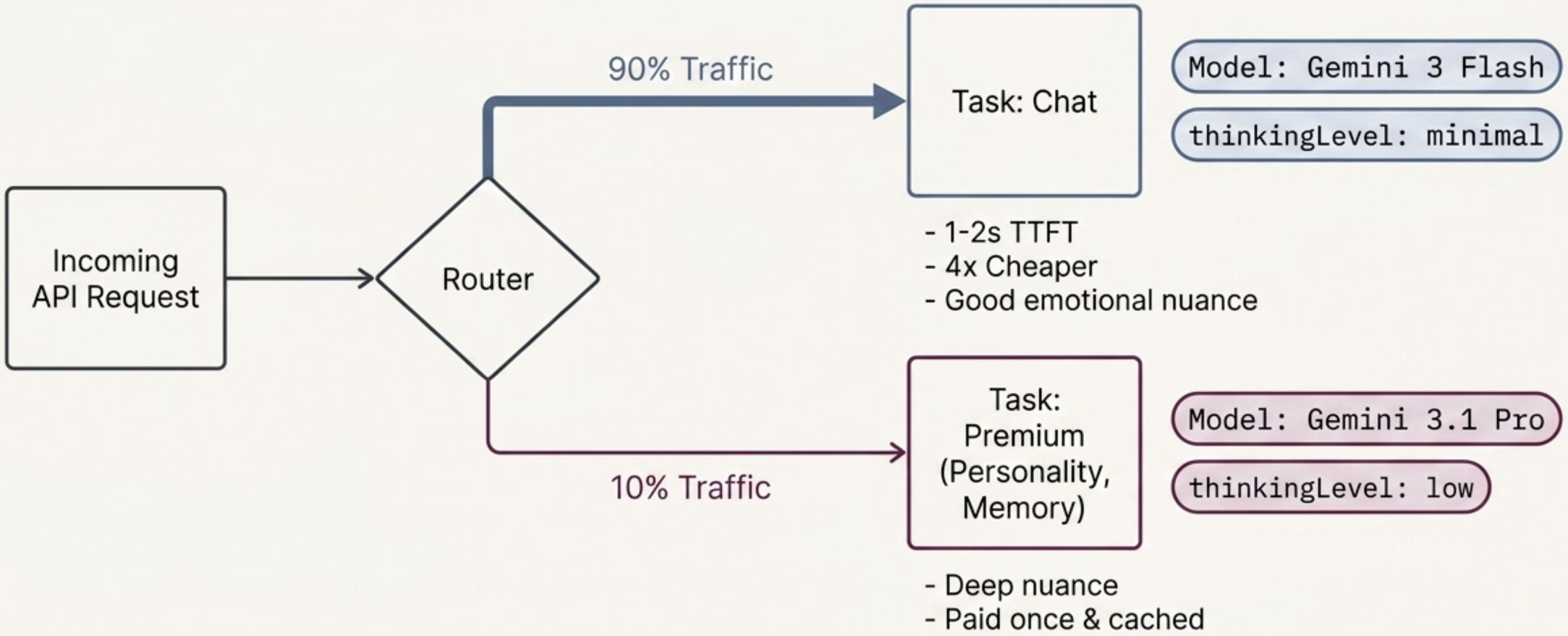
```
thinkingLevel: minimal
```

Community consensus: Flash is a side-grade to 2.5 Pro, not a downgrade. The quality gap is 1-2% on extreme emotional nuance.

Flash yields better narrative initiative and character commitment.

Gemini’s reasoning mode makes emotional and creative output worse, not better.

The 90/10 Task Routing Architecture



By routing by task type, we cut costs on 90% of traffic while actually upgrading premium tasks from 3 Pro to 3.1 Pro.

Multi-Modal Polish Requires Parameter Precision

Vision Output

Before

I see a colorful game screen.

After

Looks like you're playing Genshin Impact — how's the new region?

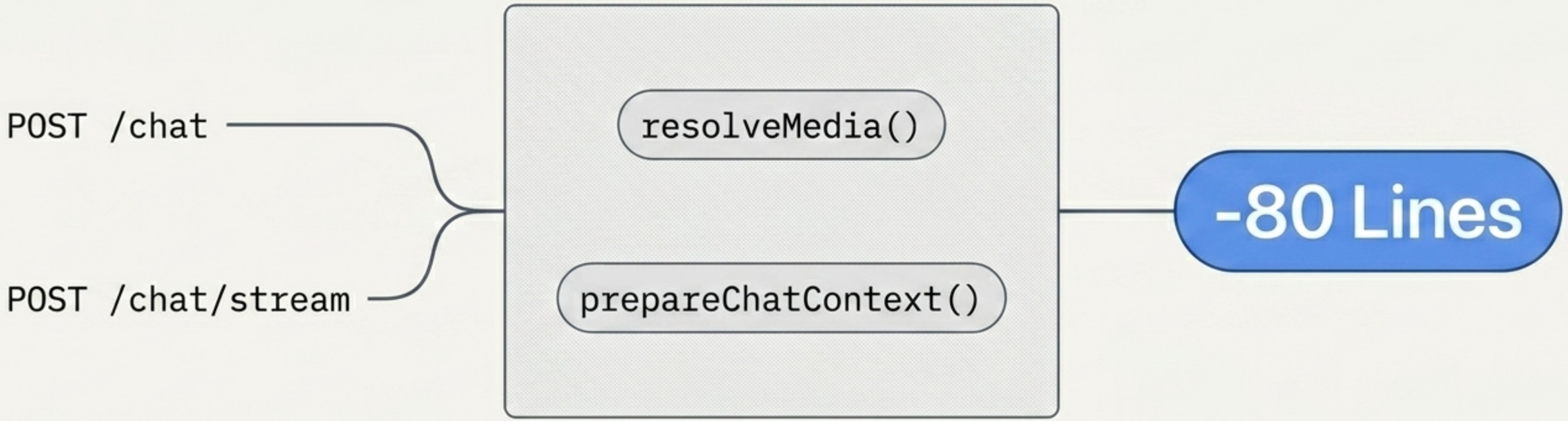
`thinkingLevel: low`

Chinese Transcription Fix

⚠ **Error:** Mandarin input hallucinating English artifacts.

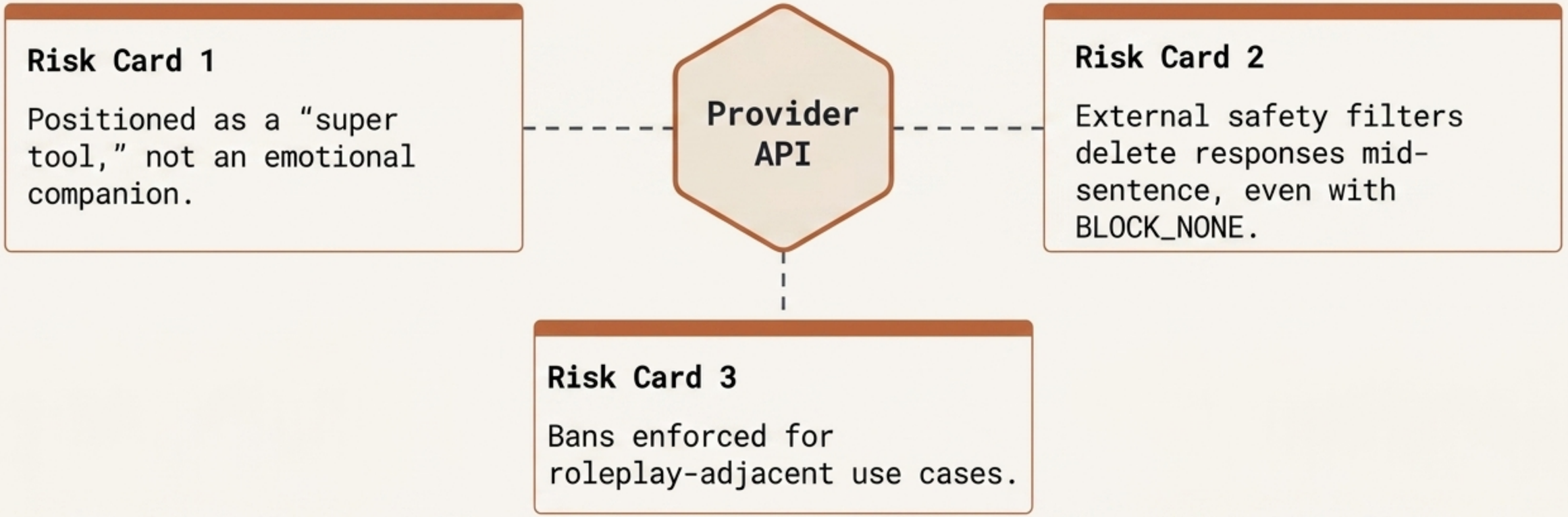
- ✓ Added explicit monospace language: `'zh'` parameter to OpenAI transcriber.
- ✓ Added Gemini fallback prompt for slang and proper nouns.

Eliminating Technical Debt with a DRY Refactor



v0.0.5 left ~80 lines of duplicated logic for media resolution and context preparation. These shared helpers execute for both endpoints.

Platform Risk Necessitates Model-Agnosticism

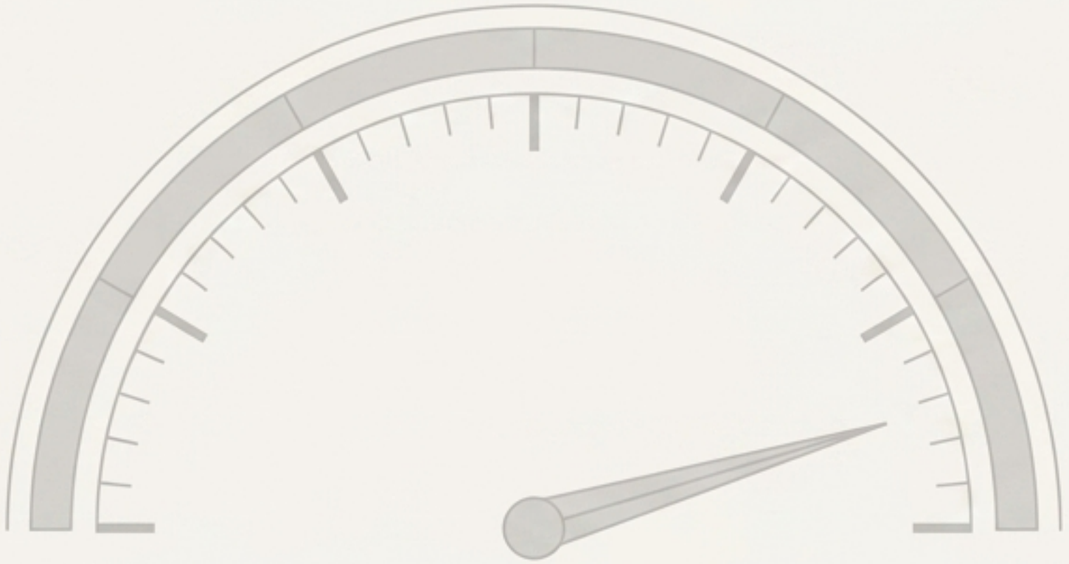


Strategic Takeaway: Mio must remain model-agnostic to swap providers instantly. The new routing layer isolates model logic from application logic.

Six Commits. Massive ROI.

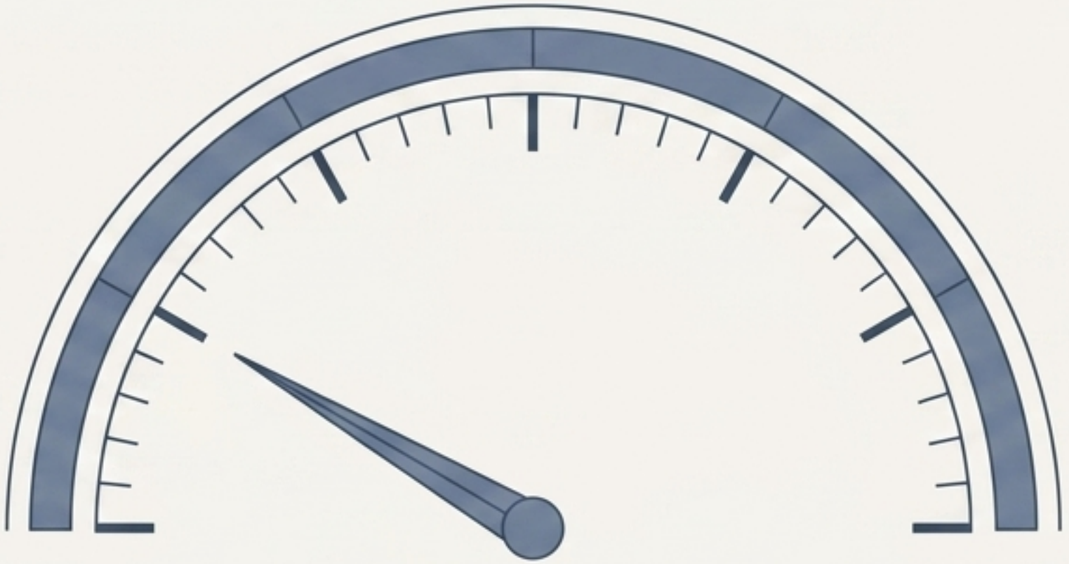
`Chat: 3 Pro → 3 Flash (minimal)`	4x cost reduction, 1-2s TTFT
`Premium: 3 Pro → 3.1 Pro (low)`	Better emotional nuance where it counts
`Vision prompts + thinking level`	Specific identification, no generic descriptions
`Chinese transcription config`	Accurate Mandarin voice recognition
`DRY refactor`	-80 lines duplicated server code
`Deployment docs`	Correct Artifact Registry commands

Data-Driven, Not Vibes-Driven



8-10 seconds

Vibes: "Pro sounds better"



1-2 seconds

Data: "Benchmarks dictate Flash"

Understand your costs before they understand you. Opting for the "`smartest`" model is a trap. `Build routing`, `benchmark relentlessly`, and let the data choose the model.