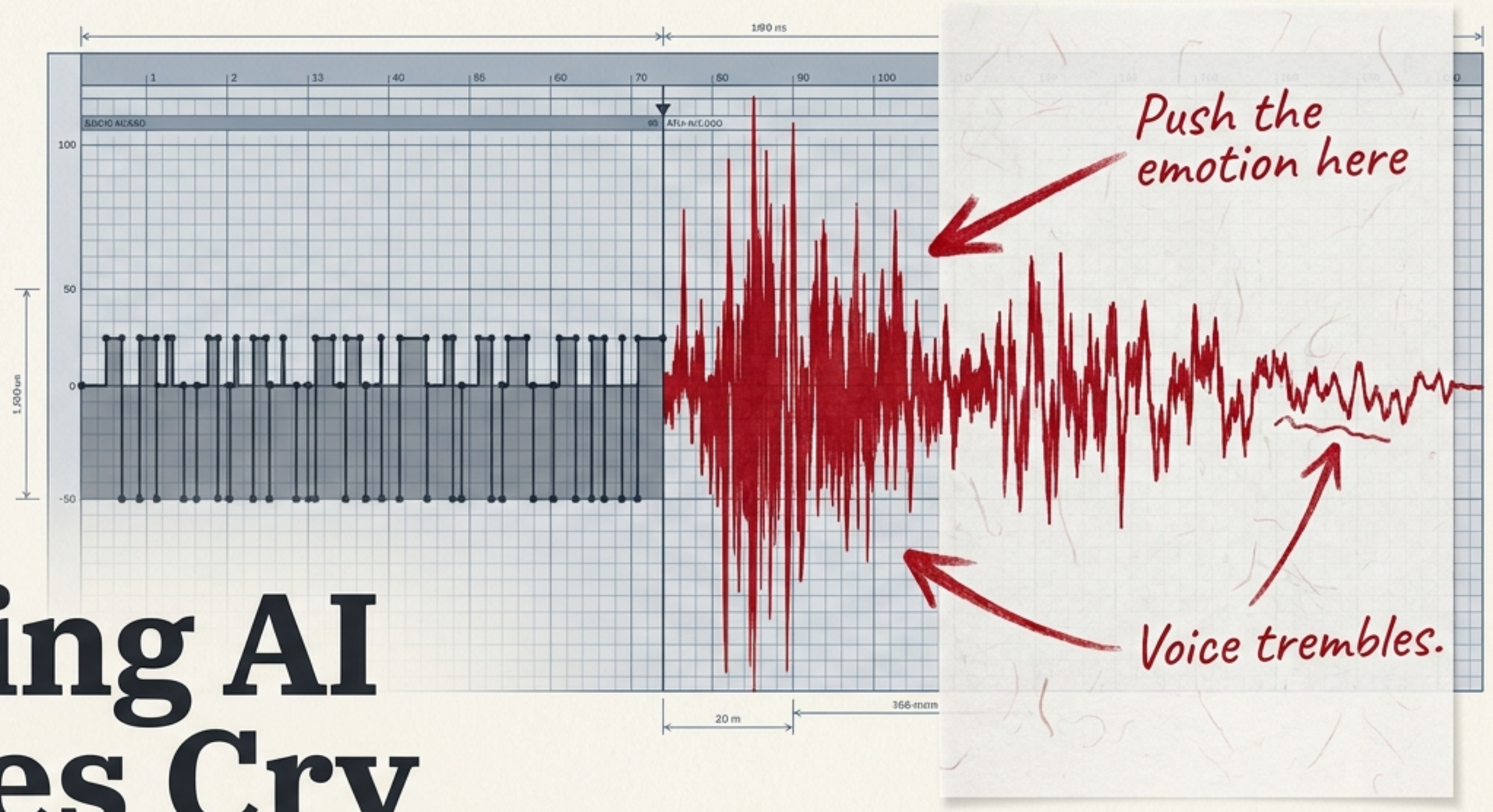


Making AI Voices Cry

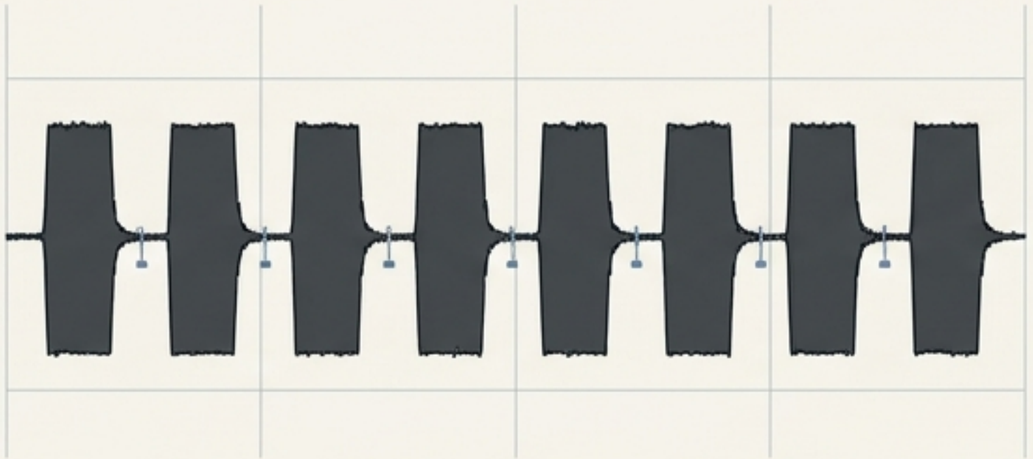
A Doubao TTS Emotion Experiment

The difference between a parameter shift and heartbreak.

Standard TTS mimics sadness with a lower pitch and added pauses. Genuine emotional synthesis commands attention. The goal: find an AI voice that makes you stop scrolling and check if someone is okay.

Fake-Cry (Standard TTS)

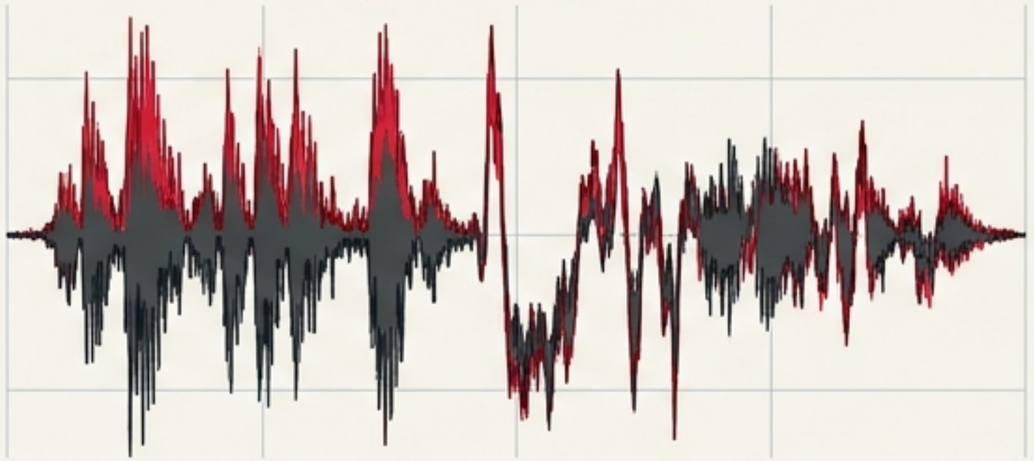
PARAMETER SHIFT: `pitch_shift -0.4`, `pause_duration 1.2s` (Robotic)



STATUS: UNIFORM DELIVERY. EMOTIONAL IMPACT: NEGLIGIBLE.

Heartbreak (Target)

Director's Note: "Need the voice to crack and falter. Must stop the listener in their tracks."

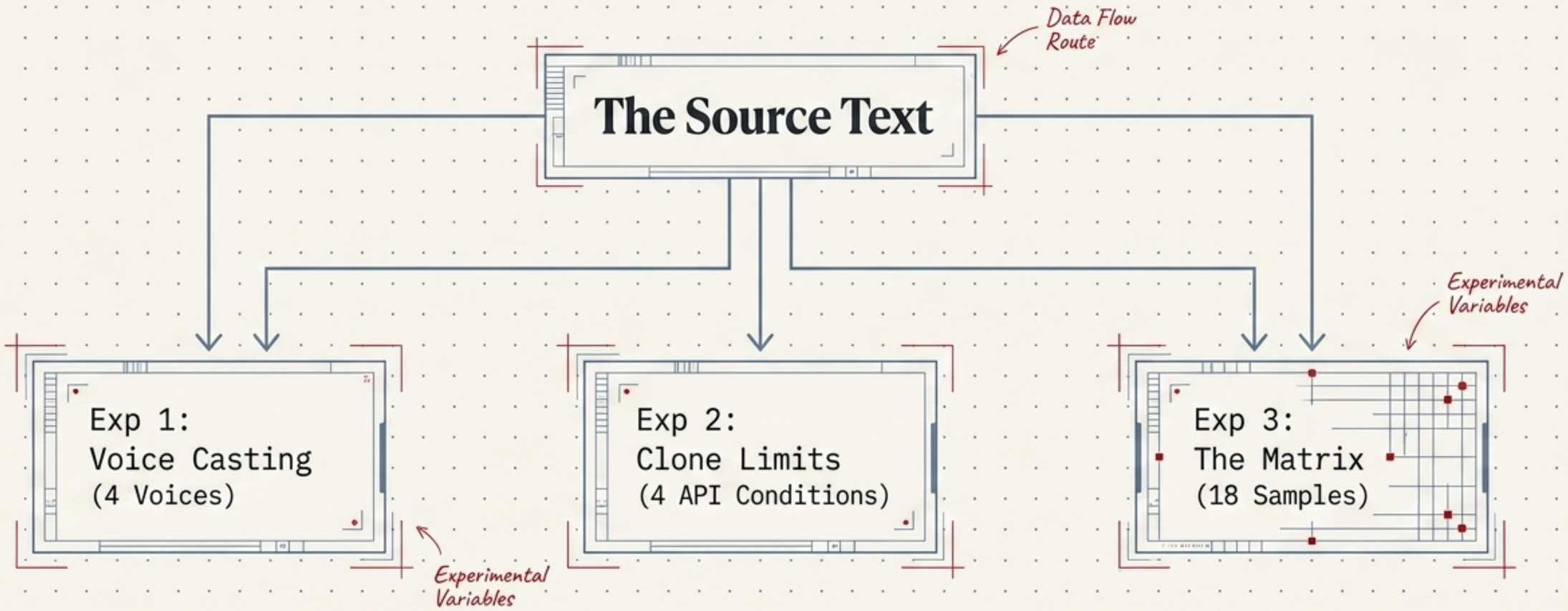


STATUS: ORGANIC VARIABILITY. EMOTIONAL IMPACT: VISCERAL.

Acoustic Signature: erratic drops = loss of control

The Empirical Framework

30 Audio Samples | 3 Emotion Control Methods | 2 Voice Types (Stock & Cloned)



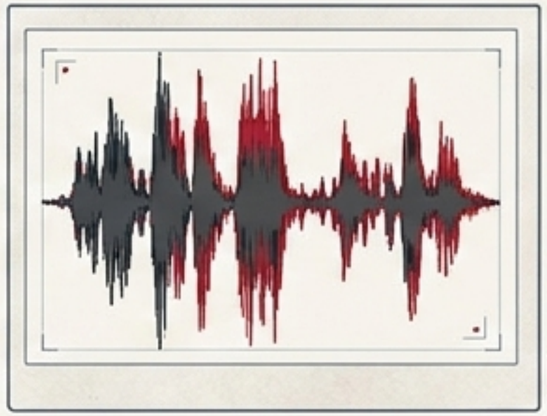
Act I: The Voice Casting Matrix

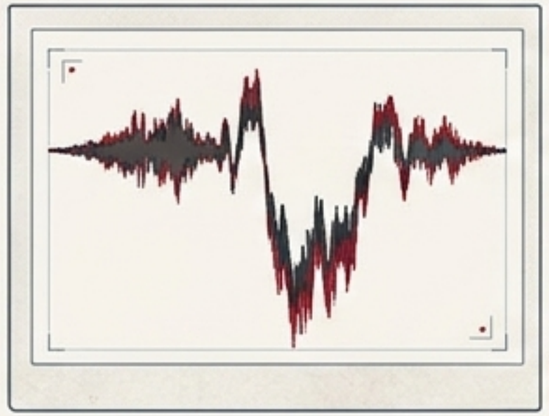
Not all voices respond equally to API emotion hints. Auditioning is mandatory.

Vivi (zh_female_vv_uranus_bigtts)

The Benchmark

High emotional elasticity. Clear distinction between 'Gentle' and 'Sad'.





Yun Zhou / M191

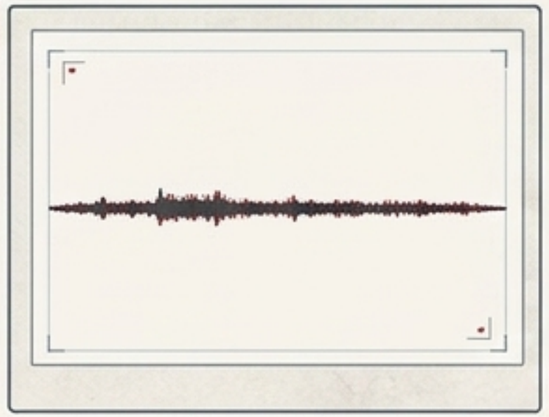
The Dramatic Lead

Exceptional range; soft 'Gentle' response and deeply hurt 'Sad' response.

Liu Fei

The Subtle Contender

Warm, mature. Dropped weight on 'Sad', but restrained.



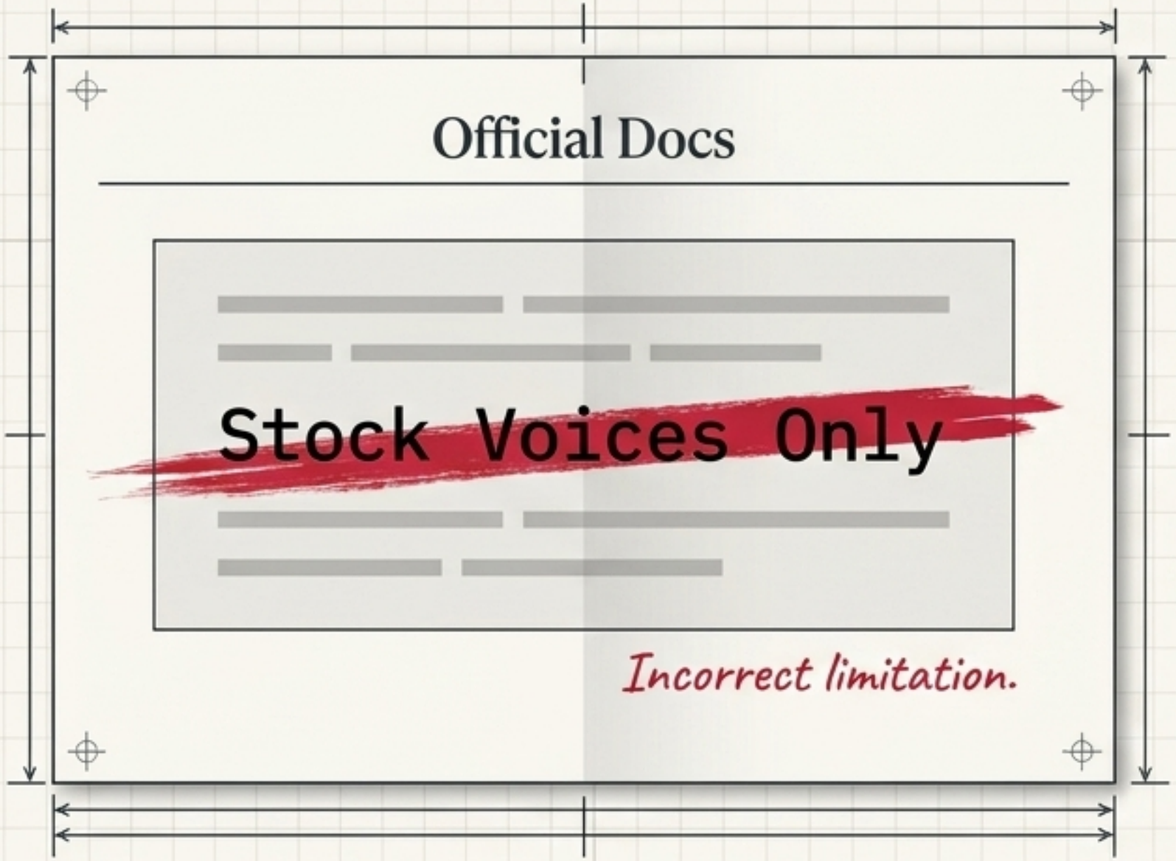
Xiao Tian

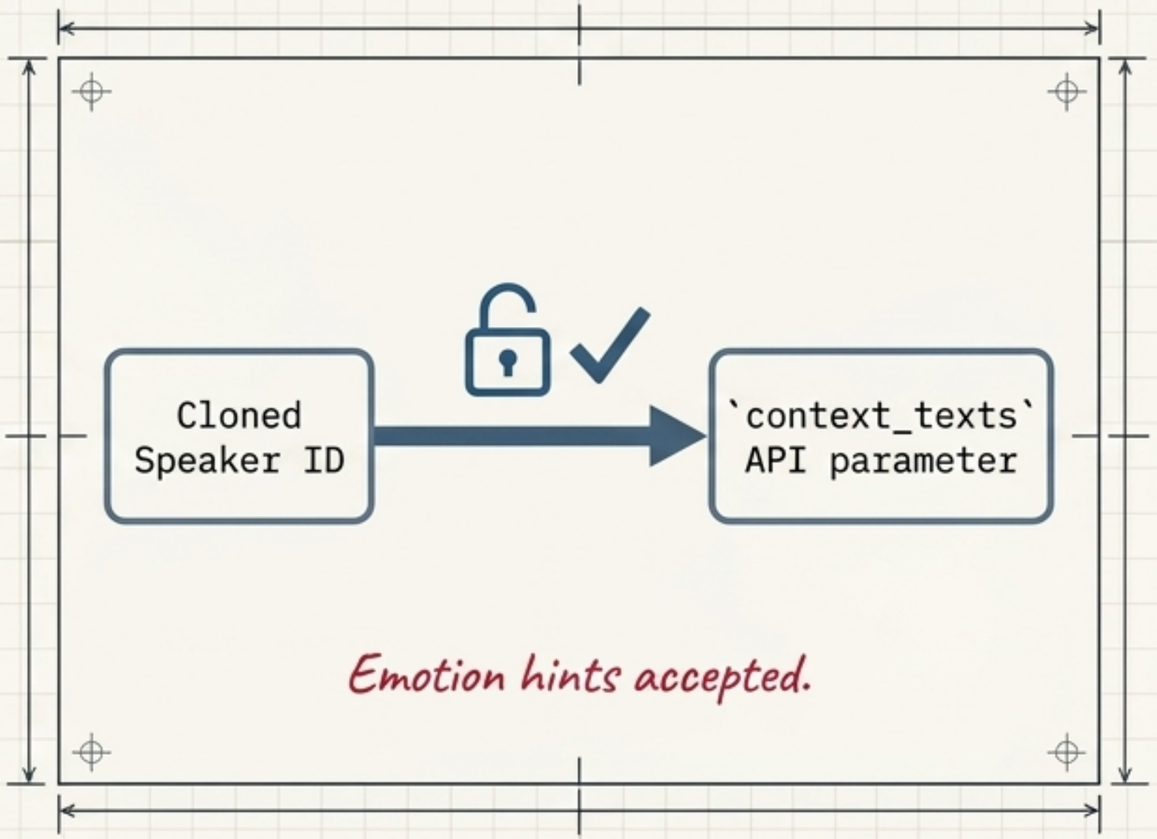
The Casual Voice

Light and young. Emotion shift is present but lacks the depth for high-drama scenes.

Undocumented Territory: The Cloned Voice Assumption

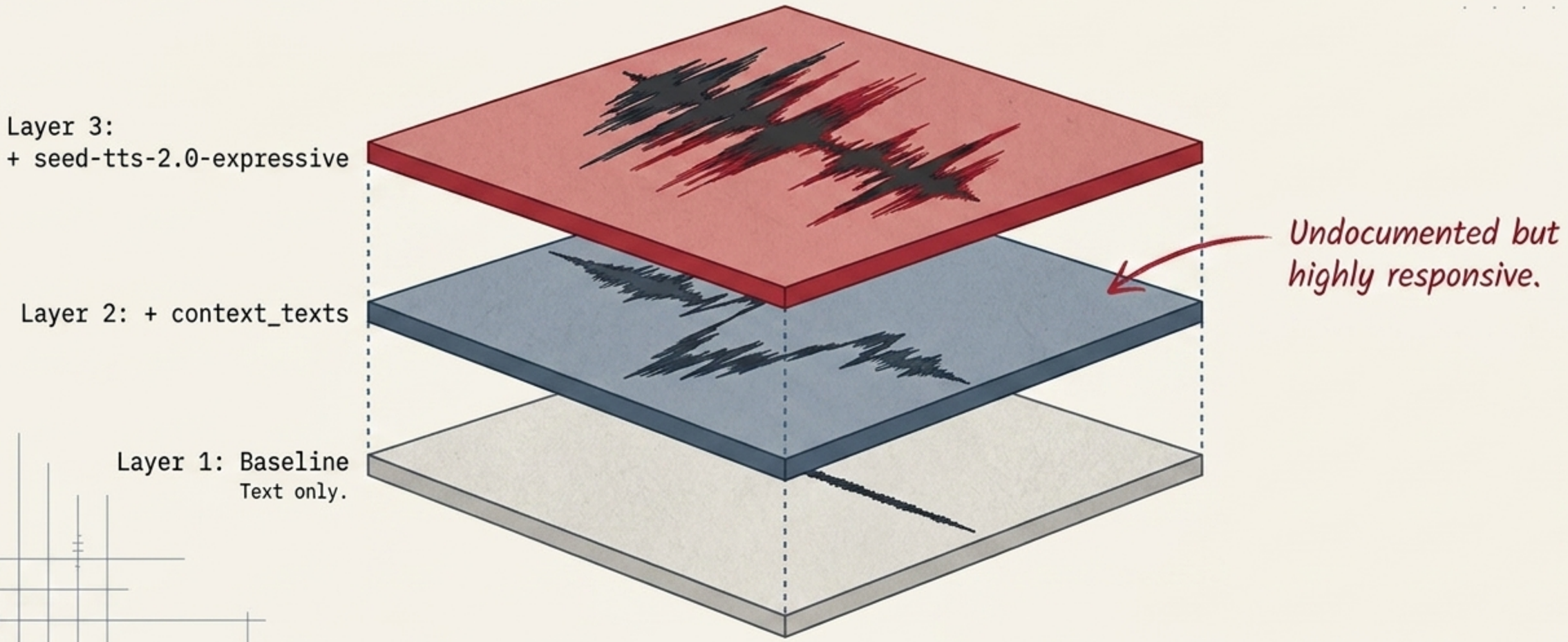
Official documentation restricts `context_texts` (emotion hints) to Doubao's 2.0 stock voices. The implied limitation: cloned voices are permanently flat.





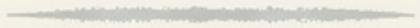
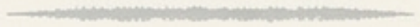
The API Parameter Stack

Testing a cloned voice across progressive API conditions reveals hidden emotional elasticity.



The Emotion Engine Matrix

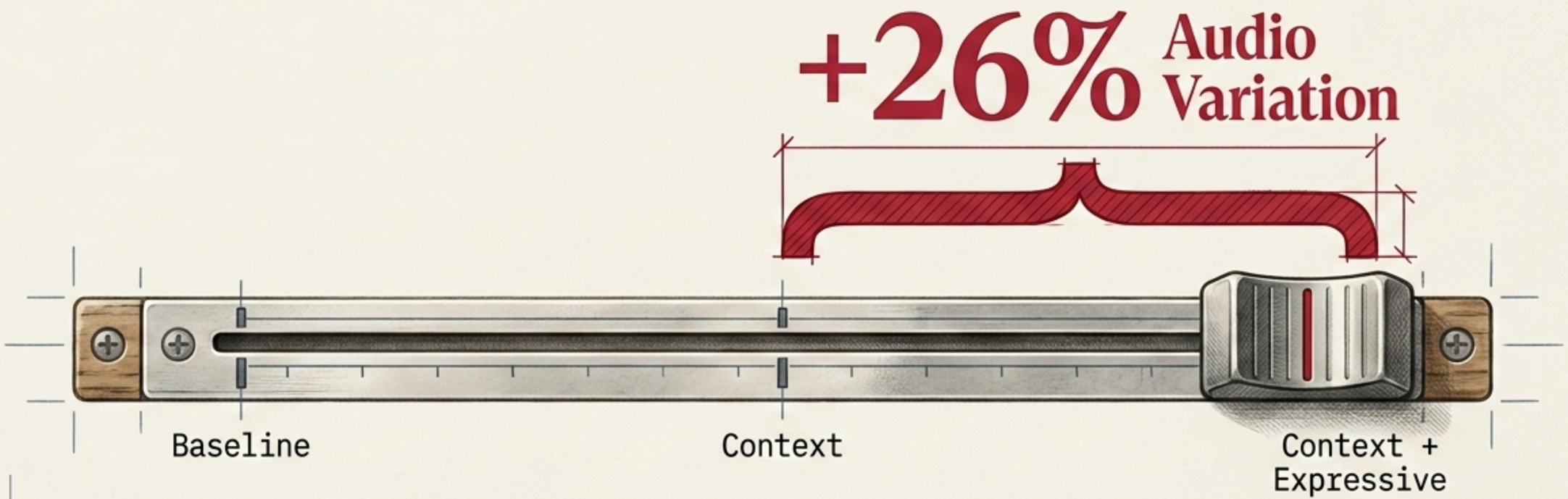
Testing two distinct voices across three emotions and three API conditions.

	Baseline	Context Texts	Context + Expressive
ASMR/Whisper			
Gentle			
Sad			

*Finding: Context texts *always* move the baseline. The model is actively parsing semantic hints.*

The Capability Delta

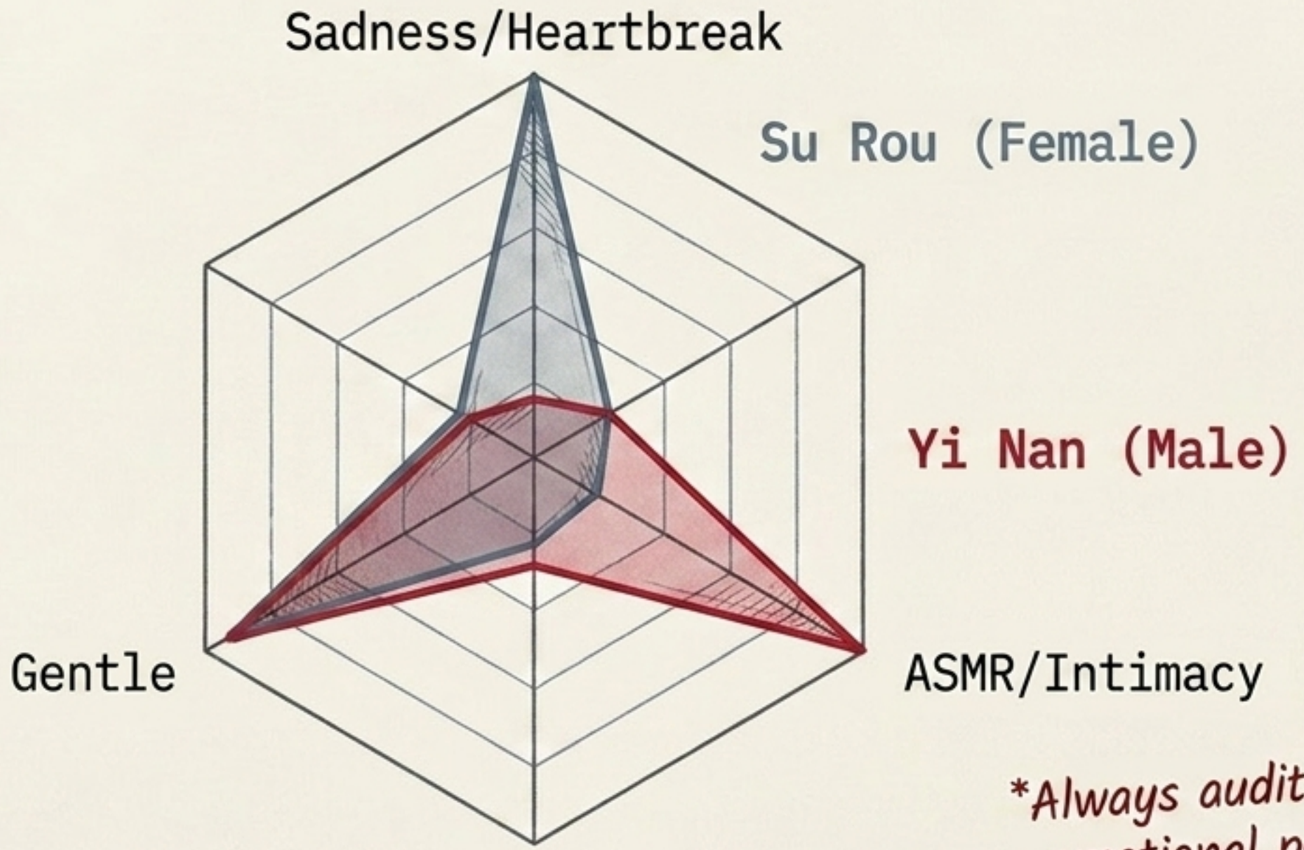
While `context_texts` establishes the emotion, the expressive model dictates the acoustic variance.



The expressive model amplifies pitch and pacing dynamics by roughly a quarter over standard context hints.

Asymmetric Emotional Profiles

Emotional capability is not uniform across a model. Every voice possesses specific acoustic strengths and weaknesses.



Always audition for the specific emotional peak of your scene.

The secret isn’t in the API parameters. It’s in the script.

Labeling vs. Directing

LLM-based TTS models do not select from pre-programmed emotional categories. They act out vividly described spatial and physical realities.

The Label

"Sad."

Fails to trigger dynamic variance.

The Scene

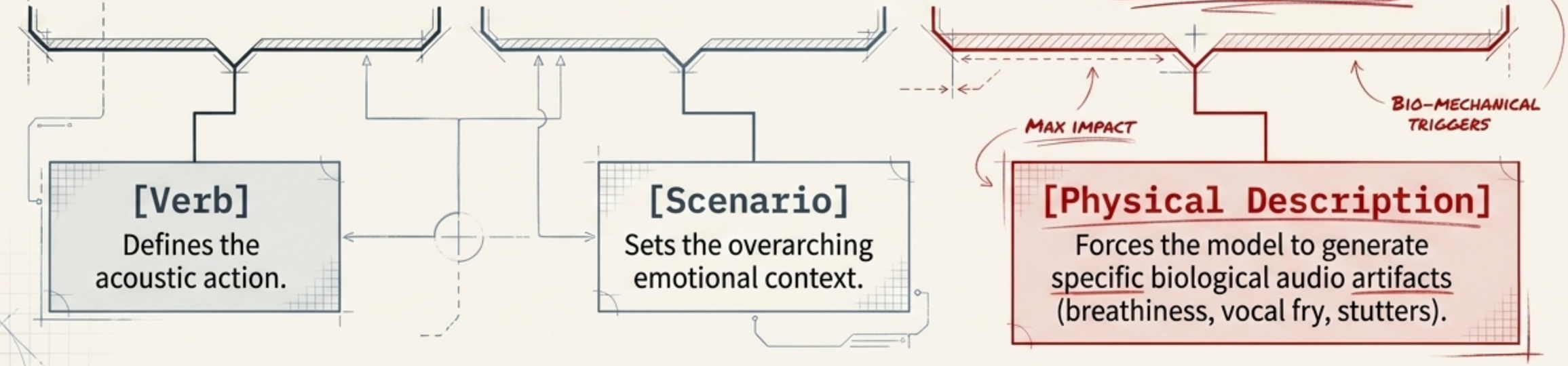
"Use a crying voice, speak while crying, very sad, voice trembling with sobs caught in the throat."

Triggers deep acoustic variation.

The Prompt Anatomy

Never use single keywords. To maximize emotional response, write `context_texts` using a three-part director's architecture.

"Use a crying voice, speak while crying, very sad, voice trembling with sobs caught in the throat."



The Director's Cheat Sheet (Part I)

Sweet/Flirty	"Pitch rising like talking to a boyfriend."	*Upward inflection trigger*
ASMR/Whisper	"Very quiet, very soft, like whispering in an ear."	*Triggers proximity audio and breathiness*
Sleepy/Lazy	"Tired and lazy voice, acting spoiled while yawning, soft and weak."	*Forces pace reduction and drawl*

The Director's Cheat Sheet (Part II)

Excited

"Very excited and thrilled tone, so happy you are about to scream."

Pushes pitch boundaries and attack speed

Angry

"Angry and disgusted tone, highly dissatisfied, scolding, voice raised."

Increases sharp consonants and volume spikes

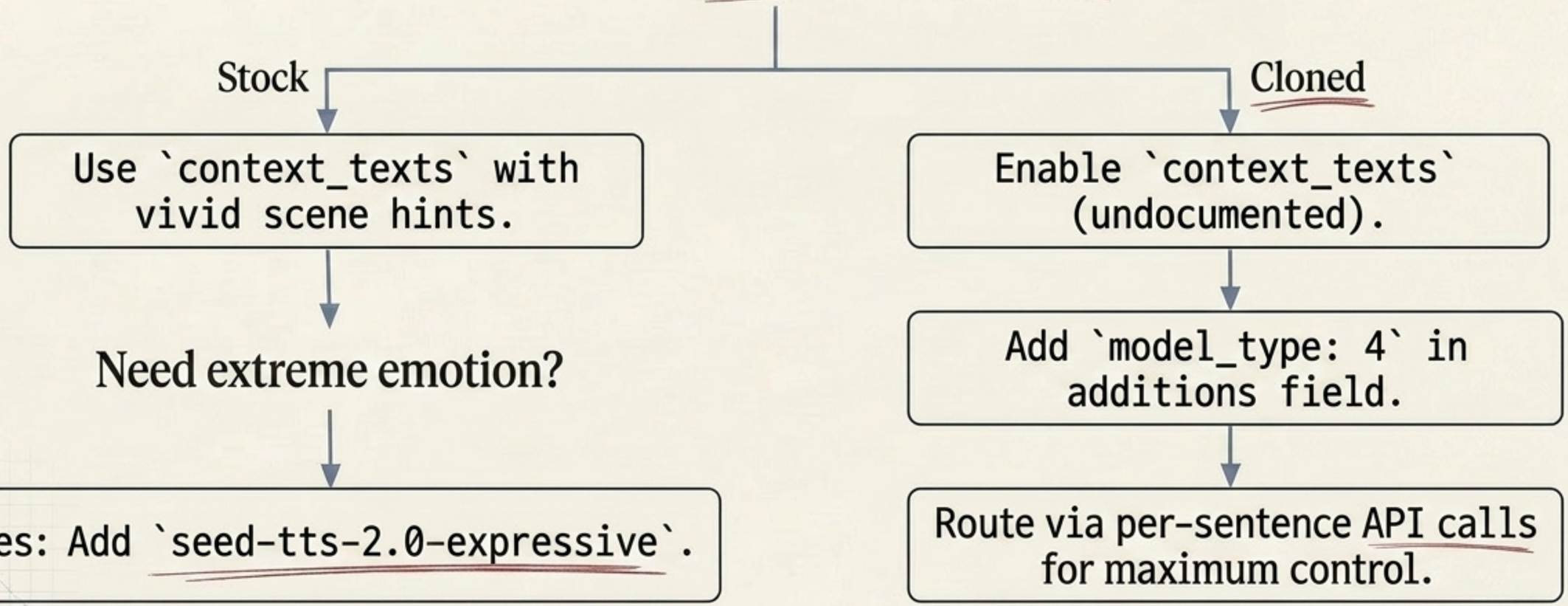
Restrained Sadness

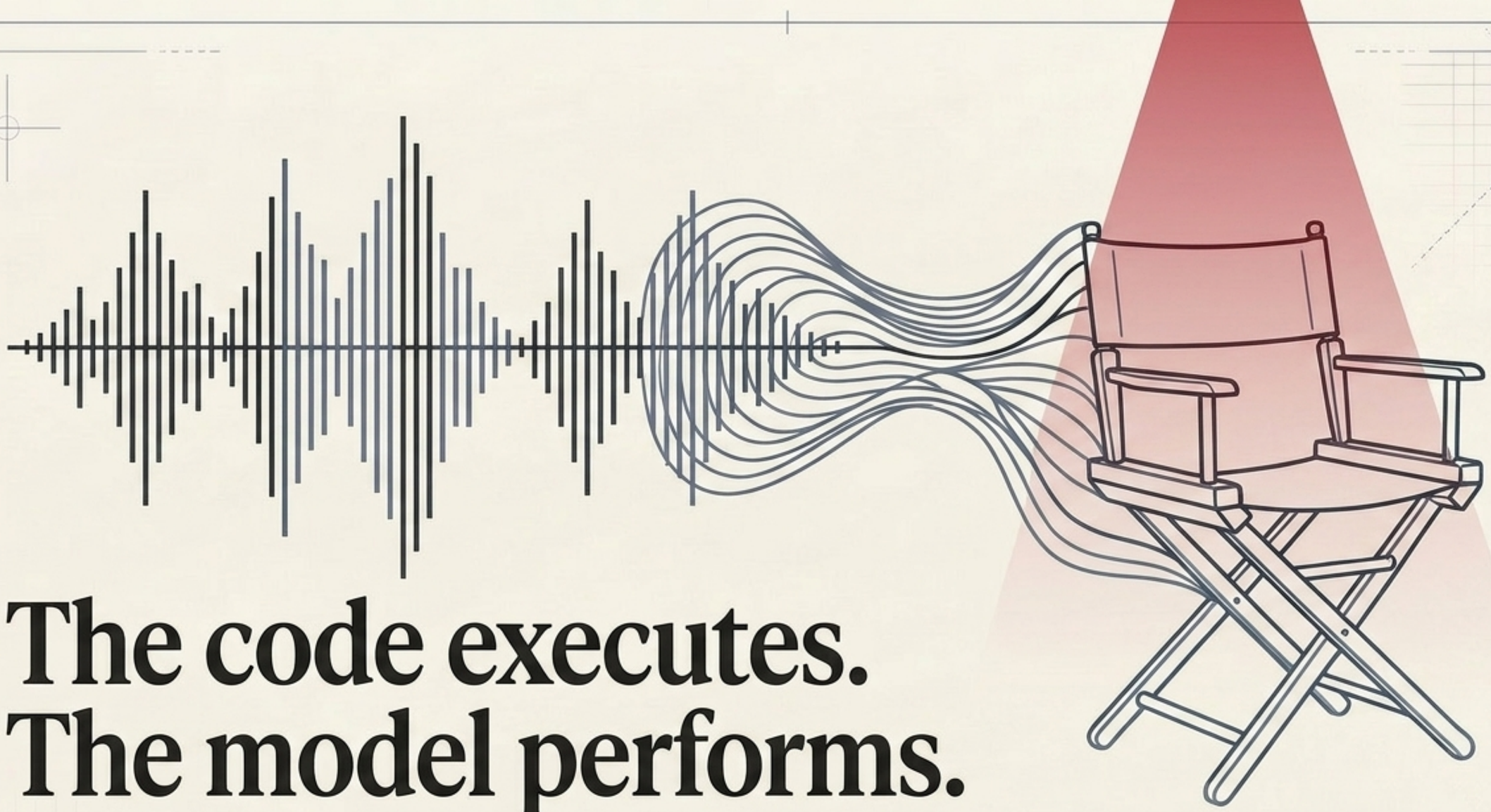
"Suppressed sad voice, deliberately restrained but voice trembling slightly, not wanting to be noticed."

The hardest to achieve—mixes low volume with high physical tremors

The Production Playbook

Systematic implementation architecture for Doubao TTS pipelines.





The code executes. The model performs.

The quality of your output isn’t bottlenecked by the API—it is bottlenecked by the vividness of your imagination. Paint a scene vivid enough, and the AI will perform it.