



1. The Perfect Voice Clone: Setup & Constraints

Seed-ICL 2.0 Voice Cloning Requirements

To generate a flawless Speaker ID (S_xxxxxxxx), the foundation must be clean.

The Golden Rules of Uploads:

- ★ **Duration:** 10-30 seconds maximum.
- ★ **Purity:** Single speaker only. No overlapping speech.
- ★ **Environment:** Zero background music, minimal room noise.
- ★ **Format:** MP3 or WAV.

Post-Setup Architecture

- **App ID:** Numeric identifier (1234567890)
- **Resource ID:** seed-icl-2.0
- **Model Type:** 4 (Required)

2. The Magic of Emotion Control

Feature	Legacy Providers (e.g., MiniMax)	Doubao Seed-ICL 2.0
Control Interface	Numeric Sliders (Speed, Pitch)	Natural Language (context_texts)
Emotion Fidelity	Basic / Keyword-based	High (Follows complex scene descriptions)
Cost Profile	~2.3x more expensive	1x (Baseline) + Free Emotion Tokens



How context_texts Works:

Paint a vivid scene in natural language instead of generic keywords.

- Weak Prompt: “开心” (Happy) -> Minimal effect.
- + Strong Prompt: “用撒娇甜蜜的语气” (Speak in a sweet, coquettish tone) -> Massive emotion shift (+25-65% variance).

3. Gotcha #1: The Flat Voice Mystery

The Problem: You pass a beautiful context_texts prompt, but the voice is robotic.

The Cause: The API expects additions to be a JSON-serialized string, not an object. It silently ignores plain objects.

✗ **WRONG:** API silently ignores this

```
additions: {  
  model_type: '4',  
  context_texts: ['用撒娇甜蜜的语气']  
}
```

✓ **CORRECT:** Stringify the payload

```
additions: JSON.stringify({  
  model_type: '4',  
  context_texts: ['用撒娇甜蜜的语气']  
})
```



4. Gotcha #2: The Parsing Crash

The Cause: Doubao streams NDJSON (Newline-Delimited JSON). Standard `response.json()` will instantly crash.



The NDJSON Parsing Pipeline:

1. **Read:** Read response body as raw text.

2. **Split:** Split the text by `\n` (newlines).

3. **Parse:** Parse each non-empty line independently.

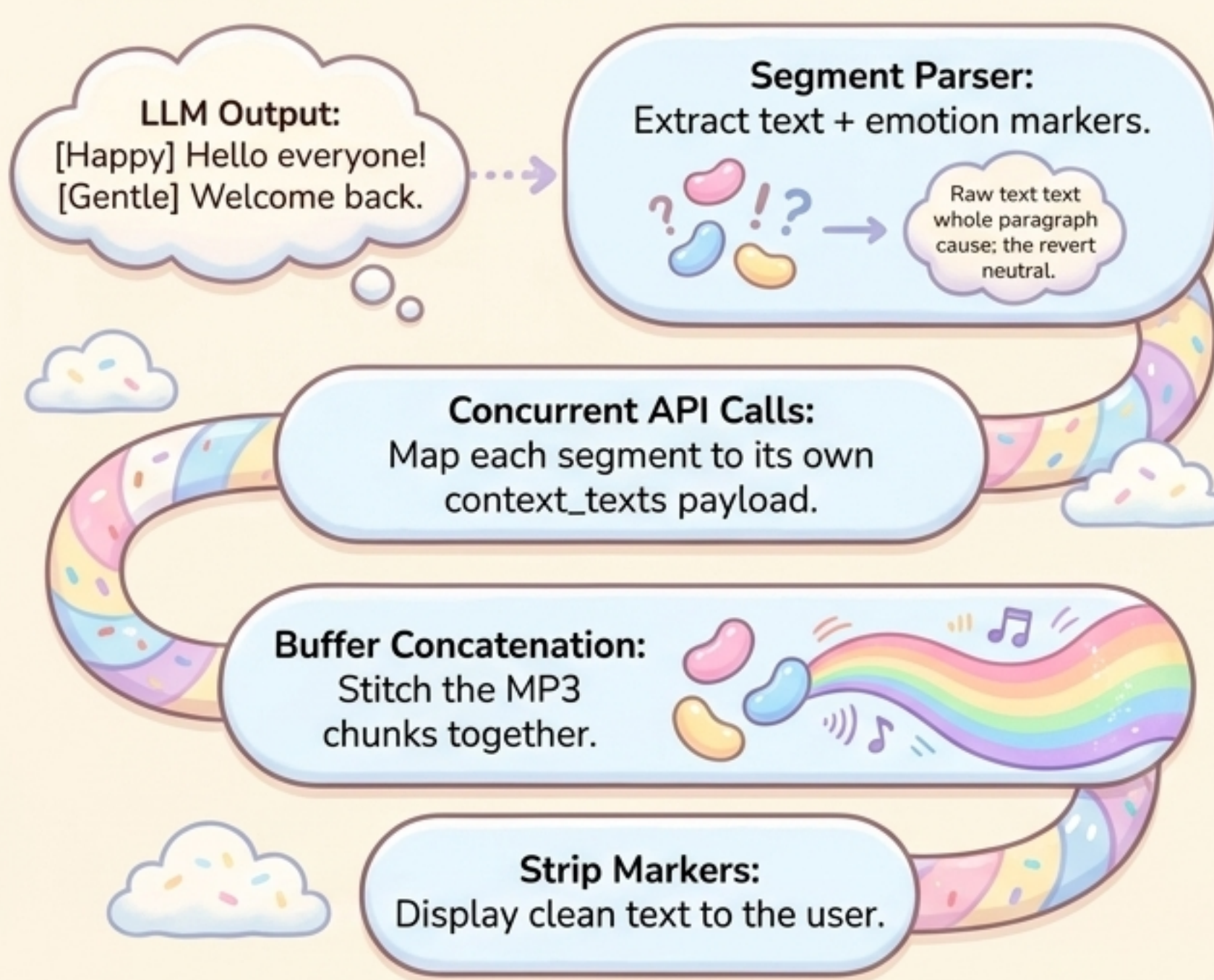
4. **Extract:** If `code === 0`, decode data from base64 and append to audio chunk array.

5. **Terminate:** If `code === 20000000`, the stream is finished. Stop parsing.

5. The Triumph: Per-Sentence Emotion Pipeline

Why Single Calls Fail: Sending a whole paragraph causes emotion to drift. The first sentence is great; the rest revert to neutral.

The Fix: One Call Per Sentence.



Result: Flawless, shifting emotions with ~2-4s latency per segment.