The Hard Half of Transcription

Turning sound to text is solved. Knowing *who* spoke is not.

- **The Problem:** Speaker Diarization.
- **The Reality Check:** A pristine 28-minute recording with exactly two people.
- **The Result:** The model confidently identified five different speakers.

V1: The Chunk-and-Stitch Drift

Gemini Chunking

Split long audio (>15m) into **12-minute pieces** to prevent the model from hallucinating. Create a **60-second overlap** to stitch them together.

Seam Drift

Every chunk **numbers its own speakers from zero**. If the stitching algorithm votes wrong on just *one* seam, every subsequent label shifts. The count **creeps endlessly upward**.

V2: Over-Splitting the “Hmm”s

Global diarization stops drift, but creates a new enemy.

- **The Shift:** Ran a global pass (Senko/Doubao) to keep speaker numbering consistent end-to-end.
- **The Bug:** General ASR models treat one-word backchannels (“mm,” “yeah,” “right”) as brand new people.
- **The Lesson:** General models over-split rather than merge. They don’t know the difference between a polite nod and a new voice.

The Solution (and its hidden trap)

Volcano's Lark Minutes natively handles exactly 2 speakers!
But beware the Auth Trap.

Old Console (AppID)

New Console (x-api-key)

Old Console

Uses: AppID + AccessToken
Result: Resource Not Granted error on 2.0 endpoints.

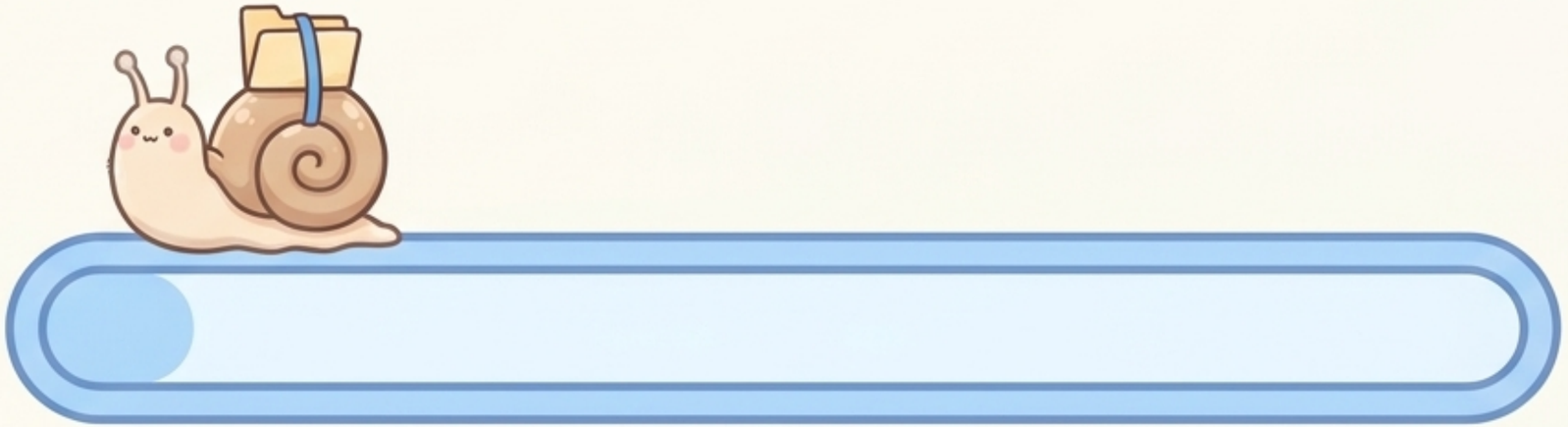
New Console

Uses: A single x-api-key
Result: Instant success for Lark Minutes & 2.0 models.

Don't mix them up! The docs bury this distinction.

The Final Boss: 34 KB/s

Lark Minutes requires a public URL, meaning local audio must be uploaded to Shanghai Object Storage first.



The Bottleneck: Cross-border single-stream uploads crawled at **34 KB/s** (4 mins for 8MB).

The Fix: Parallel multipart upload!

Cross-border throttles are per-connection, not total bandwidth. Open more connections to finish in 5 seconds!