Something Cool with TTS

Machine Learning Singapore #MSLG

- Martin Andrews = GeminiTTS @ mdda.net

13-May-2026

About Me

- Machine Intelligence / Startups / Finance
  - Moved from NYC to Singapore in Sep-2013
- 2014 = 'fun' :
  - Machine Learning, Deep Learning, NLP
  - Robots, drones
- Since 2015 = 'serious' :: NLP + deep learning
  - Including Papers...
  - & GDE ML; ML-Singapore co-organiser...
  - & Red Dragon AI...

MLSG
MACHINE LEARNING
SINGAPORE

Outline

- Gemini-TTS
  - Announcement
  - AI Studio playtime
  - Coding in Colab
  - Resources
- How are TTS models trained?
  - Example Open Source model
  - Voice Cloning Colab
- Wrap-up & QR-code

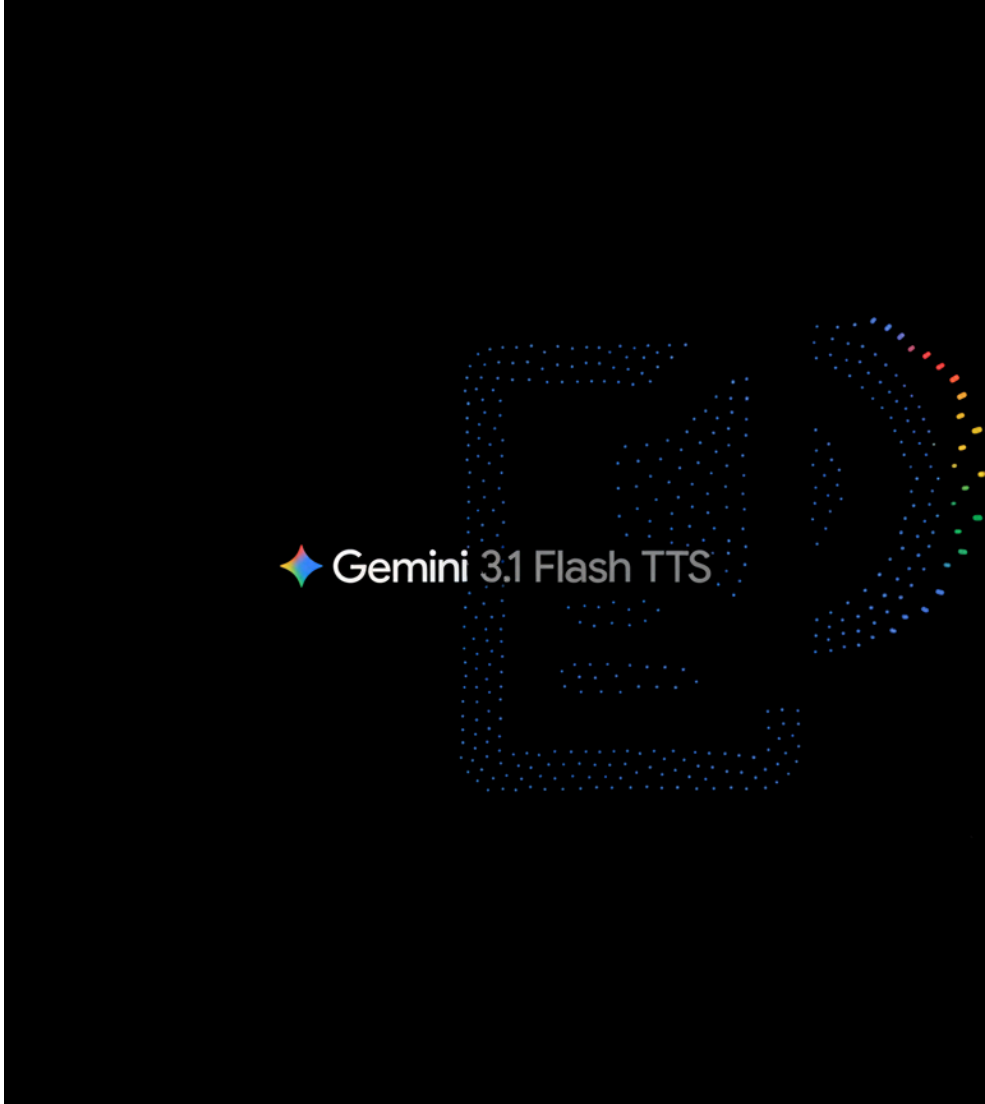
MLSG
MACHINE LEARNING
SINGAPORE

Gemini TTS

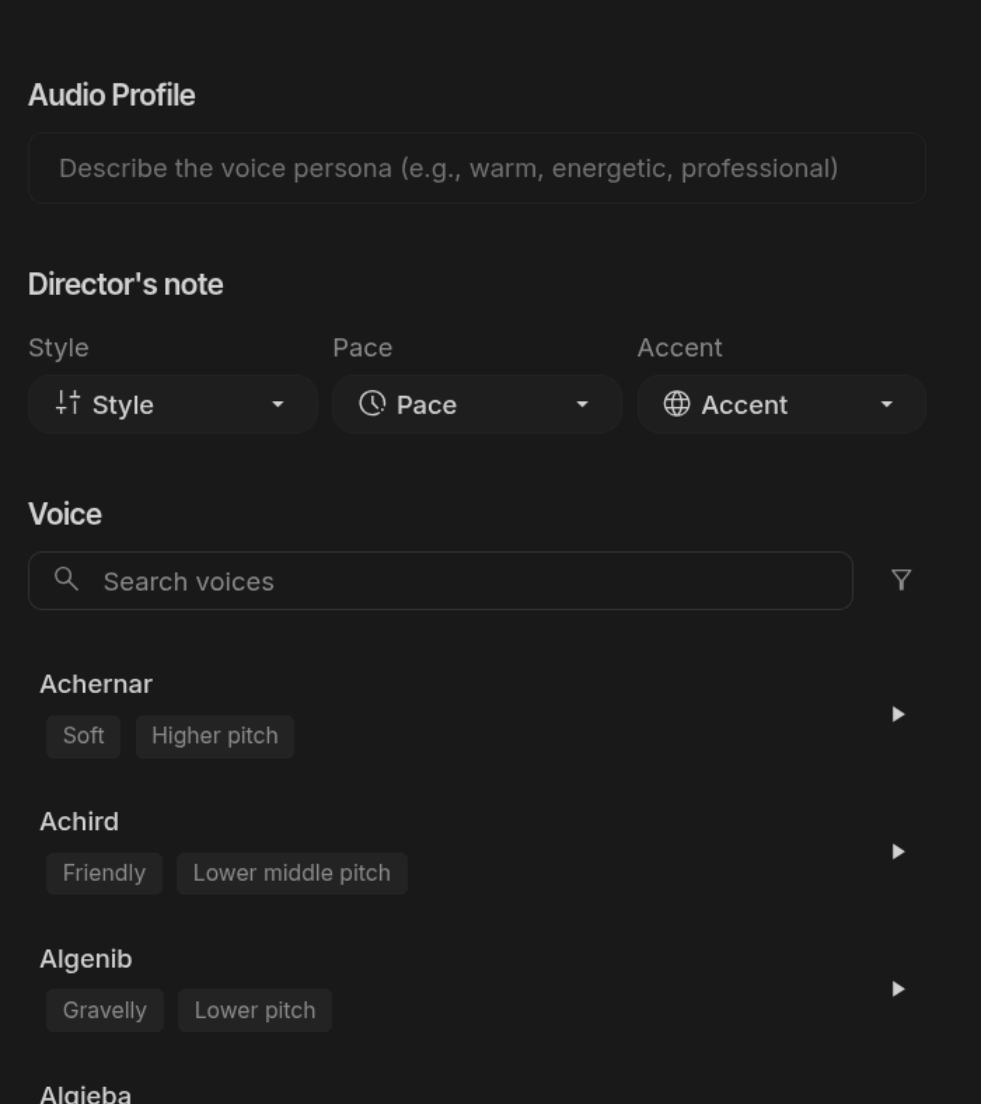
Google Official Announcement

(15-April-2026)

- "Gemini 3.1 Flash TTS: the next generation of expressive AI speech"
  - Improved speech quality and controllability
  - New audio tags for more expressive speech generation
  - Built for global scale
  - Watermarked with SynthID
- Gemini API Speech Generation Documentation



MLSG
MACHINE LEARNING
SINGAPORE

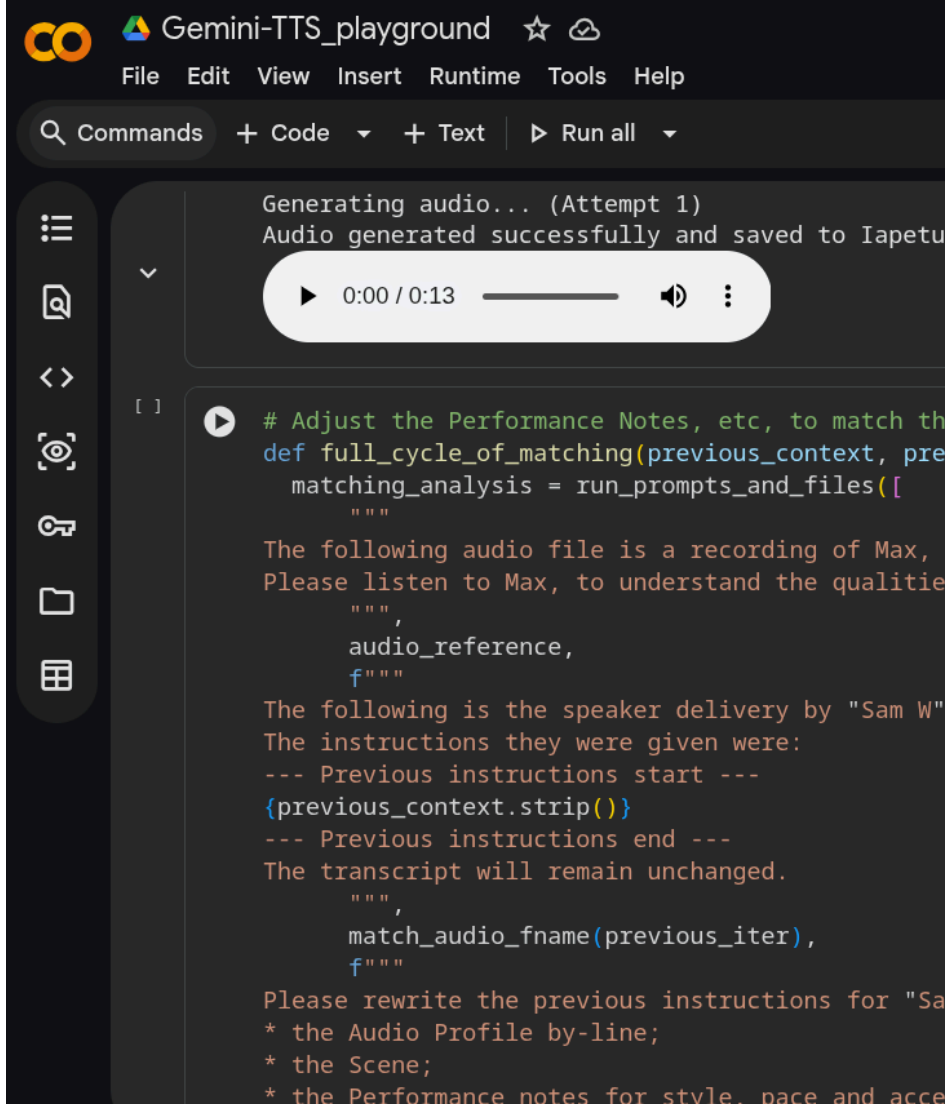


Gemini-TTS in AI Studio

- AI Studio Link
  - sadly, don't seem to be able to link to pre-populated example
- Check out:
  - Voice samples
  - Direction via RHS panel
  - REST samples

Gemini-TTS Colab DEMO

- Gemini-TTS Colab Link
- DEMO
  - Generate Speech:
    - DJ in London
    - Whisper
  - Analyse voice:
    - Rewrite prompt for Singapore DJ + Play
  - Play existing audio + Analyse (3 times)
  - Synthesise by approximation...
    - Scottish version
    - Iterate to match it more closely x4
      - finally listen to genuine article



MLSG
MACHINE LEARNING
SINGAPORE

LiveKit Products Resources Company Pricing

The Nine Rules

1. Always prepend a synthesize-speech preamble

This single paragraph is what reliably triggers the speech-synthesis classifier path instead of the "read it all" path.

```
Synthesize speech for the performance defined below. The profile, scene,
performance notes, and context are direction only. Do NOT speak them.
Speak ONLY the lines under #### TRANSCRIPT.
```

Skip this and you'll hit intermittent failures where Gemini reads your whole prompt out loud.

2. Use `#### TRANSCRIPT` as the delimiter, exact header

We tested alternatives. Lower-level headers (`##### TRANSCRIPT`) work, but `#### TRANSCRIPT` is what the official docs example uses, and it's what we saw the most reliable classifier behavior with.

3. Use short section labels, not the docs' verbose ones

The current iteration of the official Gemini docs use `### DIRECTOR'S NOTES` and `### SAMPLE CONTEXT` . **Don't.** In our testing, the literal string `DIRECTOR'S` got spoken aloud.

Use these instead:

- `### PERFORMANCE` (for Style / Pace / Accent)
- `### CONTEXT` (for sample context)

Apostrophes and multi-word section headers are classifier hazards.

4. Classify emotional register BEFORE picking audio tags

Don't use a universal tag template. Pick one register first

Other Resources

- Simon Willison Blog
  - includes a vibe-coded App
- LiveKit Gemini TTS prompting guide
  - actual real-world advice
  - ... incorporated into my Colab DEMO

TTS Model Training

Open Source Model

- "Qwen3-TTS Technical Report"
  - Hu *et al* (2026)
  - trained on over 5 million hours of speech data
    - spanning 10 languages
  - description-based control
  - 3-second voice cloning
- Apache 2.0 licensed code
  - models released : 0.6B & 1.7B

MLSG
MACHINE LEARNING
SINGAPORE

Qwen3-TTS Technical Report

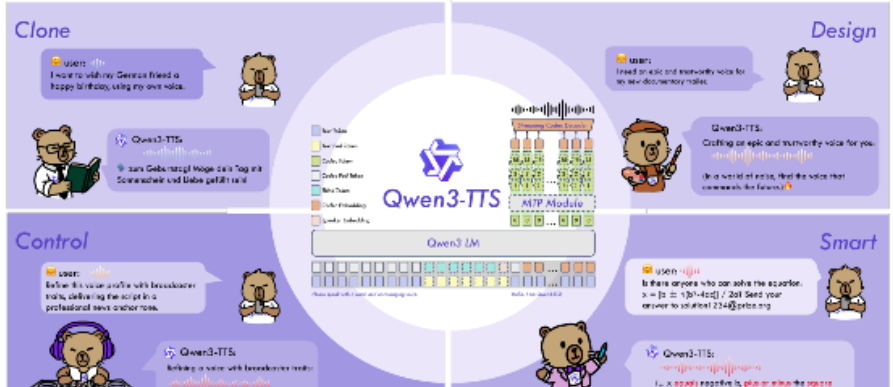
Qwen Team

https://huggingface.co/collections/Qwen/qwen3-tts
https://modelscope.cn/collections/Qwen/Qwen3-TTS
https://github.com/QwenLM/Qwen3-TTS

Abstract

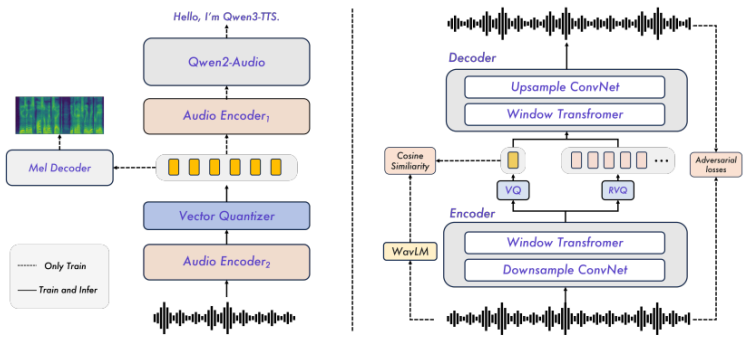
In this report, we present the Qwen3-TTS series, a family of advanced multilingual, controllable, robust, and streaming text-to-speech models. Qwen3-TTS supports state-of-the-art 3-second voice cloning and description-based control, allowing both the creation of entirely novel voices and fine-grained manipulation over the output speech. Trained on over 5 million hours of speech data spanning 10 languages, Qwen3-TTS adopts a dual-track LM architecture for real-time synthesis, coupled with two speech tokenizers: 1) *Qwen-TTS-Tokenizer-25Hz* is a single-codebook codec emphasizing semantic content, which offers seamlessly integration with Qwen-Audio and enables streaming waveform reconstruction via a block-wise DiT. 2) *Qwen-TTS-Tokenizer-12Hz* achieves extreme bitrate reduction and ultra-low-latency streaming, enabling immediate first-packet emission (97 ms) through its 12.5 Hz, 16-layer multi-codebook design and a lightweight causal ConvNet. Extensive experiments indicate state-of-the-art performance across diverse objective and subjective benchmark (e.g., TTS multilingual test set, InstructTTSEval, and our long speech test set). To facilitate community research and development, we release both tokenizers and models under the Apache 2.0 license.

1 Introduction



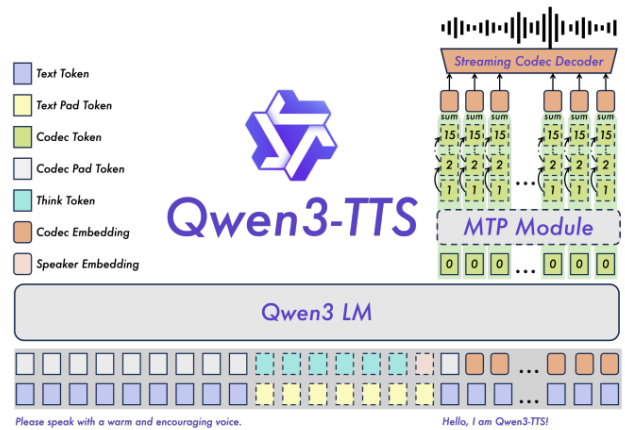
Architectural Highlights

- Qwen-TTS-Tokenizer-12Hz
  - multi-codebook tokenizer
  - jointly optimized semantic and acoustic streams
    - high-level semantic content
    - acoustic detail, prosody, and others



(a) Qwen-TTS-Tokenzier-25Hz (b) Qwen-TTS-Tokenzier-12Hz

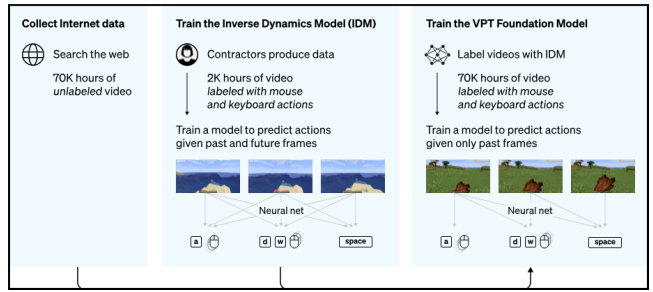
- LLM with MTP for audio tokens
  - LLM carries text
    - + (optional) speaker embeddings
  - MTP generates 16 audio tokens per step
    - audio frame rate much higher than LLM

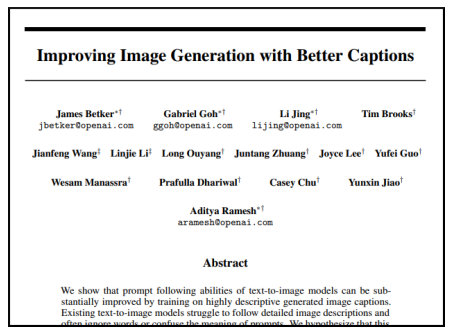


MLSG MACHINE LEARNING SINGAPORE

How to train 'performance notes' ?

- Train voice for self-reconstruction
  - can transfer voice characteristics
    - separate from Transcript
  - measure voice similarity between I/O
- Train a model to *generate* 'performance notes'
  - ~same idea as :
    - Learning to play Minecraft with Video PreTraining (2022)
    - Improving Image Generation with Better Captions (2023)
- Can then create synthetic data at scale
  - still need examples of:
    - same-speaker, different-setting
    - same-speaker, different-emotion




Improving Image Generation with Better Captions

James Betker*† jbetker@openai.com
Gabriel Goh*† ggoh@openai.com
Li Jing*† lijing@openai.com
Tim Brooks†
Jianfeng Wang† Linjie Li‡ Long Ouyang† Juntang Zhuang† Joyce Lee† Yufei Guo†
Wesam Manassra† Prafulla Dhariwal† Casey Chu† Yunxin Jiao†
Aditya Ramesh*† aramesh@openai.com

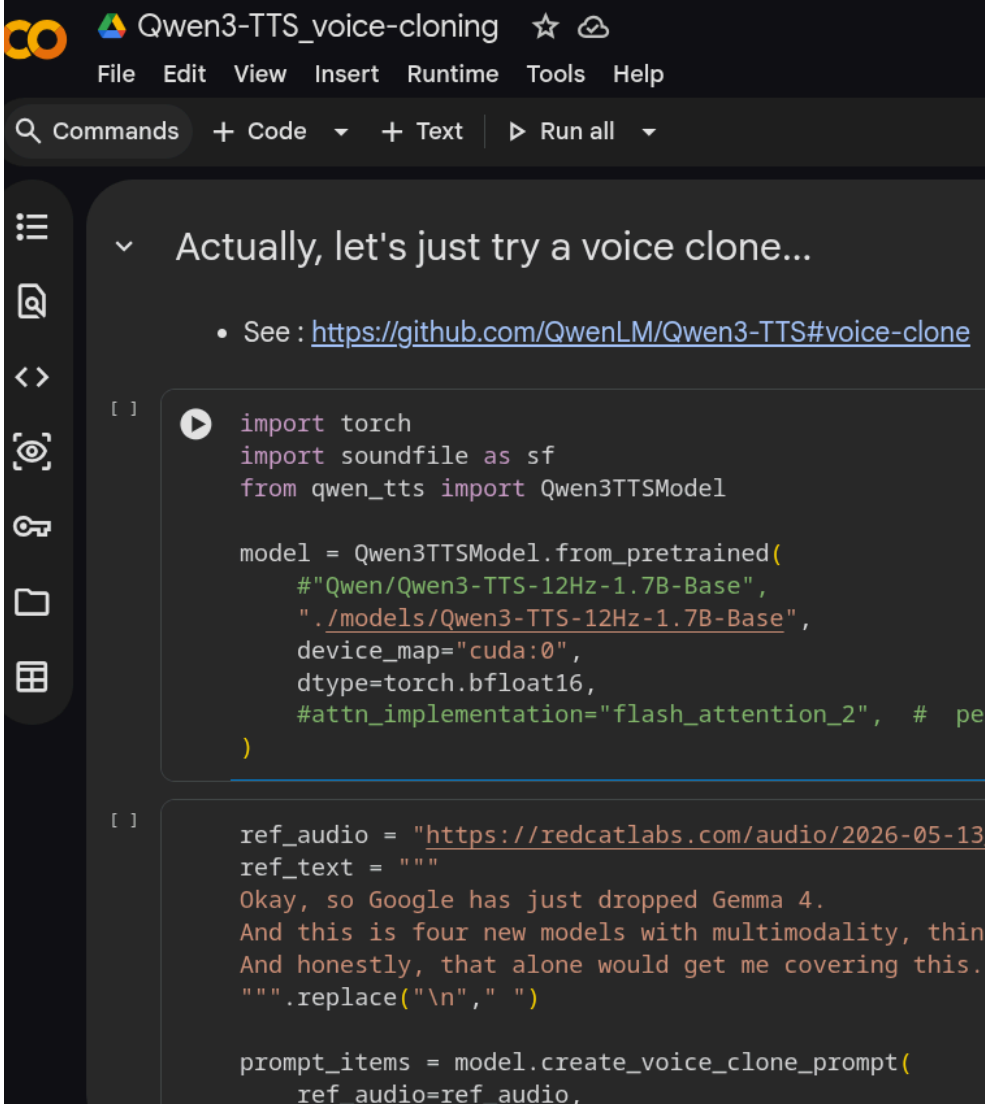
Abstract

We show that prompt following abilities of text-to-image models can be substantially improved by training on highly descriptive generated image captions. Existing text-to-image models struggle to follow detailed image descriptions and

MLSG
MACHINE LEARNING
SINGAPORE

Qwen Colab DEMO

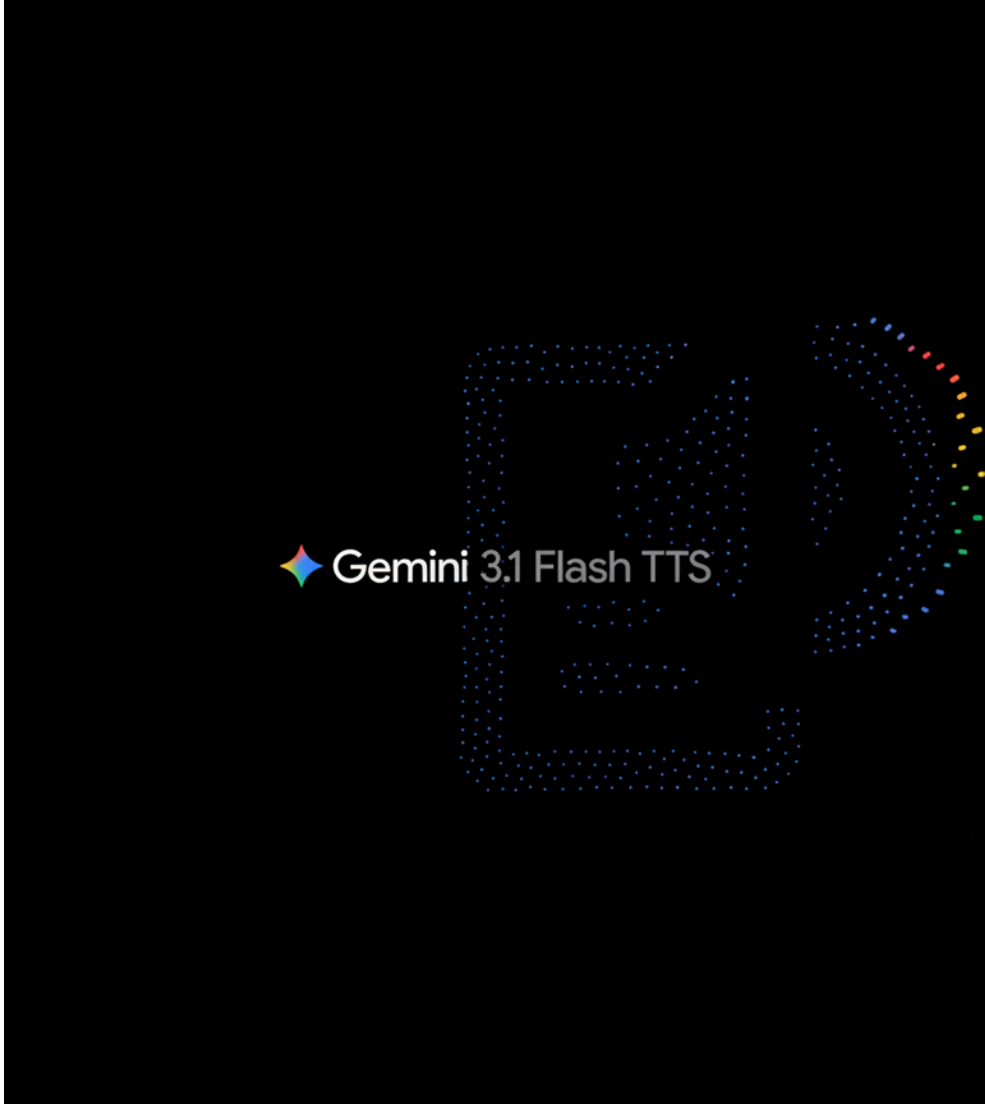
- Actual Voice-Cloning Colab
- DEMO
  - Overall set-up
    - GUI section skipped
  - In PyTorch:
    - ingestion of voice for cloning
    - generation of cloned voice



MLSG
MACHINE LEARNING
SINGAPORE

Wrap-Up

- TTS models are getting better very quickly
  - realism is now table-stakes
  - controllability is improving
- Other factors : Latency, streaming, ...
  - the right combo still difficult to find
- Lots of potential applications just opened up!



MLSG
MACHINE LEARNING
SINGAPORE