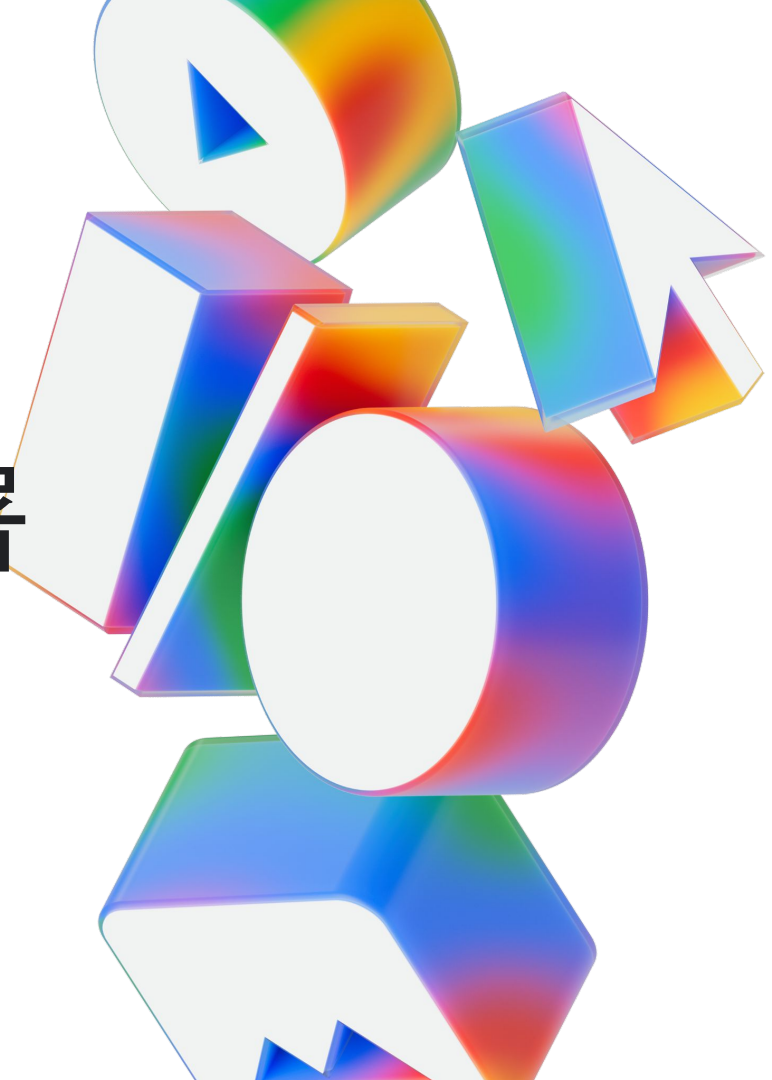
Google I/O

Gemma3端侧部署

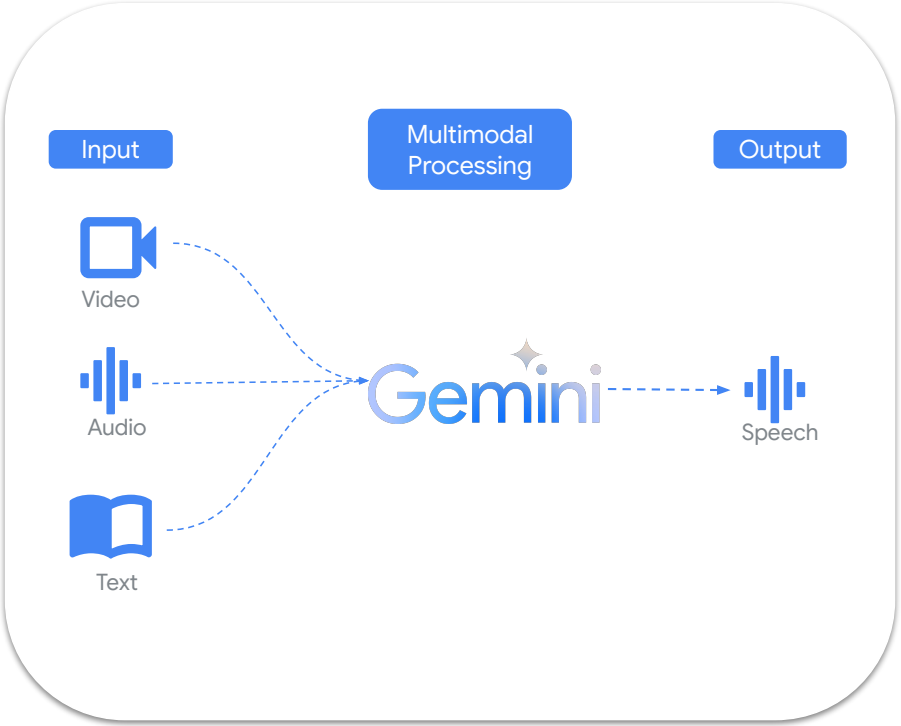


陈榆 GDE

Gemini

Multimodality in AI Studio

- Gemini in AI Studio can analyze kitchens from text descriptions, floor plans, and images, then suggest cohesive designs, color palettes, and materials using its native image generation capabilities.
- Understanding videos, native image generation, and grounding real information with Google Search are unique to Gemini.



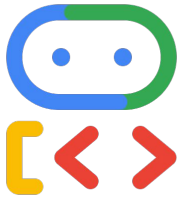
Source: [Google I/O '25 Developer Keynote]

Gemini

Agents: ADK, A2A & Agent Garden

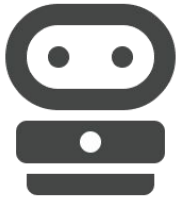
ADK Open Source

Code agents like Google (now production-ready for Python developers)



A2A GA

A2A now lets you create seamless and secure agents



Agent Garden Open Source

Sample agents and tools get you started quickly

Source: Google I/O '25 Developer Keynote

Gemini

Veo 3: Multimodal output

Vey 3 now delivers higher visual quality, a stronger understanding of physics, and better prompt adherence. With Veo 3, you can generate videos with:

- Improved quality when generating videos from text and image prompts
- Speech, such as dialogue and voice-overs
- Audio, such as music and sound effects

Source: Google I/O ‘25 Developer Keynote

Mobile

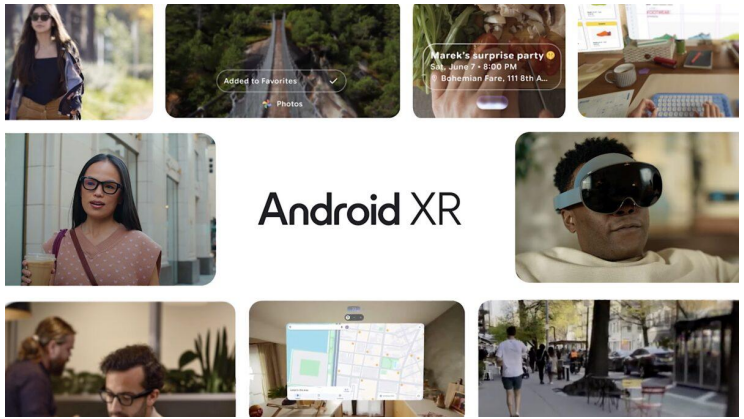
Android

Enhancing Development and User Experiences

Mobile

Android XR with Gemini integration

- A sneak peek was provided on how Gemini will work on glasses with Android XR for real-world scenarios like messaging, appointments, directions, and photos.
- Samsung's Project Moohan is the first Android XR device arriving later this year.



Source: [Google I/O '25 Developer Keynote]

Mobile

Material 3 Design for Android

- This evolution of Material Design aims to deliver more vibrant, personalized, and fluid user interfaces, bringing a new level of emotional engagement to Android apps.
- For developers, this means new tools and capabilities to create visually striking and intuitive experiences that resonate with users through enhanced customization, natural animations, and satisfying haptic feedback.

Source: [Google I/O '25 Developer Keynote]

Mobile

GenAI APIs Powered by Gemini Nano

- New multimodal, built-in AI APIs will allow users to interact with Gemini using both audio and image input, enabling on-device AI processing directly in web applications.
- These new APIs are powered by Gemini Nano, signifying a push towards more powerful and versatile on-device AI capabilities for web developers.

Source: Google I/O '25 Developer Keynote

AI

Gemma3模型

Gemma 是谷歌最 轻量级，最新的，开放 大模型，采用和Gemini 模型相同的模型架构和技术进行构建的。

Gemma 开放
遵守负责任使用AI规范
的大模型

Gemma 2 附带一个宽松的开源 许可证，允许重新分发、微调、商业使用和开 发相同类型的模型。

Gemma 2 & 1

CodeGemma

DataGemma

Recurrent Gemma

PaliGemma

ShieldGemma

Gemma 模型特点

对设计负责

这些模型结合了全面的安全措施，通过精心策划的数据集和严格的调整，有助于确保负责任且值得信赖的人工智能解决方案。

多尺寸高性能

Gemma 的 2B、7B、9B 和 27B 均取得了出色的基准测试结果，甚至优于一些较大的开源模型。

支持多框架

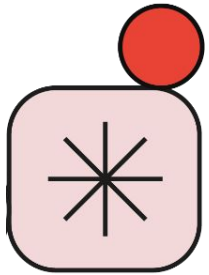
借助 Keras 3.0，您可以享受与 JAX、TensorFlow 和 PyTorch 的无缝兼容性，使您能够根据您的任务轻松选择和切换框架。

Gemma 1 and 2 区别

Feature	Gemma	Gemma 2
Model Sizes	2B, 7B	2B, 9B, 27B
Architecture Base	Transformer decoder	Transformer decoder
Attention Mechanism	Multi-Head Attention (7B), Multi-Query Attention (2B)	Grouped-Query Attention (GQA)
Position Encoding	Rotary Positioning Embeddings (RoPE)	Rotary Positioning Embeddings (RoPE)
Activation Function	GeGLU activation function	Approximated GeGLU activation function
Normalization	RMSNorm	RMSNorm (with additional pre and post-feedforward layers)
Logit Treatment	Standard	Logit Soft-Capping
Attention Pattern	Global attention	Alternating Local and Global Attention
Key Innovation	-	Alternating attention, logit soft-capping, additional normalization
Training Approach	Trained from scratch	2B and 9B distilled from 27B model
Training Data	Up to 6T tokens	Up to 13T tokens (for 27B model), 8T (9B), 2T (2B)
Vocabulary Size	256,128	256,128
Context Length	8,192 tokens	8,192 tokens
Instruction Tuning	SFT and RLHF	Extended SFT and RLHF, with model merging

模型架构

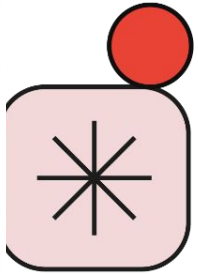
多模态:
采用SigLIP视觉编码器，将图像转换为token序列，使LLM能够处理图像信息。
通过Pan & Scan方法，支持处理任意分辨率的图像。
长文本处理:
增加上下文窗口大小到128K tokens（1B模型为32K）。
采用局部/全局注意力混合结构，降低KV缓存的内存占用。
多语言支持:
使用与Gemini 2.0相同的tokenizer，更好地支持非英语语言。
增加多语言训练数据，并采用Unimax策略处理语言不平衡问题。
知识蒸馏:使用知识蒸馏技术，将大型教师模型的知识迁移到小型学生模型中，提升模型性能。
后训练:采用一种新颖的后训练方法，提升模型在数学、推理、聊天、指令跟随和多语言等方面的能力。
采用监督微调（SFT）和强化学习人类反馈（RLHF）等技术，使模型更好地遵循指令。
使用权重平均奖励模型（WARM）等方法，提升模型的helpful, instruction-following, and multilingual abilities。
量化感知训练:
对模型进行量化，以减少内存占用和计算成本。
采用Quantization Aware Training (QAT)方法，在训练过程中模拟量化，以减少量化带来的性能损失



Gemma 各种变体型号

Parameter size	Input	Output	Architecture	Variants	Intended platforms
2B	Text	Text	Gemma 2	• Gemma 2 (base)	Mobile devices and laptops
			Gemma 1	• Gemma (base) • CodeGemma • RecurrentGemma	
3B	Text, Images	Text	Gemma 1	• PaliGemma	Mobile devices and laptops
7B	Text	Text	Gemma 1	• Gemma (base) • CodeGemma	Desktop computers and small servers
9B	Text	Text	Gemma 2	• Gemma 2 (base)	Higher-end desktop computers and servers
			Gemma 1	• RecurrentGemma	
27B	Text	Text	Gemma 2	• Gemma 2	

參數	Full 32bit	BF16 (16 位元)	SFP8 (8 位元)	Q4_0 (4 位元)	INT4 (4 位元)
Gemma 3 1B (僅限文字)	4 GB	1.5 GB	1.1 GB	892 MB	861 MB
Gemma 3 4B	16 GB	6.4 GB	4.4 GB	3.4 GB	3.2 GB
Gemma 3 12B	48 GB	20 GB	12.2 GB	8.7 GB	8.2 GB
Gemma 3 27B	108 GB	46.4 GB	29.1 GB	21 GB	19.9 GB



Ref: Gemma Docs - ai.google.dev

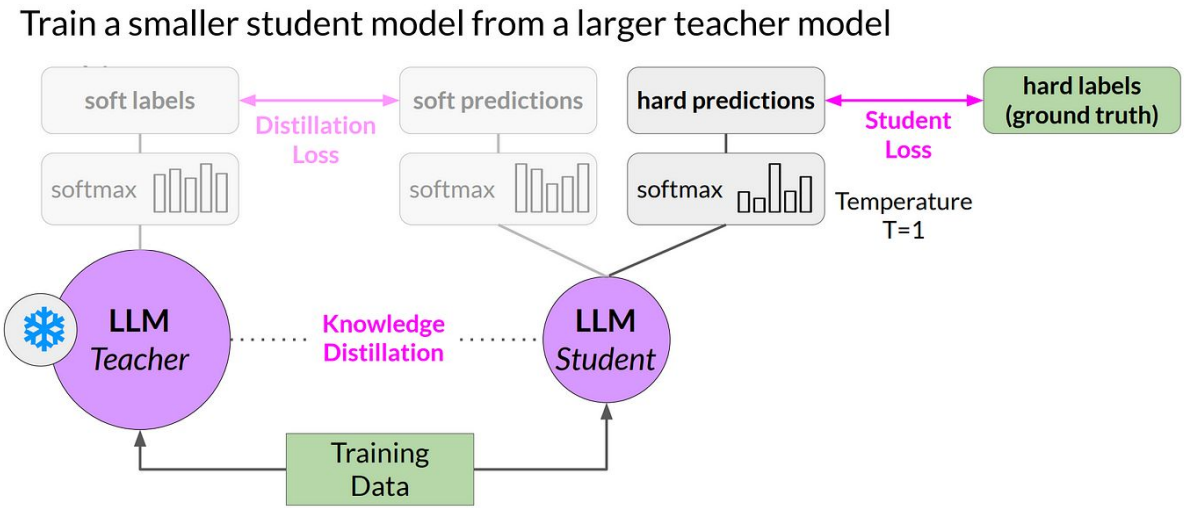
支持的软件平台，框架，硬件

Gemma3特点

- **多模态能力**：Gemma 3引入了视觉理解能力，可以处理图像和文本信息，这为LLM的应用开辟了新的方向。
- **长文本处理能力**：Gemma 3支持128K tokens的上下文长度，这使得模型可以处理更长的文档和对话，提升了模型的应用范围。
- **多语言支持**：Gemma 3增强了多语言支持能力，可以更好地服务于全球用户。
- **轻量化设计**：Gemma 3在保持高性能的同时，注重模型大小和计算效率，使其可以在消费级硬件上运行。
- **开源开放**：Gemma 3以开源的方式发布，促进了AI技术的普及和发展。

知识蒸馏

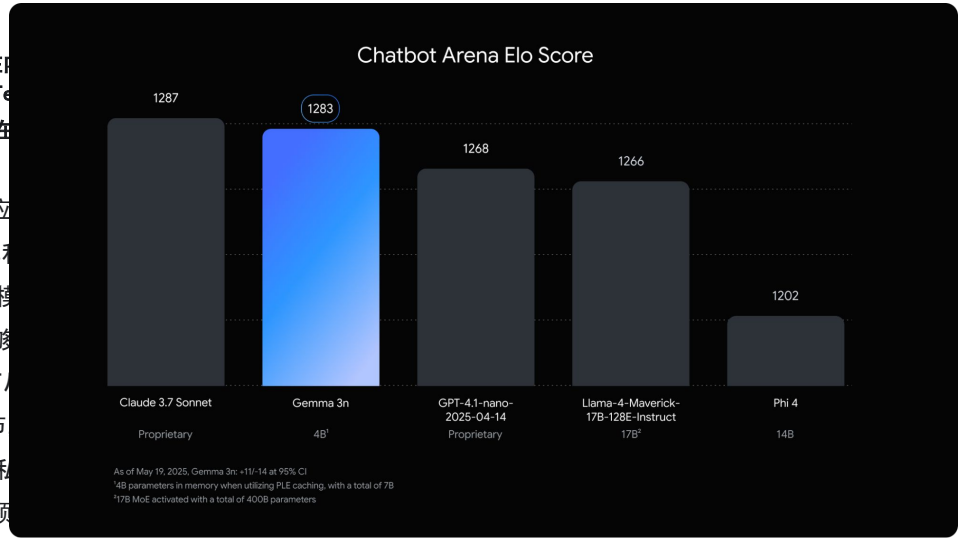
知识蒸馏是一种常用技术，用于训练较小的 学生 模型以模仿较大但表现更好的 教师 模型的行为。这是通过将大语言模型的下一个 Token 预测任务与教师提供的 Token 概率分布 (例如 GPT-4、Claude 或 Gemini) 结合起来，从而为学生提供更丰富的学 习信号。



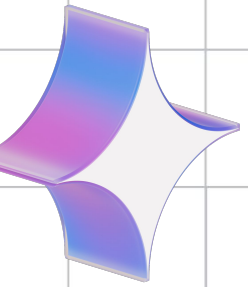
Ref: Machine Learning Q and AI (Book)

Gemma 3n 模型

为了推动下一代设备端人工智能的发展，并支持包括提升 Gemini Nano 性能在内
一代基础架构是与高通技术公司（Qualcomm Technologies）、联发科（MediaTe
企业密切合作创建的，并针对闪电般快速的多模态人工智能进行了优化，能够在

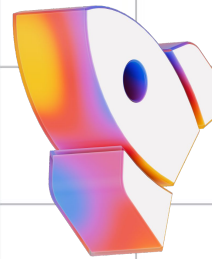


- **设备端性能优化与效率提升**：Gemma 3n 在移动端的启动响应
，并且通过诸如每层嵌入 (Per Layer Embeddings)、KVC 共享
- **多合一灵活性**：Gemma 3n 是一款主动内存占用为 4B参数的模
内存占用子模型（得益于 MatFormer 训练）。这一特性使其能够
们还在 Gemma 3n 中引入了混合匹配 (mix'n'match) 能力，可
例——并实现质量/延迟的关联权衡。关于这一研究的更多细节
- **隐私优先且离线可用**：本地执行支持以下特性：既尊重用户隐私
- **扩展的音频多模态理解能力**：Gemma 3n 能够理解和处理音频
持模型执行高质量的自动语音识别（转录）和翻译（语音转翻译文本）。此外，该模型接受跨模态的交错输入，从而能够理解
复杂的多模态交互（公共版本即将推出）。
- **增强的多语言能力**：Gemma 3n 提升了多语言性能，尤其在日语、德语、韩语、西班牙语和法语中表现突出。其在多语言基
准测试中展现出强劲实力，例如在 WMT24++ (ChrF) 中达到了 50.1% 的分数。



CodeLab

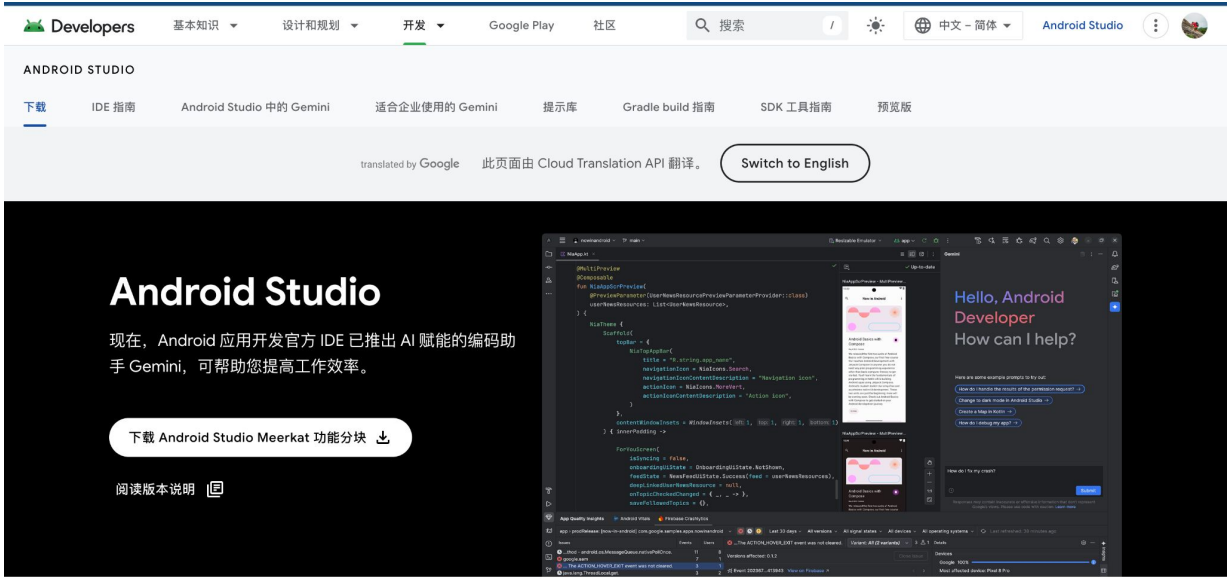
Gemma3 模型端侧部署



CodeLab

准备工作

- Android Studio
- 一台手机
- 数据线
- 科学上网



https://developer.android.com/studio?hl=zh-cn

CodeLab

部署环境

- Google AI Edge 端侧推理解决方案
- 下载模型
- 转换模型
- 导入模型到工程中
- API集成

https://ai.google.dev/edge?hl=zh-cn

CodeLab

部署环境

- Google AI Edge 端侧推理解决方案
- 下载模型
- 转换模型
- 导入模型到工程中
- API集成

```python
from mediapipe.tasks.python.genai.bundler import llm_bundler

def build_gemma3_1b_it_block_q4():
  output_file = "/content/gemma3_1b_finetune_q4_block32_ekv1024.task"
  tflite_model = "/content/gemma3_1b_finetune_q4_block32_ekv1024.tflite"
  tokenizer_model = (
      "/content/tokenizer.model"
  )
  config = llm_bundler.BundleConfig(
      tflite_model=tflite_model,
      tokenizer_model=tokenizer_model,
      start_token="",
      stop_tokens=[""],
      output_filename=output_file,
      enable_bytes_to_unicode_mapping=False,
      prompt_prefix="user\n",
      prompt_suffix="\nmodel\n",
  )
  llm_bundler.create_bundle(config)

# Build the MediaPipe task bundle.
build_gemma3_1b_it_block_q4()
```

CodeLab

部署环境

- Google AI Edge 端侧推理解决方案
- 下载模型
- 转换模型
- 导入模型到工程中
- API集成

PyTorch 模型转换

★ **注意：** AI Edge Torch 库目前处于开发初期阶段。该 API 不稳定，并且存在一些已知问题。

您可以使用 AI Edge Torch Generative API 将 PyTorch 生成式模型转换为与 MediaPipe 兼容的格式。您可以使用此 API 将 PyTorch 模型转换为多签名 LiteRT (TensorFlow Lite) 模型。如需详细了解如何映射和导出模型，请访问 AI Edge Torch 的 GitHub 页面。

使用 AI Edge Torch Generative API 转换 PyTorch 模型涉及以下步骤：

1. 下载 PyTorch 模型检查点。
2. 使用 AI Edge Torch Generative API 编写、转换模型，并将其量化为与 MediaPipe 兼容的文件格式（`.tflite`）。
3. 使用 tflite 文件和模型分词器创建任务软件包（`.task`）。

Torch 生成式转换器仅适用于 CPU，并且需要 Linux 机器具有至少 64 GB 的 RAM。

https://github.com/google-ai-edge/ai-edge-torch

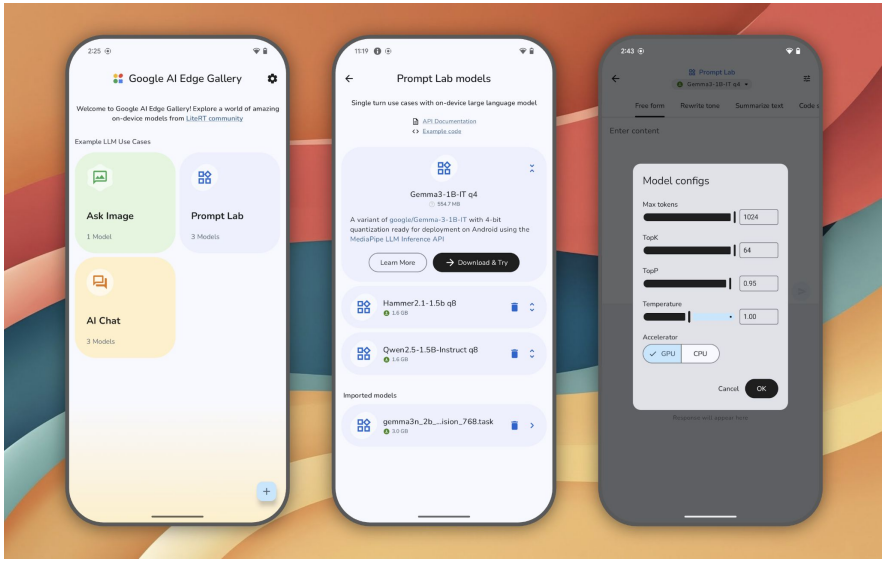
CodeLab

工程文件

- 下载文件
- 使用Android Studio打开

（以/Users/用户名/AndroidProject/gallery-main/Android/src）为项目路径

- 编译工程文件（可能需要科学上网）
- 下载到真机



https://github.com/google-ai-edge/gallery

CodeLab

下载模型

- gemma3-1B-IT
- gemma3-3n-E4B

Gemma3-1B-IT_multi-prefill-seq_q4_ekv4096.liter…	584 MB LFS	Rename Gemma3-1B-IT_multi-prefill-seq_int4_ekv…	9 days ago
Gemma3-1B-IT_multi-prefill-seq_q8_ekv1280.task	1.05 GB LFS	Add files using upload-large-folder tool	about 1 month ago
Gemma3-1B-IT_multi-prefill-seq_q8_ekv2048.task	1.07 GB LFS	Upload Gemma3-1B-IT_multi-prefill-seq_q8_ekv20…	about 1 month ago
Gemma3-1B-IT_multi-prefill-seq_q8_ekv4096.task	1.05 GB LFS	Add files using upload-large-folder tool	about 1 month ago
Gemma3-1B-IT_seq128_f32_ekv1280.task	4.01 GB LFS	Add files using upload-large-folder tool	about 1 month ago
Gemma3-1B-IT_seq128_f32_ekv4096.task	4.01 GB LFS	Add files using upload-large-folder tool	about 1 month ago
Gemma3-1B-IT_seq128_q4_block128_ekv1280.task	676 MB LFS	Add files using upload-large-folder tool	about 1 month ago
Gemma3-1B-IT_seq128_q4_block128_ekv4096.task	676 MB LFS	Add files using upload-large-folder tool	about 1 month ago
Gemma3-1B-IT_seq128_q4_block32_ekv1280.task	709 MB LFS	Add files using upload-large-folder tool	about 1 month ago
Gemma3-1B-IT_seq128_q4_block32_ekv4096.task	709 MB LFS	Add files using upload-large-folder tool	about 1 month ago
Gemma3-1B-IT_seq128_q8_ekv1280.task	1.02 GB LFS	Add files using upload-large-folder tool	about 1 month ago
Gemma3-1B-IT_seq128_q8_ekv4096.task	1.02 GB LFS	Add files using upload-large-folder tool	about 1 month ago
README.md	17.5 kB	Update README.md	10 days ago
gemma3-1b-it-int4-web.task	700 MB LFS	Upload 2 files	about 1 month ago
gemma3-1b-it-int4.litertlm	555 MB LFS	Upload gemma3-1b-it-int4.litertlm	10 days ago
gemma3-1b-it-int4.task Safe	555 MB LFS	Rename gemma3-1B-it-int4.task to gemma3-1b-it-…	3 months ago
gemma3-1b-it-int8-web.task	1.01 GB LFS	Upload 2 files	about 1 month ago
notebook.ipynb	2.35 kB	Update notebook.ipynb	9 days ago
tokenizer.model Safe	4.69 MB LFS	Add files using upload-large-folder tool	about 2 months ago

https://huggingface.co/litert-community/Gemma3-1B-IT

https://huggingface.co/google/gemma-3n-E4B-it-litert-preview/tree/main

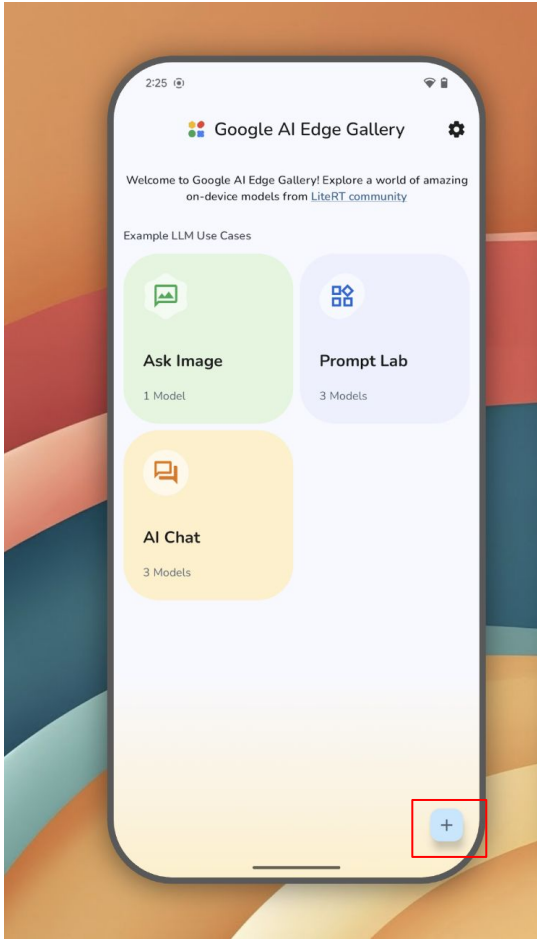
CodeLab

推送模型到真机

```
$ adb shell rm -r /data/local/tmp/llm/ loaded models
$ adb shell mkdir -p /data/local/tmp/llm/
$ adb push gemma3.task /data/local/tmp/llm/gemma3.task
```

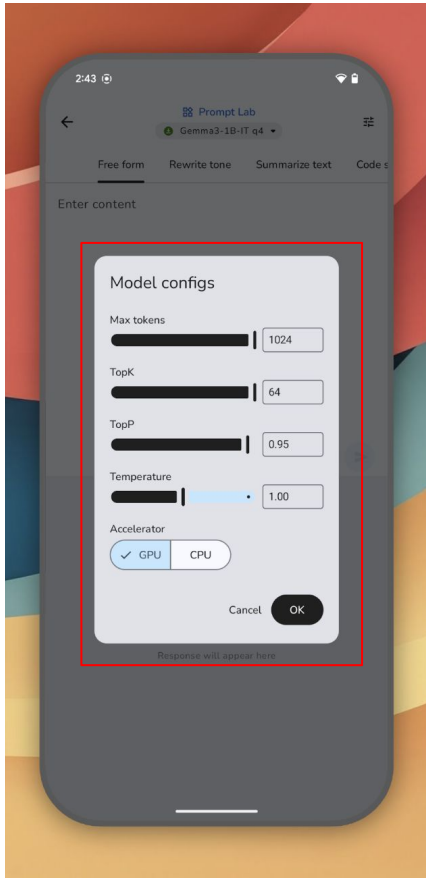
CodeLab

APP上导入模型



CodeLab

选择合适的推理硬件



CodeLab

选做（微调Gemma并转为.task模型）

https://github.com/google-ai-edge/mediapipe-samples/blob/main/codelabs/litert_inference/Gemma3_1b_fine_tune.ipynb

CodeLab

参考资料

1.https://ai.google.dev/edge/mediapipe/solutions/genai/llm_inference/android?hl=zh-cn
2.https://github.com/google-ai-edge
3.https://developers.googleblog.com/en/google-ai-edge-small-language-models-multimodality-rag-function-calling/