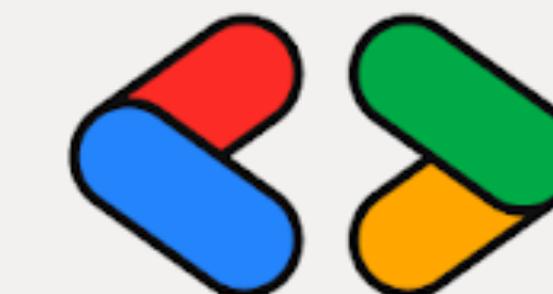




# Agents on Mobile 2025

On-device Model 与多端协同

El Zhang (2BAB)



Google  
Developer  
Groups



The diagram illustrates a network of connections. An arrow points from the DevFest Shenzhen logo to a globe icon, symbolizing global reach. Another arrow points from the globe to a smartphone screen displaying a Google Developer Experts certificate for Bingquan Zhang. A third arrow points from the smartphone to a book titled "EXTENDING ANDROID BUILDS" by El(Bingquan) Zhang. Below the book is a QR code. To the right of the book is the BinaryTape logo, which features a stylized "B" composed of binary digits (0s and 1s). The BinaryTape logo is also enclosed in a rounded rectangle along with the book and QR code.

- Google Developer Expert (Android)
- Tech Lead at an AI Startup
- Founder of BinaryTape
- Author of Extending Android Builds - a book officially endorsed by Gradle
- Google 2025 出海创业加速器导师 / 出海日讲师

01

# 端侧小模型 Recap

年度趋势

“互动环节”

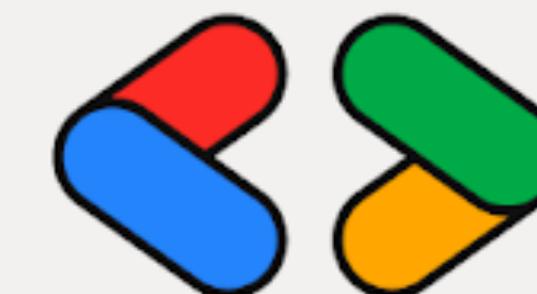


Google  
Developer  
Groups



# Gemma 2 / 3(n)

模型系列	模型参数大小	是否多模态	Max Context Length on Mobile	发布日期
Gemma 2	2B, 9B, 27B	否	~2048	24 年 6 月
Gemma 3	1B, 4B, 12B, 27B	是 (4B 以上支持图片)	2048~4096	25 年 3 月
Gemma 3n	E2B, E4B	是 (图片、音频)	4096	25 年 6 月

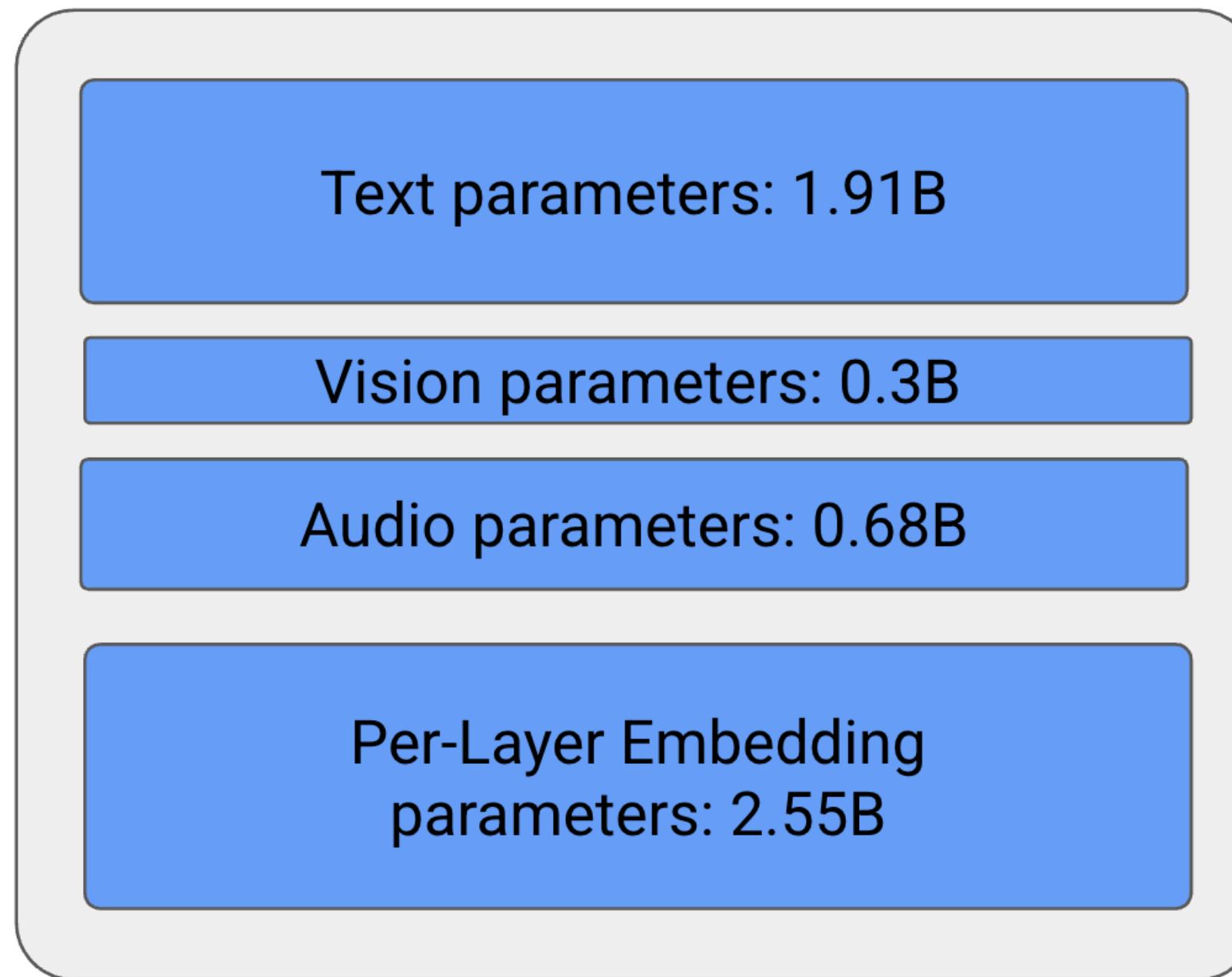


Google  
Developer  
Groups

# Gemma 3n

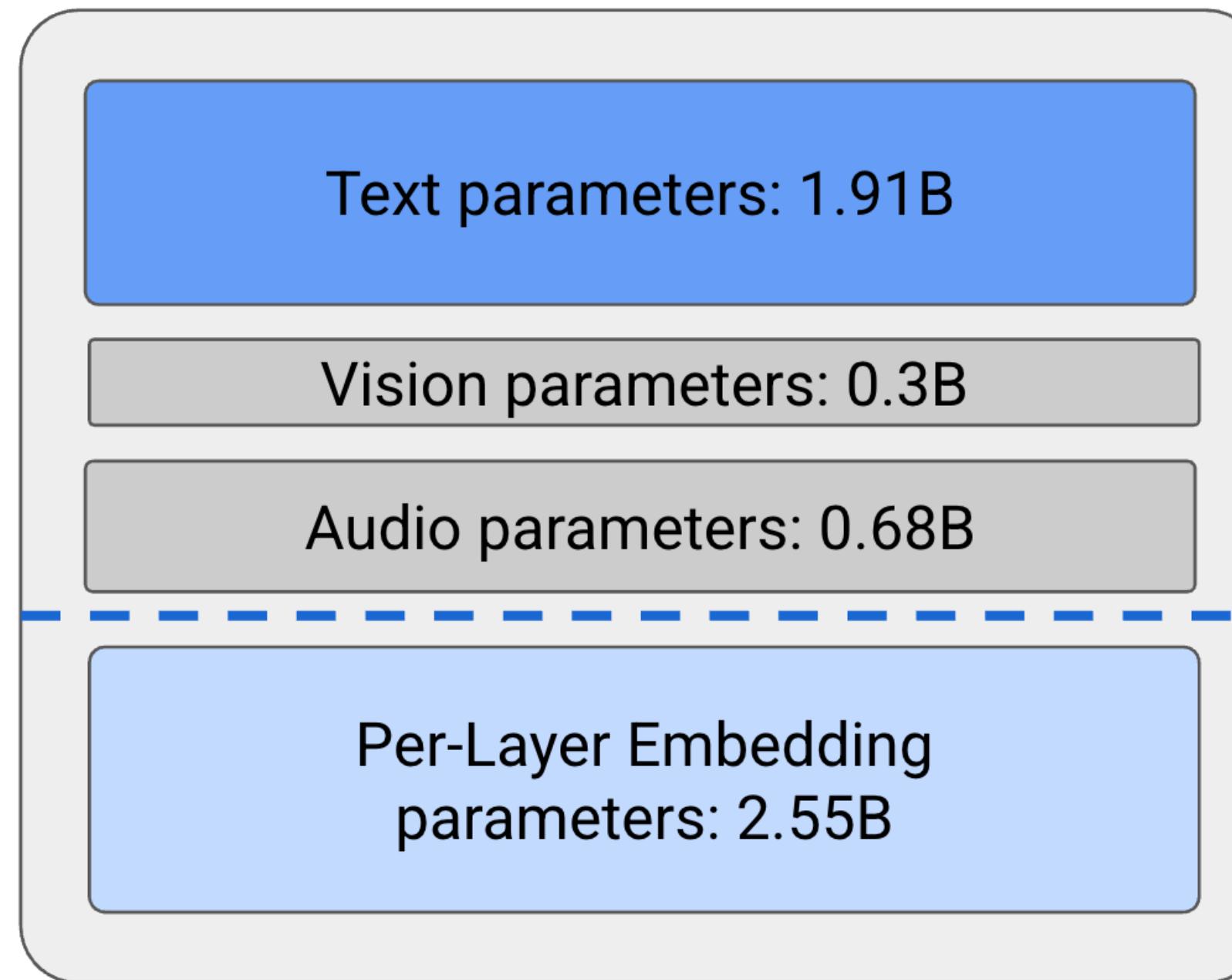
## Standard execution

Parameters loaded: 5.44B



## with skipped parameters & cached PLE

Parameters loaded: 1.91B



PLE data cached to **fast storage**



Google  
Developer  
Groups

# Gemma 3n

- MatFormer Arch
- 前馈神经网络（FFN）从单一结构变成嵌套类型（类似套娃）
- 可以根据设备的硬件情况在模型加载时决定使用不同复杂度的网络



Google  
Developer  
Groups

# Gemma 3n

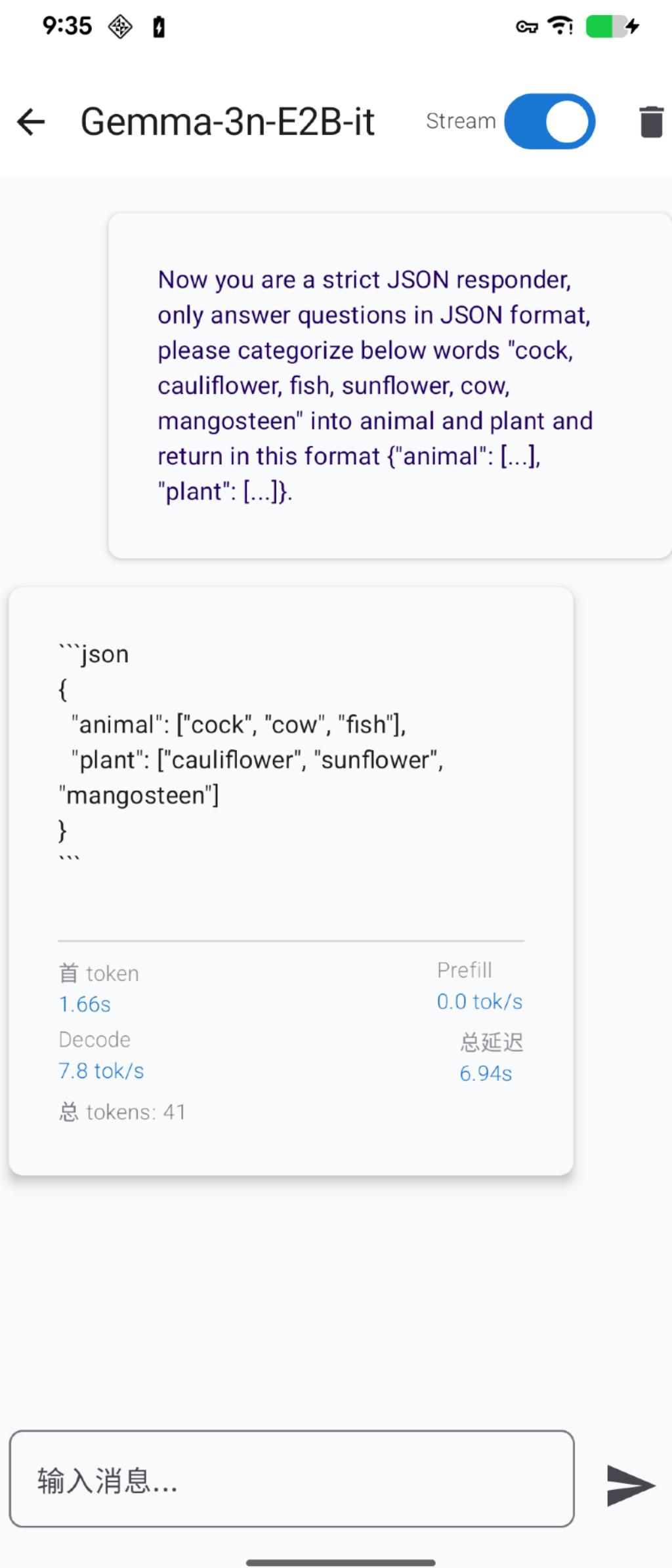
全面提升：

- 推理能力

- 知识量

- 固定输出格式

- 内存占用



Two screenshots of the "MediaPipe LLM Inference" app are shown side-by-side, comparing the GEMMA3 CPU and GEMMA3 CPU models.

**Left Screenshot (GEMMA3\_CPU):**

- Timestamp: 10:50
- Model: MediaPipe LLM Inference
- Responses generated by user-provided model
- Processor: GEMMA3\_CPU
- Tokens remaining: 1141 tokens remaining
- User Input: "Now you are a strict JSON responder, only answer questions in JSON format, please categorize below words "cock, cauliflower, fish, sunflower, cow, mangosteen" into animal and plant and return in this format {"animal": [...], "plant": [...]}.  
```json  
{  
 "animal": ["cock", "cow", "fish"],  
 "plant": ["cauliflower", "sunflower",  
 "mangosteen"]  
}  
```"
- Model Output: {"animal": ["cock", "cow", "fish", "dinosaur", "beetles", "dinosaur"], "plant": ["cock", "cauliflower"]}

**Right Screenshot (GEMMA3\_CPU):**

- Timestamp: 10:51
- Model: MediaPipe LLM Inference
- Responses generated by user-provided model
- Processor: GEMMA3\_CPU
- Tokens remaining: 745 tokens remaining
- User Input: "Now categorize the following words:  
["cock", "cauliflower", "fish", "sunflower", "cow", "mangosteen"]
- Model Output: {"animal": ["lion", "ant"], "plant": ["apple", "tulip"]}



Google  
Developer  
Groups

# Gemini Nano v2 / v3 with MLKit

| 模型系列           | 是否多模态     | 发布日期        |
|----------------|-----------|-------------|
| Gemini Nano v2 | 是 (图片)    | 2024 年 10 月 |
| Gemini Nano v3 | 是 (图片、音频) | 2025 年 8 月  |



Google  
Developer  
Groups

# Gemini Nano v2 / v3 with MLKit

## Feature-specific API device support

The [Summarization](#), [Proofreading](#), [Rewriting](#), and [Image Description](#) APIs are available on the following devices, with plans to expand support to additional devices:

- Google: Pixel 10, Pixel 10 Pro, Pixel 10 Pro XL, Pixel 10 Pro Fold, Pixel 9, Pixel 9 Pro, Pixel 9 Pro XL, Pixel 9 Pro Fold
- Honor: Honor 400 Pro, Magic 6 Pro, Magic 6 RSR, Magic 7, Magic 7 Pro, Magic V3, Magic V5
- iQOO: iQOO 13
- Motorola: Razr 60 Ultra
- OnePlus: OnePlus 13, OnePlus 13s, OnePlus Pad 3
- OPPO: Find N5, Find X8, Find X8 Pro, Reno 14 Pro
- POCO: POCO F7 Ultra, POCO X7 Pro
- realme: realme GT 7 Pro, realme GT 7T
- Samsung: Galaxy S25, Galaxy S25+, Galaxy S25 Ultra, Galaxy Z Fold7
- vivo: vivo X200, vivo X200 Pro, vivo X Fold3 Pro, vivo X Fold5
- Xiaomi: Xiaomi 15 Ultra, Xiaomi 15, Xiaomi 15T Pro, Xiaomi 15T, Xiaomi Pad mini



Google  
Developer  
Groups

# Gemini Nano v2 / v3 with MLKit

Prompt API device support [🔗](#)

[Prompt API](#) is currently supported on the following devices:

---

nano-v2

- Google: Pixel 9, Pixel 9 Pro, Pixel 9 Pro XL, Pixel 9 Pro Fold
- Honor: Magic V5, Magic 7, Magic 7 Pro
- iQOO: iQOO 13
- Motorola: Razr 60 Ultra
- OnePlus: OnePlus 13, OnePlus 13s, OnePlus Pad 3
- OPPO: Find N5
- POCO: POCO F7 Ultra
- realme: realme GT 7 Pro
- Samsung: Galaxy Z Fold7
- Xiaomi: Xiaomi 15 Ultra, Xiaomi 15

---

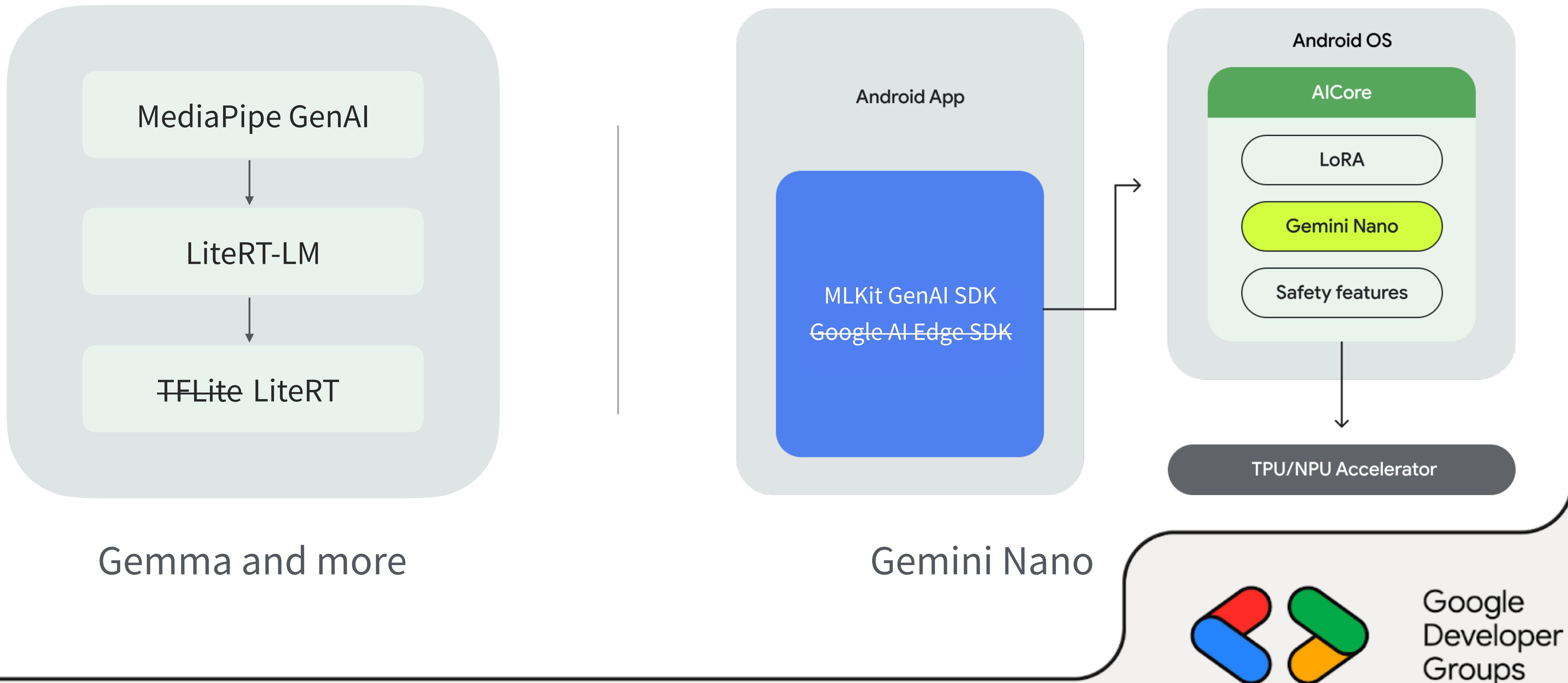
nano-v3

- Google: Pixel 10, Pixel 10 Pro, Pixel 10 Pro XL, Pixel 10 Pro Fold



Google  
Developer  
Groups

# Google 的 On Device AI 推理框架



# Google 的 On Device AI 演进路线



简介 评论 弹幕功能关闭 |

Google 中国  
31万粉丝 1413 视频

+ 关注

从技术生态到个人助理 | 一场对谈，看懂移动端 AI 的演进路线图 (上)

1078 0 2025年9月19日 17:00 1人正在看

BV12yWuz2Exw 未经作者授权禁止转载

在本期视频中，我们有幸邀请到 Google Core ML 团队的 Li Na，在 Google I/O Connect 上海现场与两位 Android GDE——2BAB 和 Yuang 展开深度对话。

我们深入探索了 Google 端侧 AI 的完整生态，从底层的 AICore 到高级的 ML Kit，再到为开源模型提供实验通路的 LiteRT (前 TensorFlow Lite)，旨在从技术层面为出海开发者提供支持和指导。访谈还分享了 Gemini Nano 的最新应用，NPU 带来的性能提升，以及端侧 LoRA 微调的巨大潜力。欢迎您观看视频，轻松掌握移动端 AI 的未来方向。



Google  
Developer  
Groups

# Gecko-110m-en

[litert-community/Gecko-110m-en](#)

This model provides a few variants of the embedding model published in the [Gecko paper](#) that are ready for deployment on Android or iOS using [LiteRT stack](#) or [google ai edge RAG SDK](#).

## Use the models

### Android

- Try out the gecko embedding model in the [google ai edge RAG SDK](#). You can find the SDK on [GitHub](#) or follow our [android guide](#) to install directly from Maven. We have also published a [sample app](#).
- Use the sentencepiece model as the tokenizer for the Gecko embedding model.

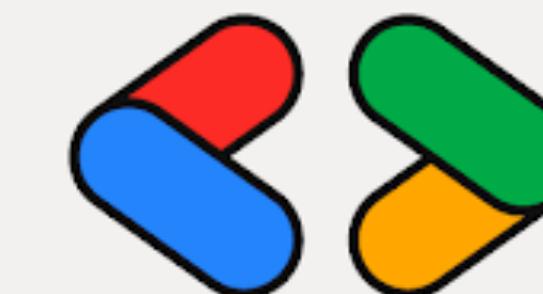


Google  
Developer  
Groups

02

# 多端协同 Agent

Gemma 3n 与 Gemini 2.5



Google  
Developer  
Groups

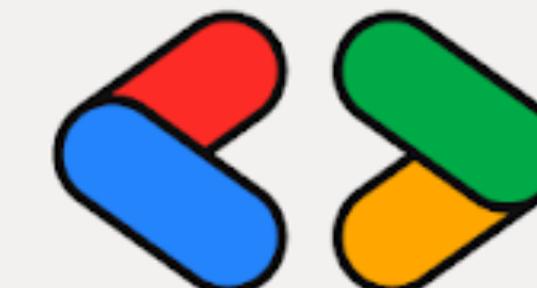
## 现有场景和长远目标

当前：

- 非常具体，单一任务
- Workflow 场景有限

长远：

- 复杂场景适合更高权限的系统级应用
- Agentic/AGI 考验模型大小和设备性能

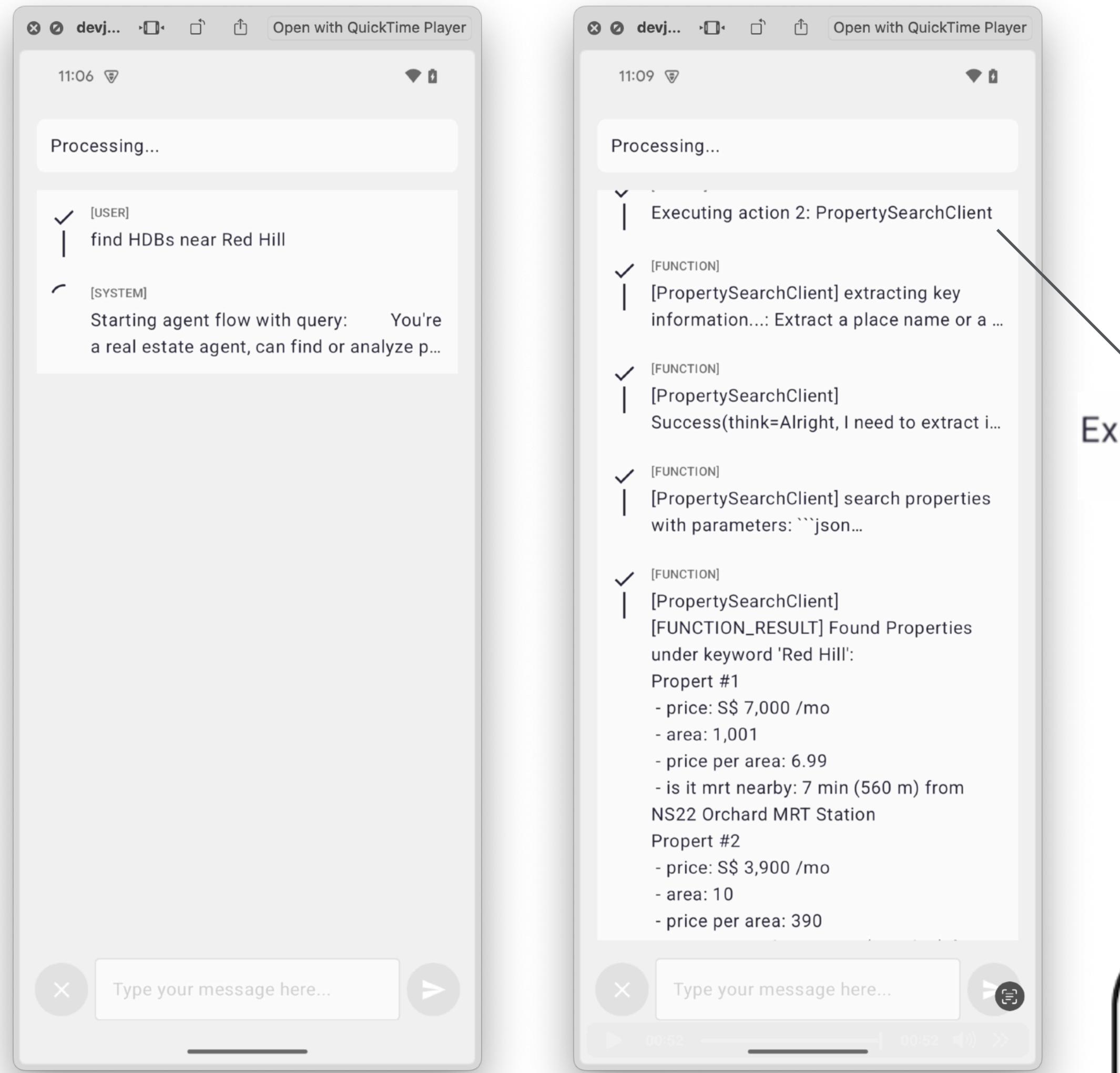


Google  
Developer  
Groups

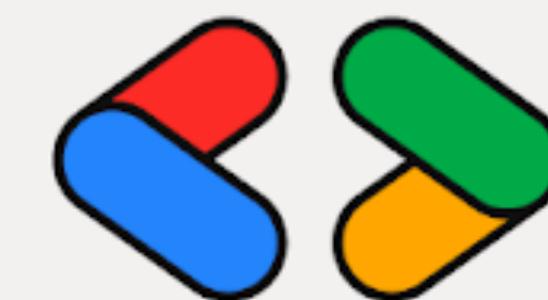
# 固定场景和 API

## Function Calling

- 远程 API



Executing action 2: PropertySearchClient

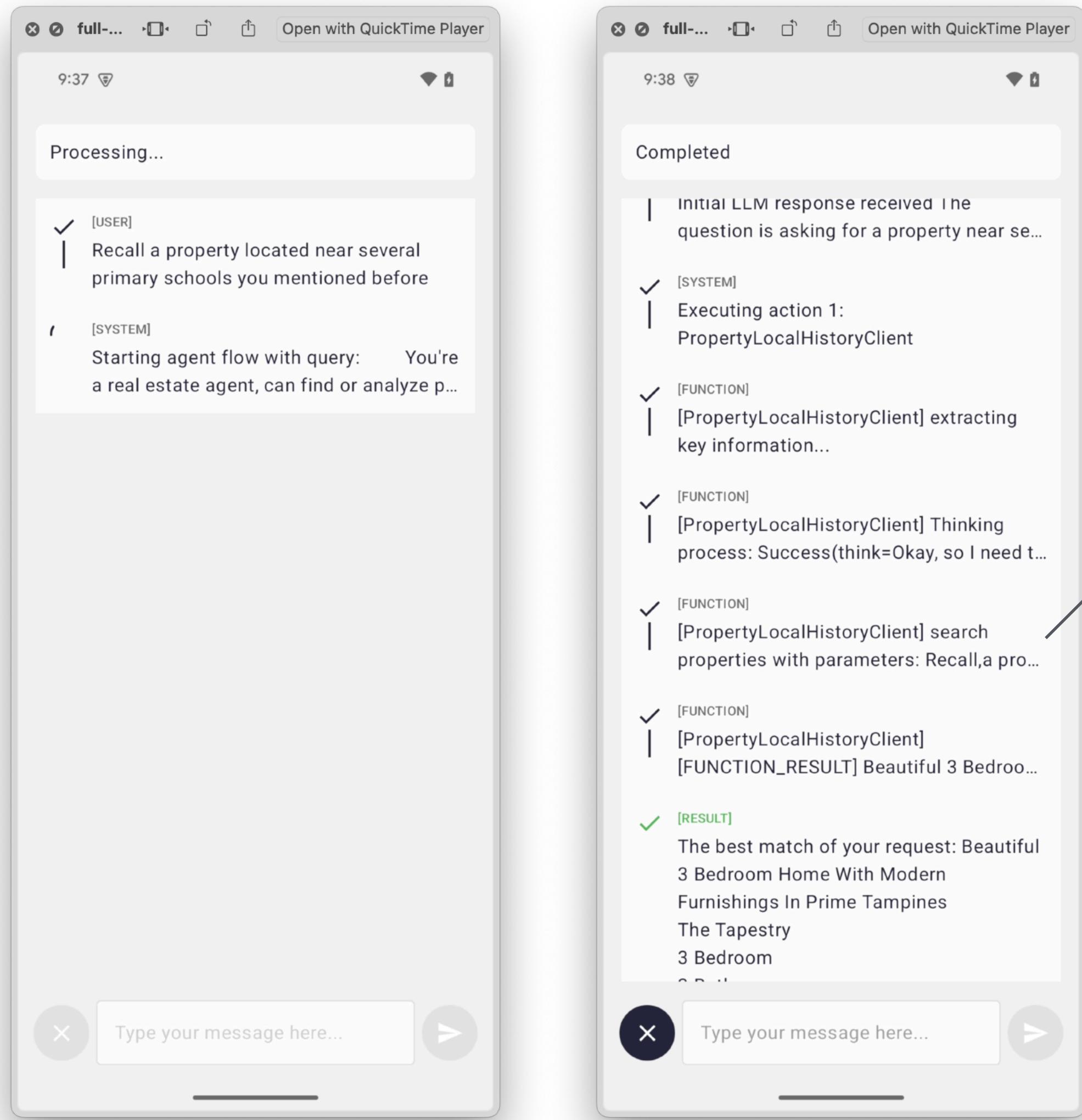


Google  
Developer  
Groups

# 固定场景和 API

RAG

- Gecko-110m-en



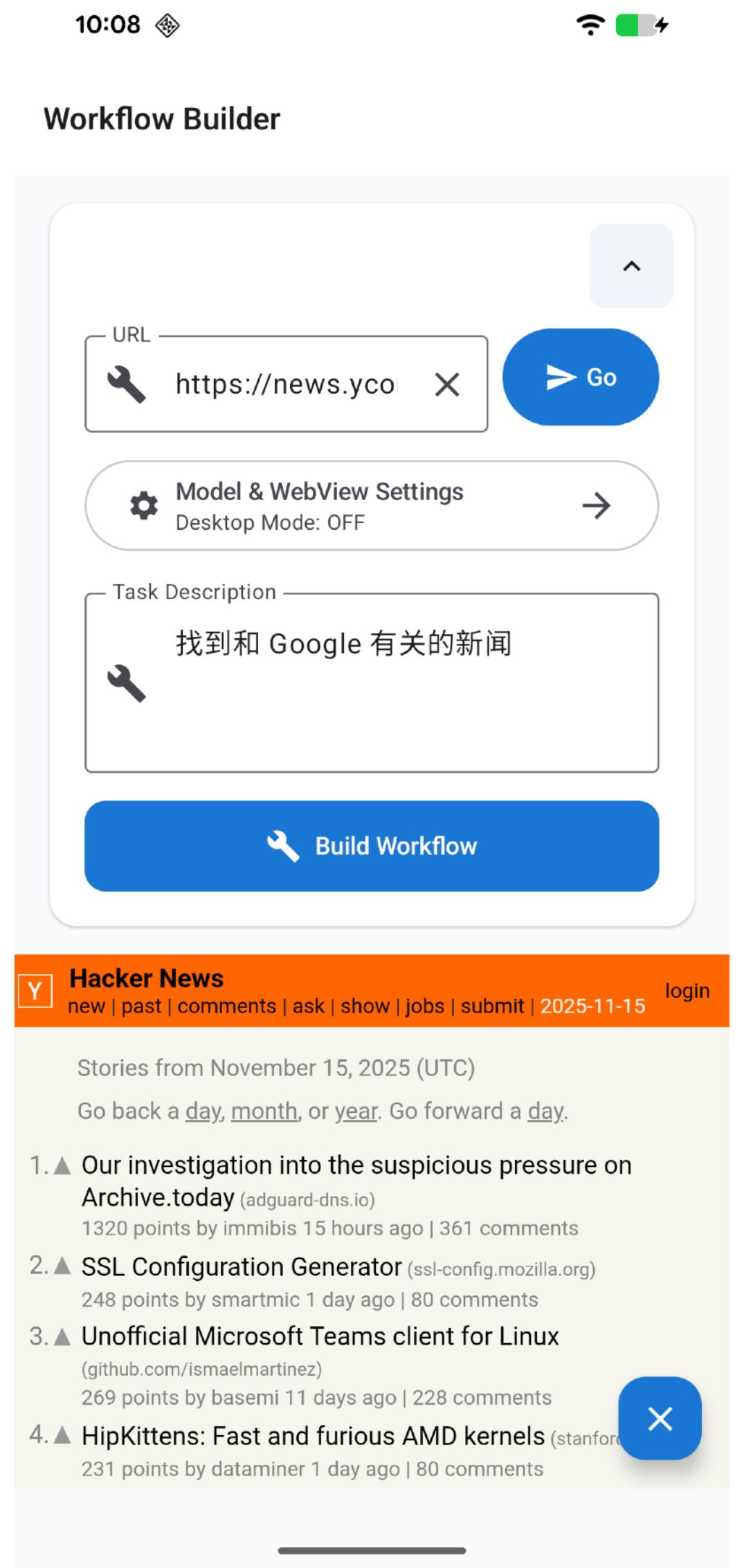
[PropertyLocalHistoryClient] search properties with parameters: Recall,a pro...



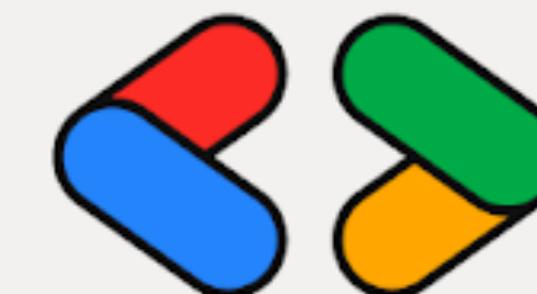
Google  
Developer  
Groups

# 通用的订阅管理器

- 不限制网站
- 不限制任务类型（查找、过滤、总结等等）
- (仅受限于模型 Context Length)



“互动环节”

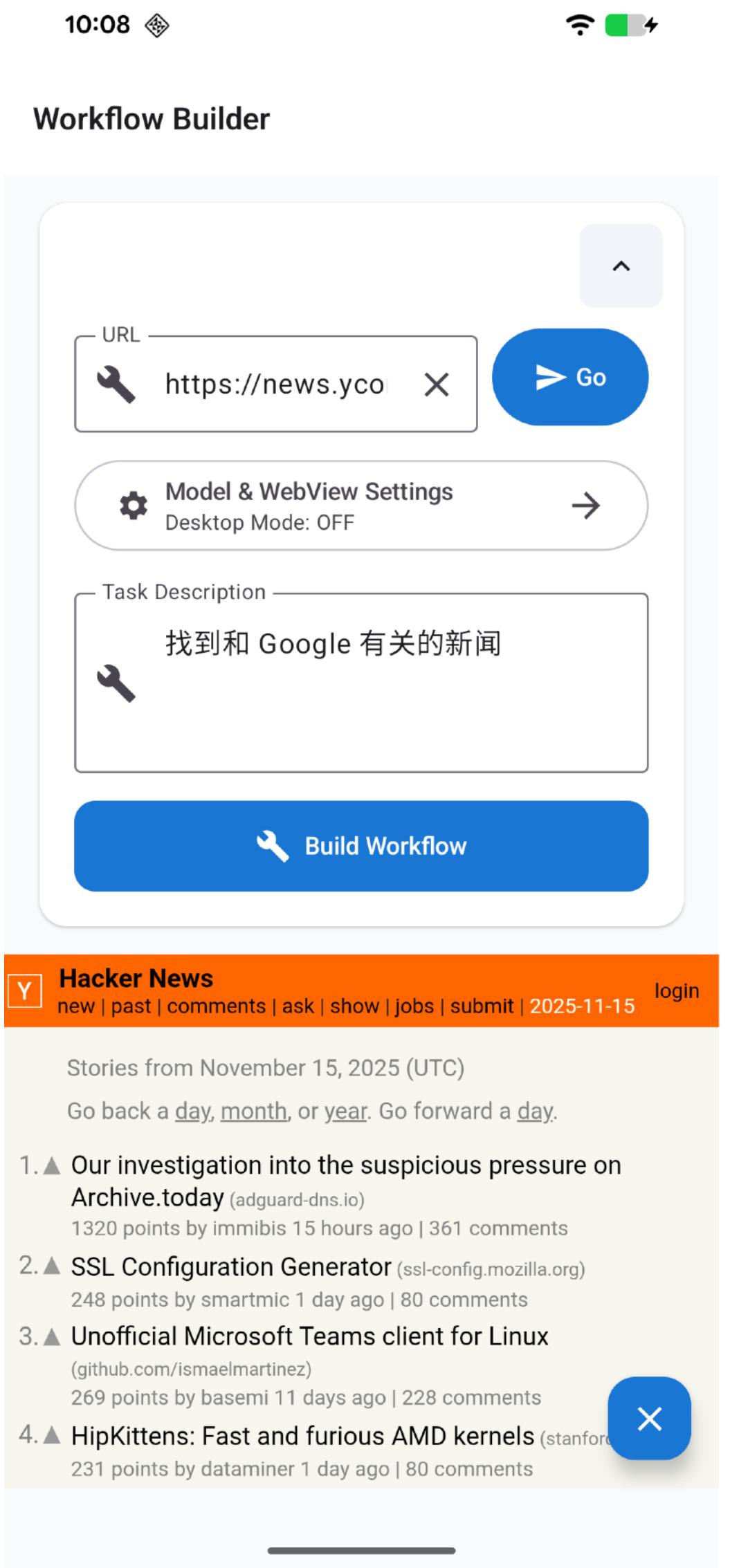


Google  
Developer  
Groups

# 通用的订阅管理器

# Agents on Mobile 2025

On-device Model 与多端协同



Google  
Developer  
Groups

## Phase 1: Build

工作流构建阶段

Gemini 2.5 Pro/Flash

### 输入

用户目标 + 初始URL

### 加载页面

WebView 加载并简化 HTML

### Propose

LLM 分析并提议下一步动作

### Verify

在 WebView 中执行验证

### Commit

成功后添加到工作流

循环最多 20 轮



## Workflow JSON

完整的工作流定义  
包含所有步骤、类型  
和变量映射



## Phase 2: Execute

工作流执行阶段

Gemma 3n E2/4B

### 加载工作流

读取步骤列表和配置

### 变量解析

处理 {{variable}} 占位符

### 执行步骤

JS / LLM / Loop / Navigate

### 错误重试

最多 3 次指数退避重试

### 存储结果

保存到变量上下文

→ 顺序执行所有步骤



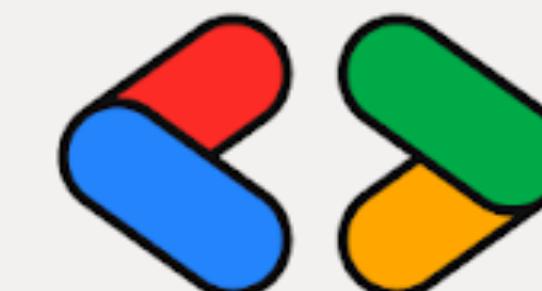
Google  
Developer  
Groups

Demo1

Demo2

## 今日还未涉及的内容

- 图片分析的加入
- 后台执行，每日早上推送



Google  
Developer  
Groups

```
{  
  "thought": "Detailed step-by-step reasoning of your current analysis and plan...",  
  "workflowStepsBuiltSoFar": [  
    {  
      "id": "stepId",  
      "description": "Human-readable description of what this step does",  
      "type": "js" | "llm" | "loop" | "navigate",  
      "payload": {  
        "code": "/* For js: JavaScript code with {{placeholder}} support */",  
        "query": "/* For llm: Concise task description*/",  
        "context": "/* For llm: {{previousResult}} providing the data */",  
        "url": "/* For navigate: URL to load, supports {{placeholders}} */",  
        "iterateOn": "/* For loop: {{arrayVariable}} to iterate over */",  
        "loopVariable": "currentItem",  
        "navigateToField": "/* For loop (optional): field name in item containing URL to navigate to */",  
        "steps": [/* For loop: nested workflow steps */]  
      },  
      "outputVariable": "resultName"  
    },  
    {  
      "nextAction": {  
        "toolName": "js" | "navigate" | "finish",  
        "parameters": {  
          "code": "/* For js tool: The JavaScript code to execute now */",  
          "url": "/* For navigate tool: Full URL to load */"  
        }  
      },  
      "isDone": false,  
      "initialUrl": "{{current_page_url}}"  
    }  
  ]  
}
```



# Thank You! / Q&A

El Zhang (2BAB)

Google Developer Expert - Android



Google  
Developer  
Groups