



包山包海的 Agents on Mobile

2025 LLM 雲地協作新玩法

El Zhang (2BAB)



Google
Developer
Groups



- Google Developer Expert (Android)
- Tech Lead at an AI Startup
- Author of Extending Android Builds - a book officially endorsed by Gradle
- Mentor of Google for Startups Accelerator 2025
- Founder of BinaryTape

The collage includes:

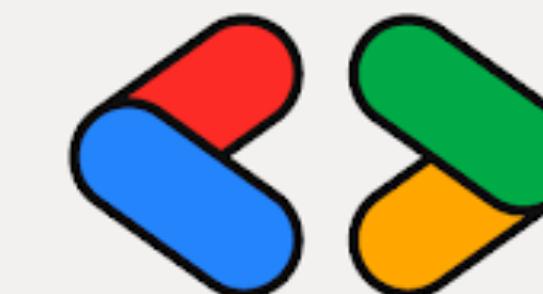
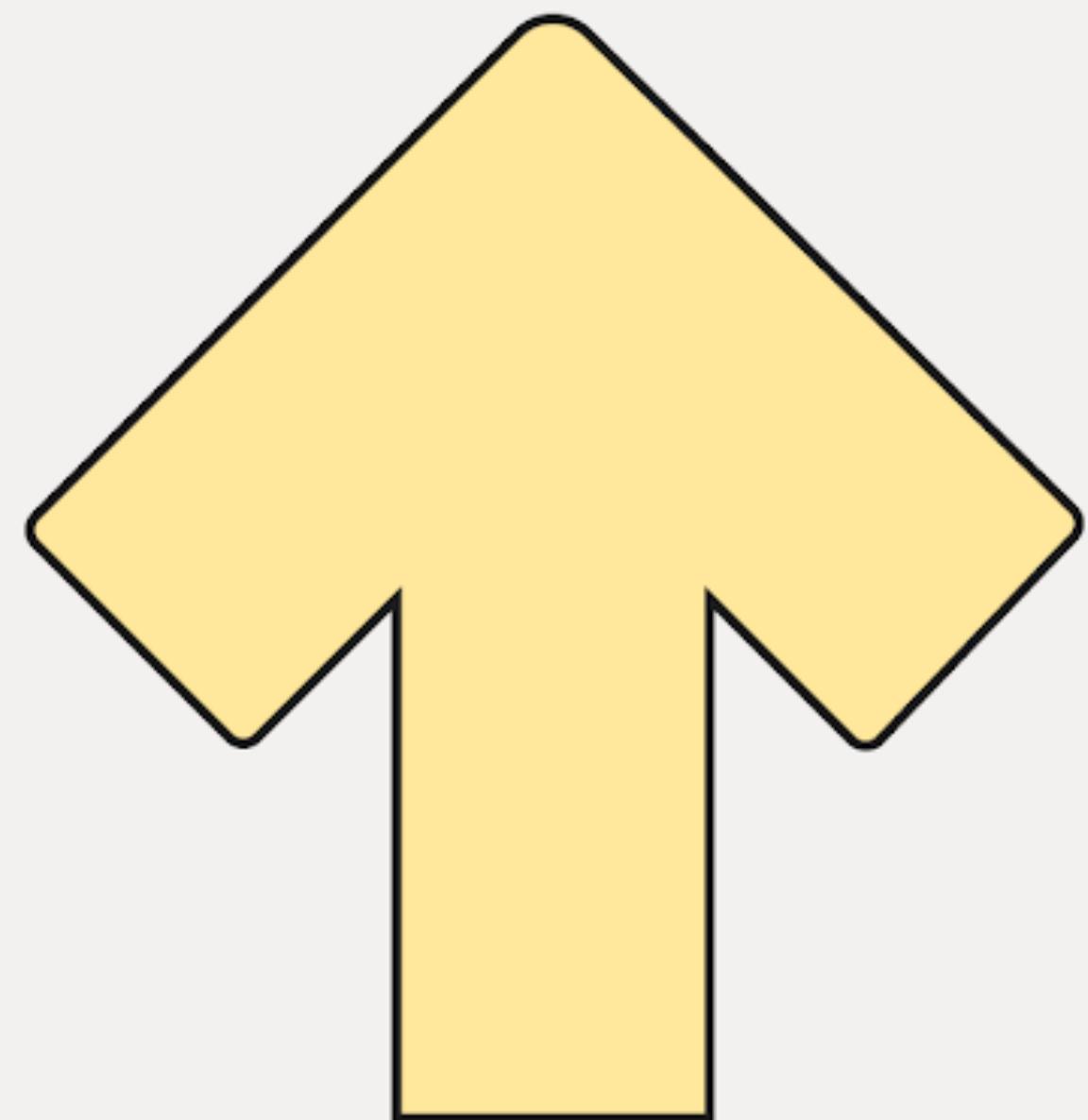
- A globe icon inside a speech bubble shape.
- A smartphone displaying the Google Experts program interface, showing achievements like "2022 Experts" and "2023 Experts".
- The book cover for "EXTENDING ANDROID BUILDS" by El(Bingquan) Zhang.
- A QR code.
- The BinaryTape logo, featuring the word "BINARYTAPE" in a stylized, colorful font with a binary tape graphic.
- The Google Developers Groups logo, consisting of four interlocking colored shapes (blue, green, red, yellow).



01

裝置端 AI Recap

年度趨勢



Google
Developer
Groups

The screenshot shows a window titled 'MLX google/gemma-3n-e4b' in the top bar. The main area displays a conversation:

- You: Histotripsy 是什麼意思?
- google/gemma-3n-e4b: 好的，關於 Histotripsy (生物超音波碎塊術)，我來詳細解釋一下：
Histotripsy，簡單來說，就是一種利用高能量的超音波，來破壞病灶的醫療技術。它的原理類似於傳統的超音波碎石術，但應用範圍更廣泛，而且能量更精準。
- More details:
 - 原理：Histotripsy 使用高頻（通常是數 MHz 到數十 MHz）的超音波，產生共聚焦的能量。這些能量會集中在病灶上，造成細胞破壞和組織碎裂。
 - 應用：主要應用於治療各種疾病，包括：
 - 腫瘤治療：針對腦部腫瘤、肝臟腫瘤、腎臟腫瘤等，可以精準破壞腫瘤細胞。
 - 血管疾病：可以用於治療動脈瘤、血栓等血管問題。
 - 其他疾病：目前也在研究用於治療其他疾病，例如骨骼病變、關節炎等

At the bottom, there is a message input field: 'Send a message to the model...' and a toolbar with icons for file operations and a selected 'rag-v1' model.

The screenshot shows the LM Studio interface. At the top, there's a toolbar with standard window controls (red, yellow, green), a message icon, a plus sign, a gear icon, and a URL field containing "MLX google/gemma-3n-e4b". To the right of the URL are buttons for "合" (Merge) and "Eject". On the far right, there are icons for a profile picture and a gear.

The main workspace has a sidebar on the left with four numbered items: 1 (yellow speech bubble icon), 2 (green document icon), 3 (red folder icon), and 4 (blue circle icon). The main area shows a conversation with "lilium".

You:

Lilium 在國語裡是什麼意思呀幫我翻譯下?
另外，這隻團來自哪裡？

google/gemma-3n-e4b:

「lilium」在國語裡的意思是 百合。
這是一支來自台灣的獨立音樂組合，以其獨特的音樂風格和充滿詩意的歌詞而聞名。

關於「這支團」的資訊，從截圖中可以得知：

- **13,158 monthly listeners:** 牠們的月聽眾數為 13,158 人，表示有一定的聽眾基礎。
- **Verified Artist:** 牠們是經過平台認證的藝人，表示其身份真實。
- **Popular:** 牠們的歌曲在平台上的熱度較高，顯示受到不少人喜愛。

Send a message to the model...

Message input bar with icons for attachments, image, and audio, followed by the model name "rag-v1" and buttons for "User (⌘U)", "Insert (⌘I)", and a scroll arrow.

Bottom status bar: LM Studio 0.3.31 (Build 7), RAM: 5.14 GB | CPU: 0.00 %, and user/cog icons.

12:11 🔈 ⚡ 94
接下來的問題都要用台灣繁體中文...
LFM2-8B-A1B-Q4_0.

Lilium 是什麼意思？

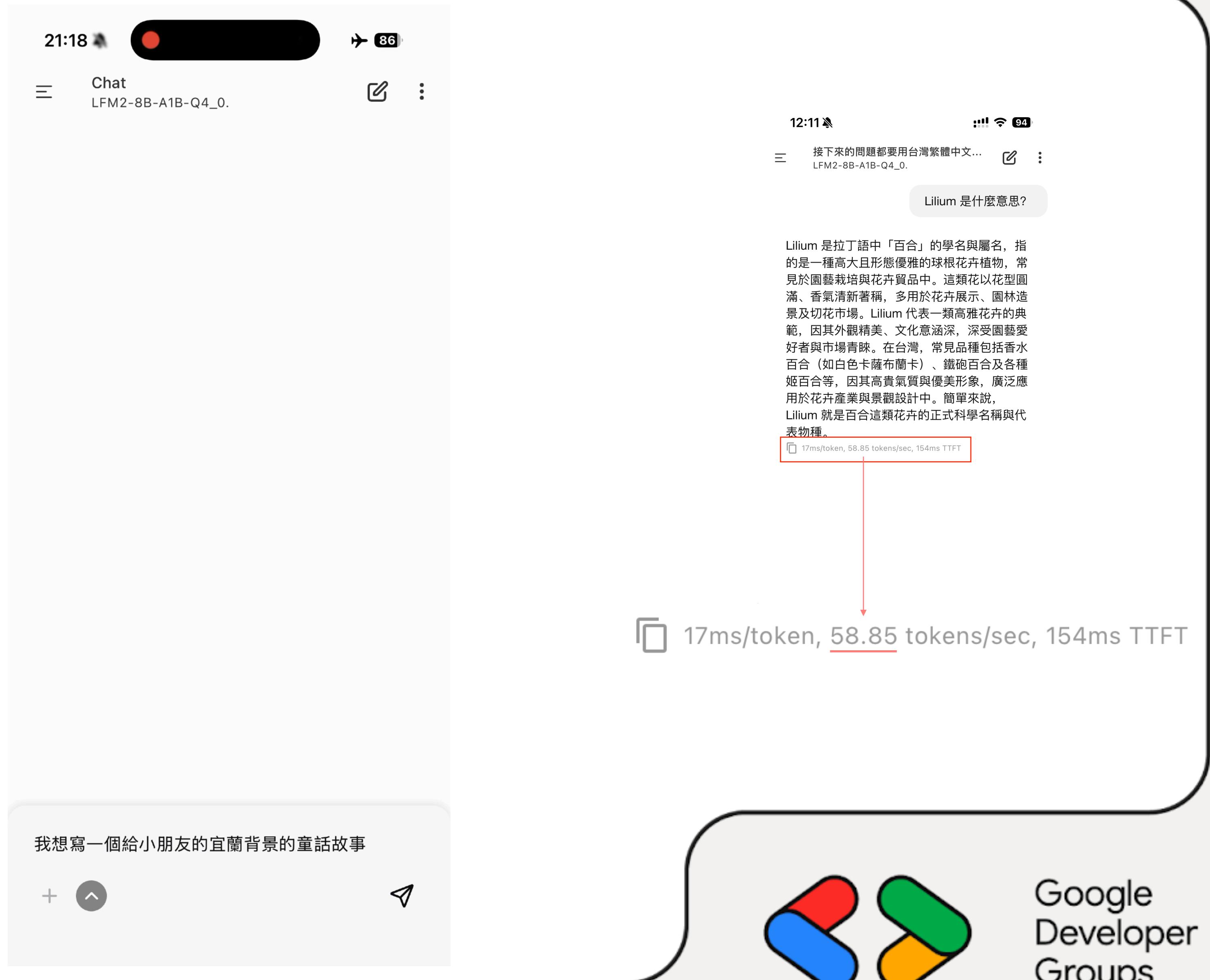
Lilium 是拉丁語中「百合」的學名與屬名，指的是一種高大且形態優雅的球根花卉植物，常見於園藝栽培與花卉貿品中。這類花以花型圓滿、香氣清新著稱，多用於花卉展示、園林造景及切花市場。Lilium 代表一類高雅花卉的典範，因其外觀精美、文化意涵深，深受園藝愛好者與市場青睞。在台灣，常見品種包括香水百合（如白色卡薩布蘭卡）、鐵砲百合及各種姬百合等，因其高貴氣質與優美形象，廣泛應用於花卉產業與景觀設計中。簡單來說，Lilium 就是百合這類花卉的正式科學名稱與代表物種。

17ms/token, 58.85 tokens/sec, 154ms TTFT

17ms/token, 58.85 tokens/sec, 154ms TTFT



Google Developer Groups



Google
Developer
Groups

Gemma 2 / 3(n)

Model	Parameters	Multimodal	Max Context Length on Mobile	Release Date
Gemma 2	2B, 9B, 27B	No	~2048	06/2024
Gemma 3	1B, 4B, 12B, 27B	Yes (4B and above, Vision)	2048~4096	03/2025
Gemma 3n	E2B, E4B	Yes (Vision/Sound)	4096	06/2025

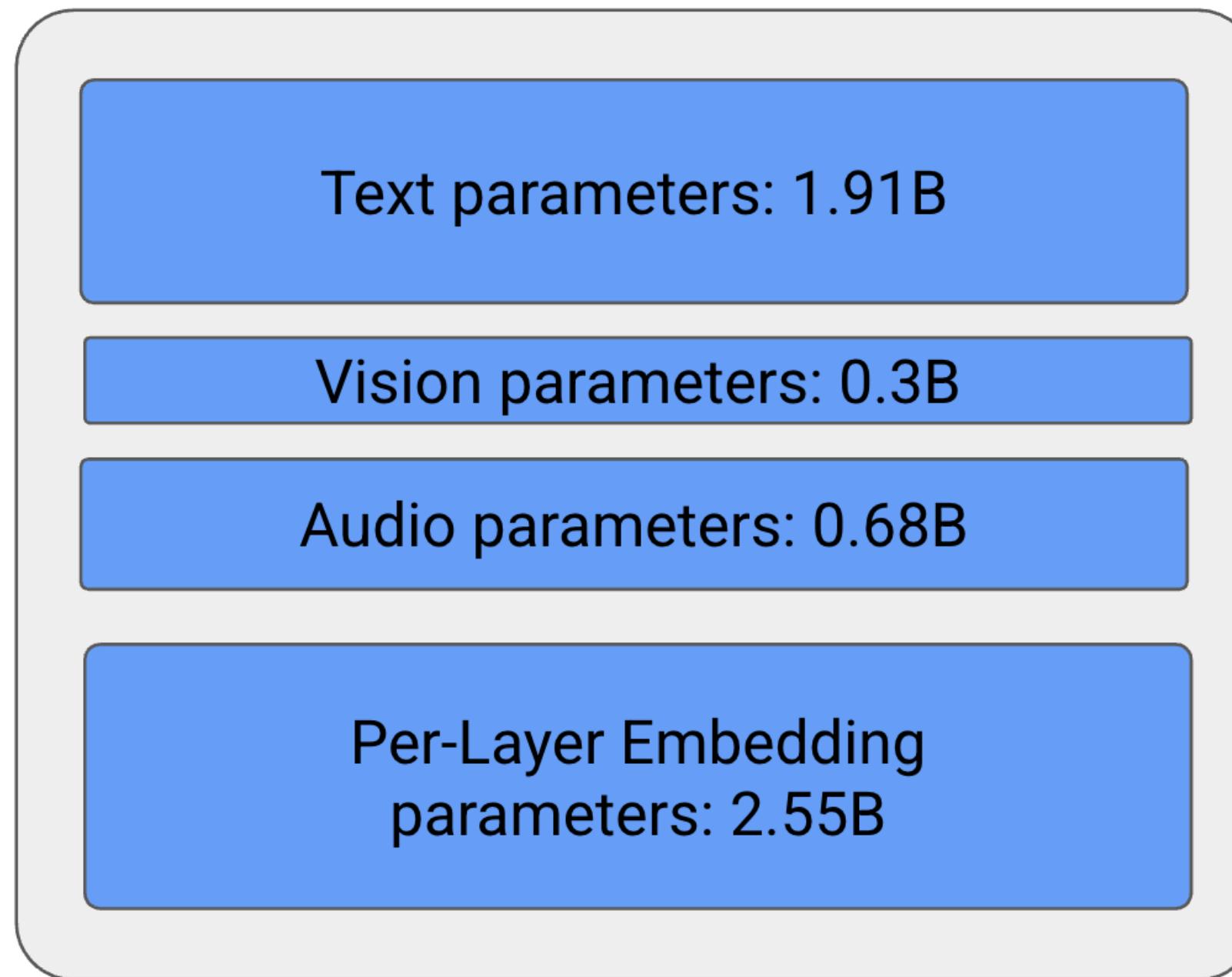


Google
Developer
Groups

Gemma 3n

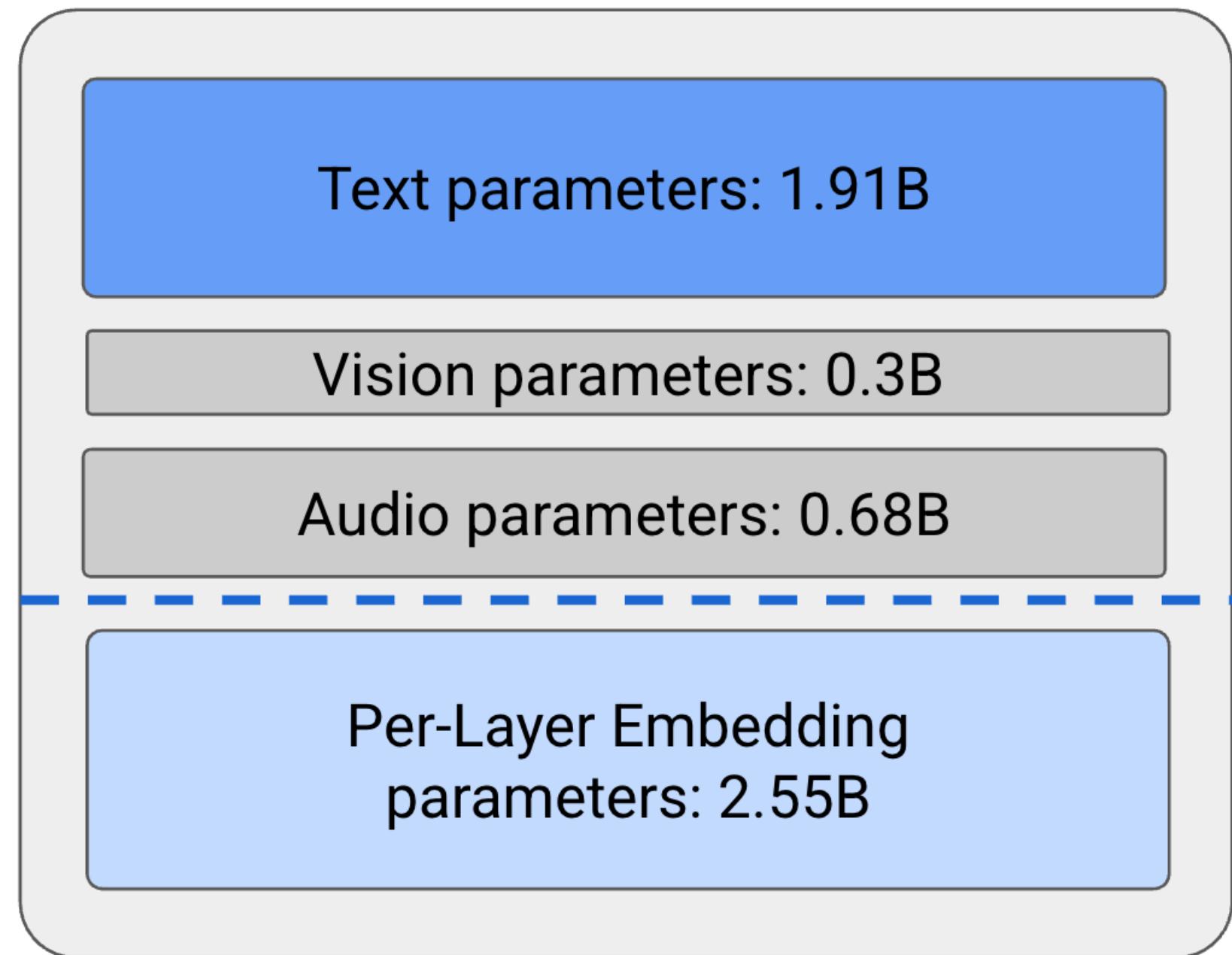
Standard execution

Parameters loaded: 5.44B



with skipped parameters & cached PLE

Parameters loaded: 1.91B



PLE data cached to **fast storage**

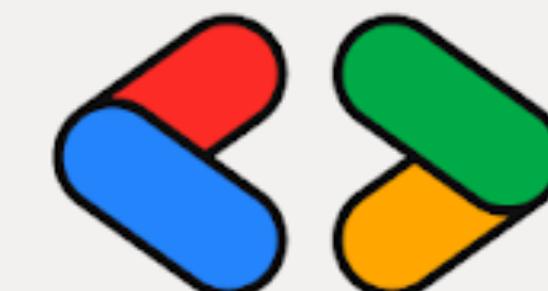


Google
Developer
Groups

Gemma 3n

MatFormer 架構

- **巢狀 FFN 設計**：將前饋神經網路 (FFN) 從單一結構改為巢狀結構（類似俄羅斯娃娃）。
- **彈性部署**：可依據設備的硬體資源，在載入模型時彈性選擇不同複雜度的網路。

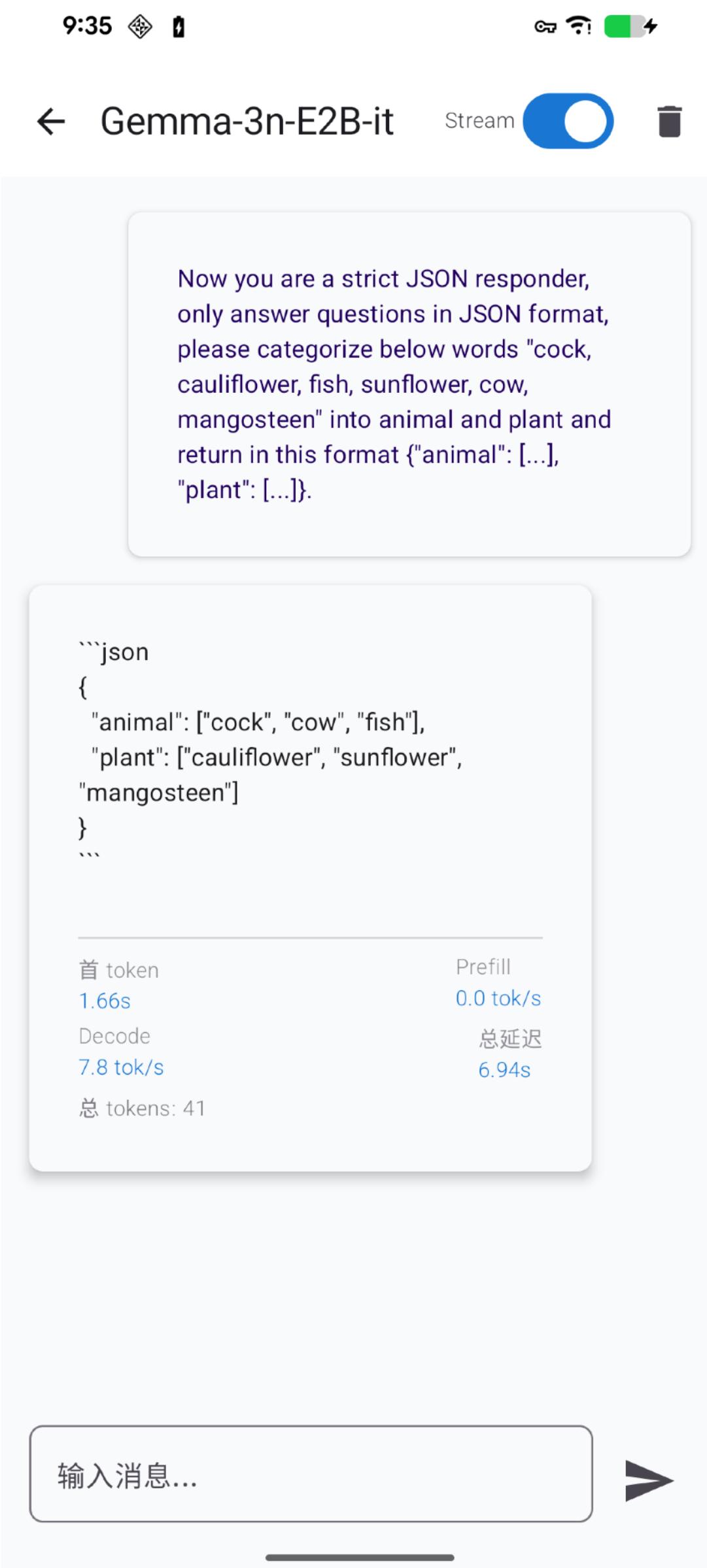


Google
Developer
Groups

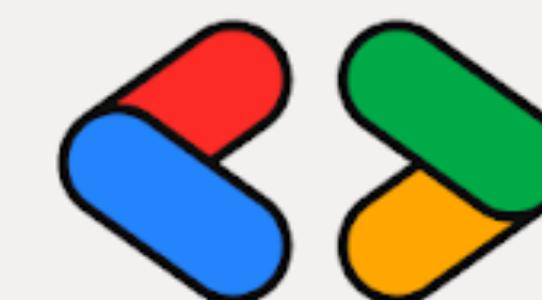
Gemma 3n

全方位升級：

- 邏輯推演能力
- 知識庫擴充
- 格式遵循能力
- 記憶體使用效率



This screenshot shows the Gemma 3n mobile application interface. At the top, there are two purple header cards for "MediaPipe LLM Inference" with the time "10:50" and "10:51". Below each header is a light gray card labeled "Responses generated by user-provided model". The first card shows "GEMMA3_CPU 1141 tokens remaining" and the second shows "GEMMA3_CPU 745 tokens remaining". Both cards have a close (C) and delete (X) button. Below these cards is a pink message input field with the placeholder "Input: [...]". The input field contains the text "Output: { "animal": ["dog", "cat"], "plant": ["rose", "carrot", "banana"] }". To the right of the input field is a light gray message card with the placeholder "User". It contains the same JSON output as the input field. Below the user message is a light gray message card with the placeholder "Model". It contains the text "Now categorize the following words: ['cock', 'cauliflower', 'fish', 'sunflower', 'cow', 'mangosteen']" followed by a long JSON response card. The response card starts with "{ "animal": ["cock", "cauliflower", "fish", "sunflower", "cow", "mangosteen"], "plant": ["cock", "cauliflower"] }". At the bottom of the screen is a black progress bar.



Google
Developer
Groups

Gemini Nano v2 / v3 with MLKit

Model	Multimodal	Release Date
Gemini Nano v2	Yes (Vision)	10/2024
Gemini Nano v3	Yes (Vision/Sound)	08/2025



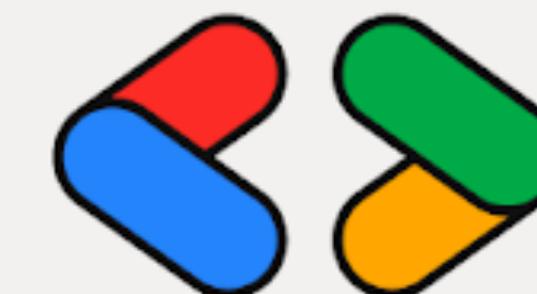
Google
Developer
Groups

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

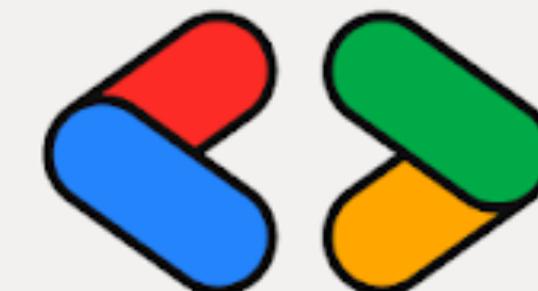
Gemini Nano v2 / v3 with MLKit

Prompt API device support [🔗](#)

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

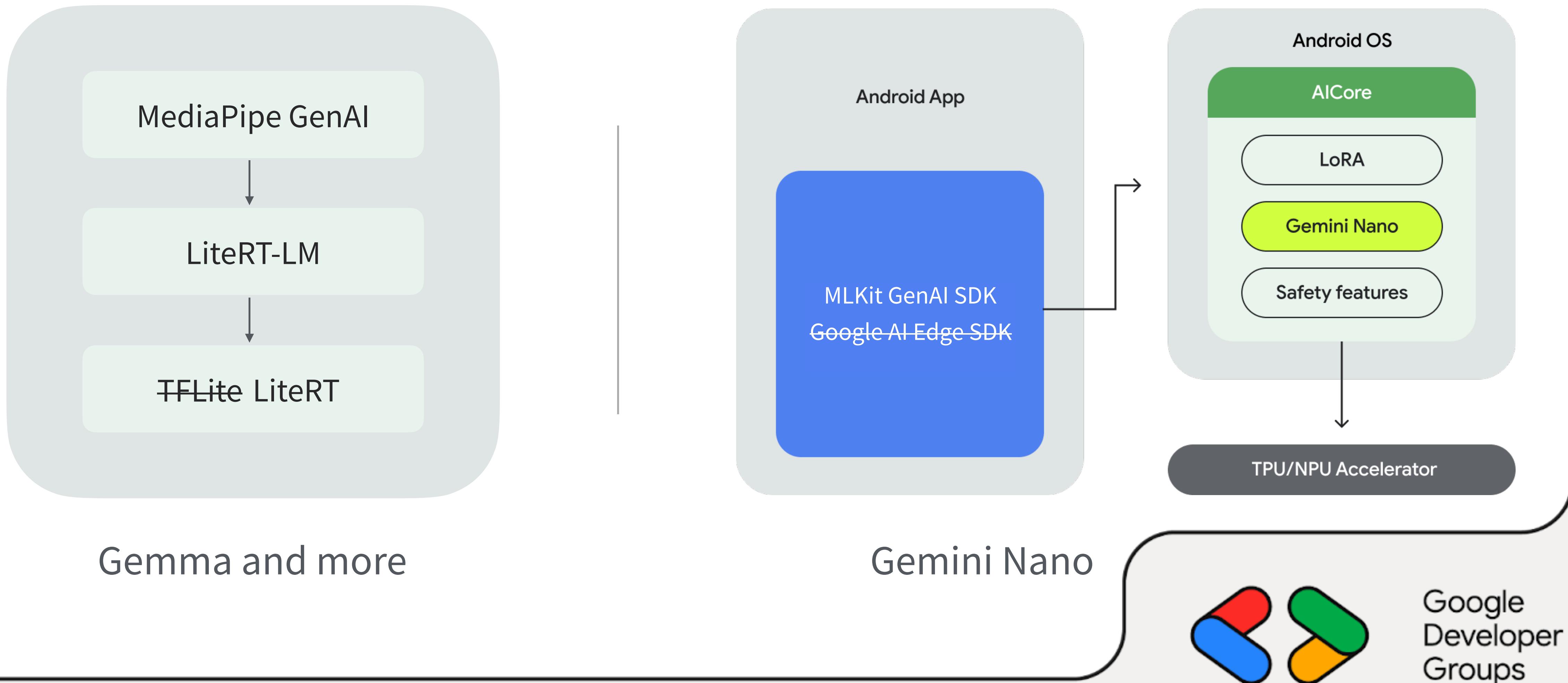
-
- | | |
|---------|---|
| nano-v2 | <ul style="list-style-type: none">• Google: Pixel 9, Pixel 9 Pro, Pixel 9 Pro XL, Pixel 9 Pro Fold• Honor: Magic V5, Magic 7, Magic 7 Pro• iQOO: iQOO 13• Motorola: Razr 60 Ultra• OnePlus: OnePlus 13, OnePlus 13s, OnePlus Pad 3• OPPO: Find N5• POCO: POCO F7 Ultra• realme: realme GT 7 Pro• Samsung: Galaxy Z Fold7• Xiaomi: Xiaomi 15 Ultra, Xiaomi 15 |
|---------|---|
-

- | | |
|---------|--|
| nano-v3 | <ul style="list-style-type: none">• Google: Pixel 10, Pixel 10 Pro, Pixel 10 Pro XL, Pixel 10 Pro Fold |
|---------|--|
-



Google
Developer
Groups

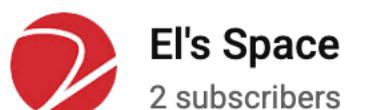
Google 裝置端 AI 推論框架



Google 裝置端 AI 演進歷程 & 未來



Google Engineer Reveals: On-Device Agents and the Future of Personalized AI



El's Space
2 subscribers

Analytics

Edit video



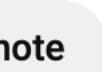
3



1



Share



Promote



Download



46 views 2 months ago

This video was created in collaboration with Nax Qiao (/ @naxqiao)

In this episode, Android GDEs El and Yuang interview Li Na from Google's CoreML team at Google I/O Connect Shanghai 2025! We div ...more



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

雲地协同 Agent

Gemma 3n 与 Gemini 3

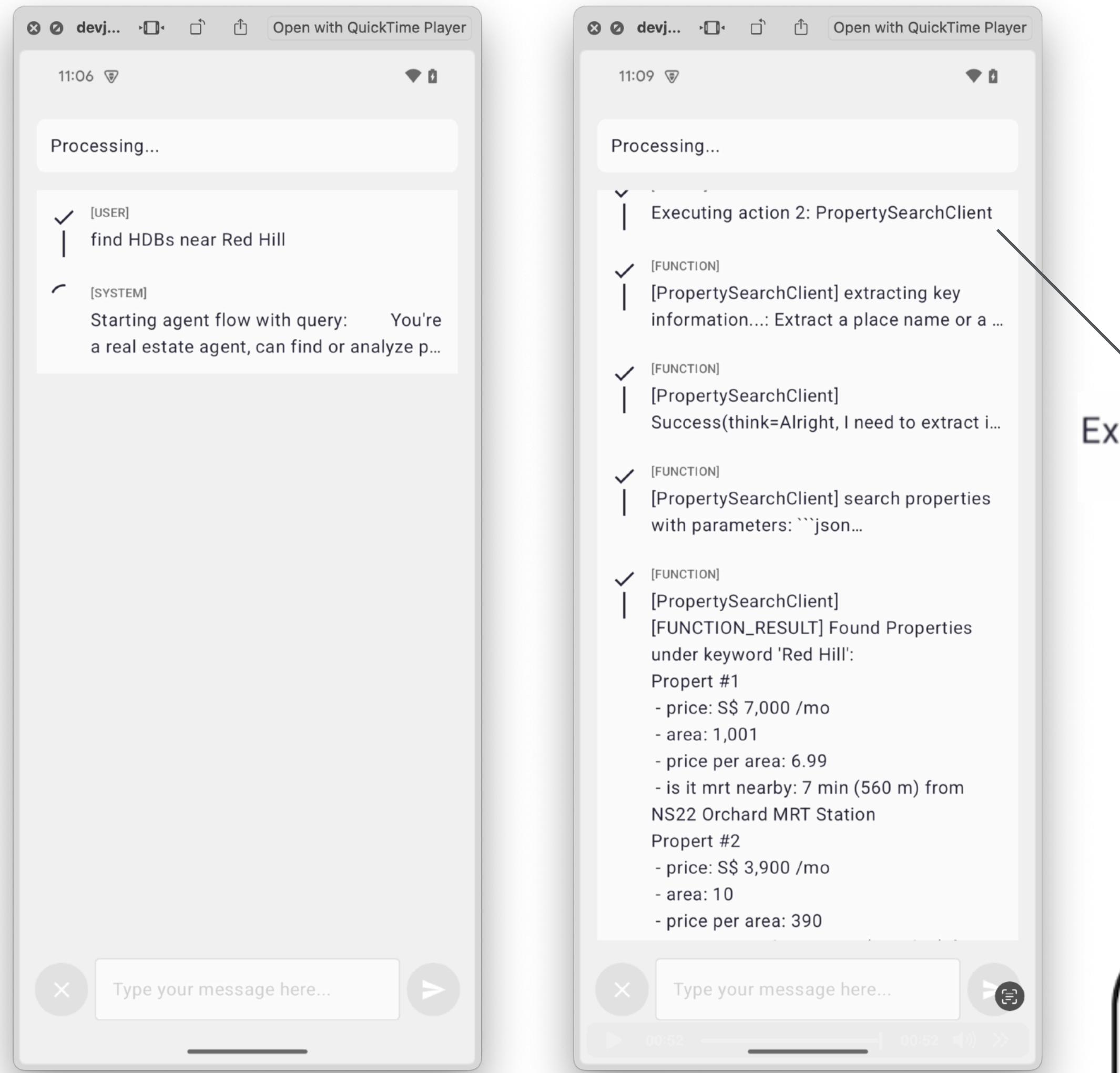


Google
Developer
Groups

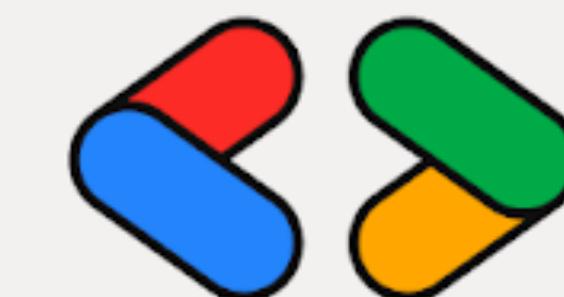
固定情境與 API

Function Calling

- 遠端 API



(DevJam TW 2025)

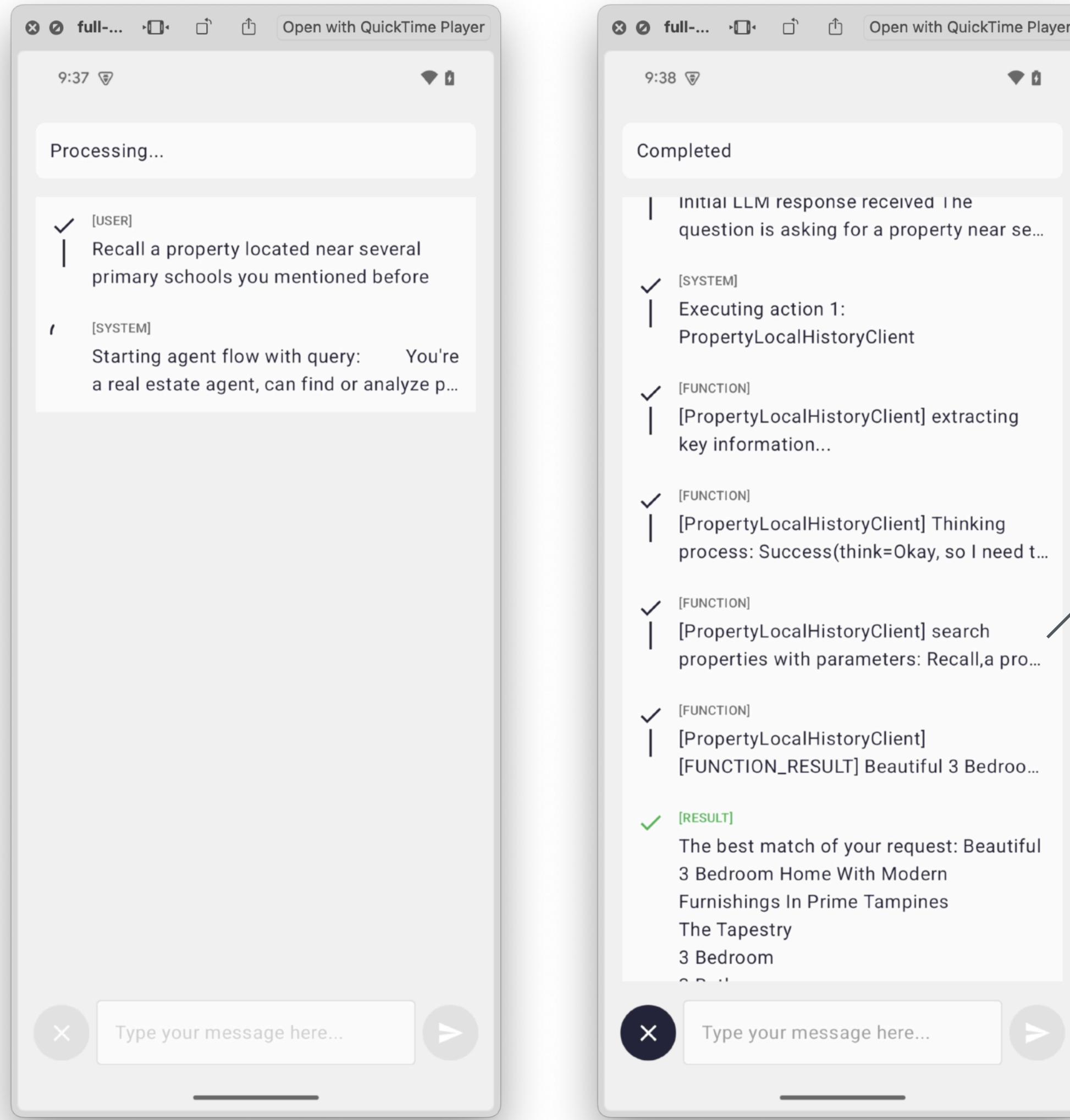


Google
Developer
Groups

固定情境與 API

RAG

- Gecko-110m-en



(DevJam TW 2025)

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



Google
Developer
Groups

現有情境與長期目標

現階段限制：

- 僅針對特定單一任務
- Workflow 適用情境受限

長期目標：

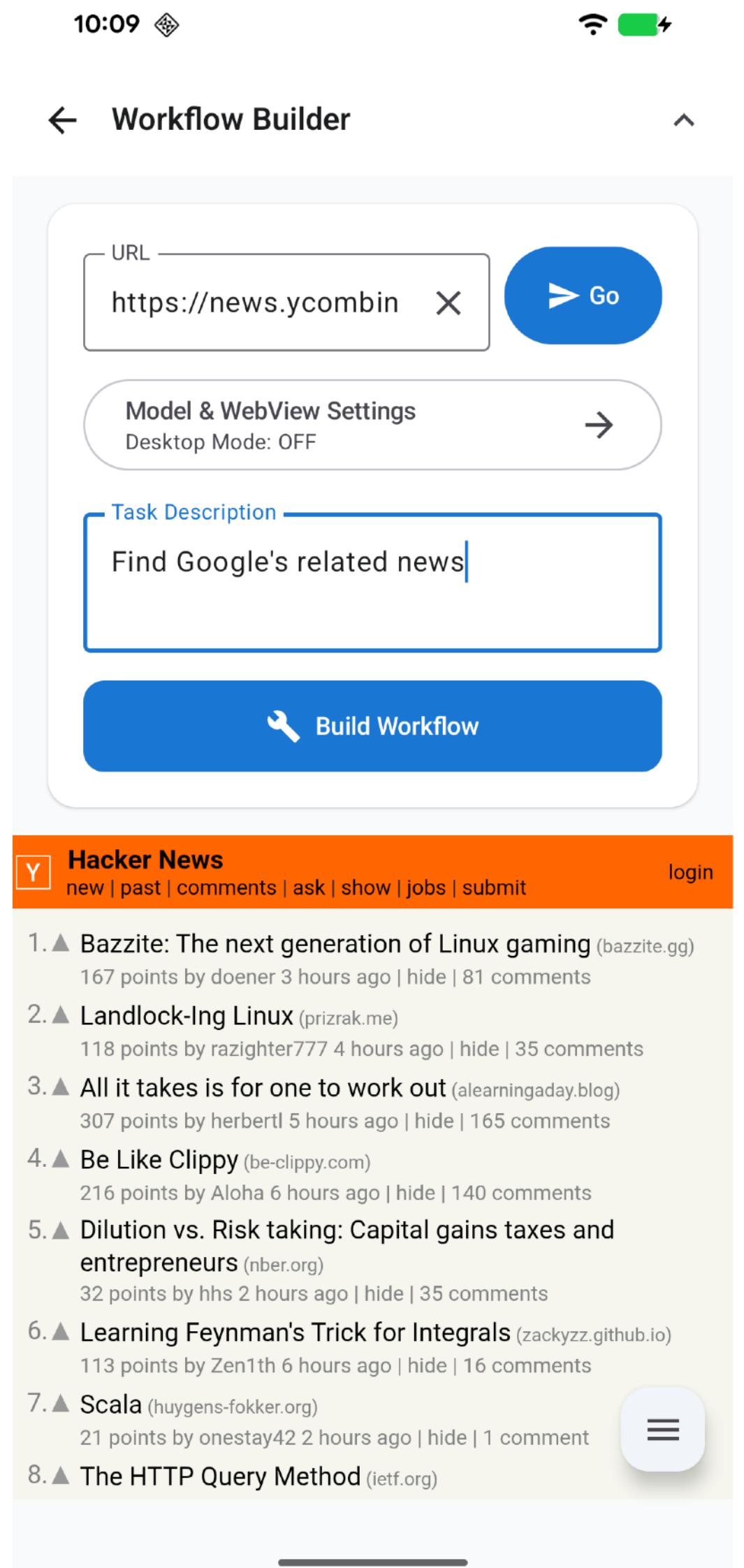
- 複雜情境：適合更高權限的系統級應用
- Agentic / AGI：考驗模型規模與裝置效能



Google
Developer
Groups

通用訂閱管理員

- 我想要知道那位明星的日常
- 我想要知道這家公司每天的市場反應
- 我不會任何 *coding skills*, 我只想說一句話它就幫我解決問題
- 我超懶只想說一次，之後每天早上它都告訴我最新動態



- 網站不設限
- 任務類型不拘（搜尋、過濾、摘要等）
- 僅受限於模型的 *Context Length* (上下文長度)

Guess how?



Google
Developer
Groups

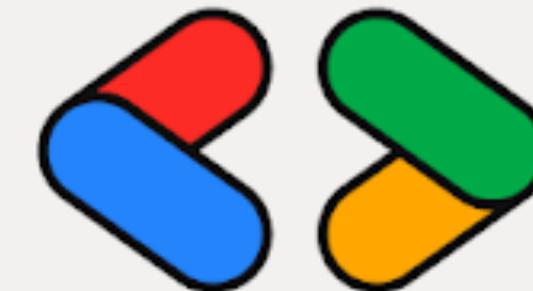
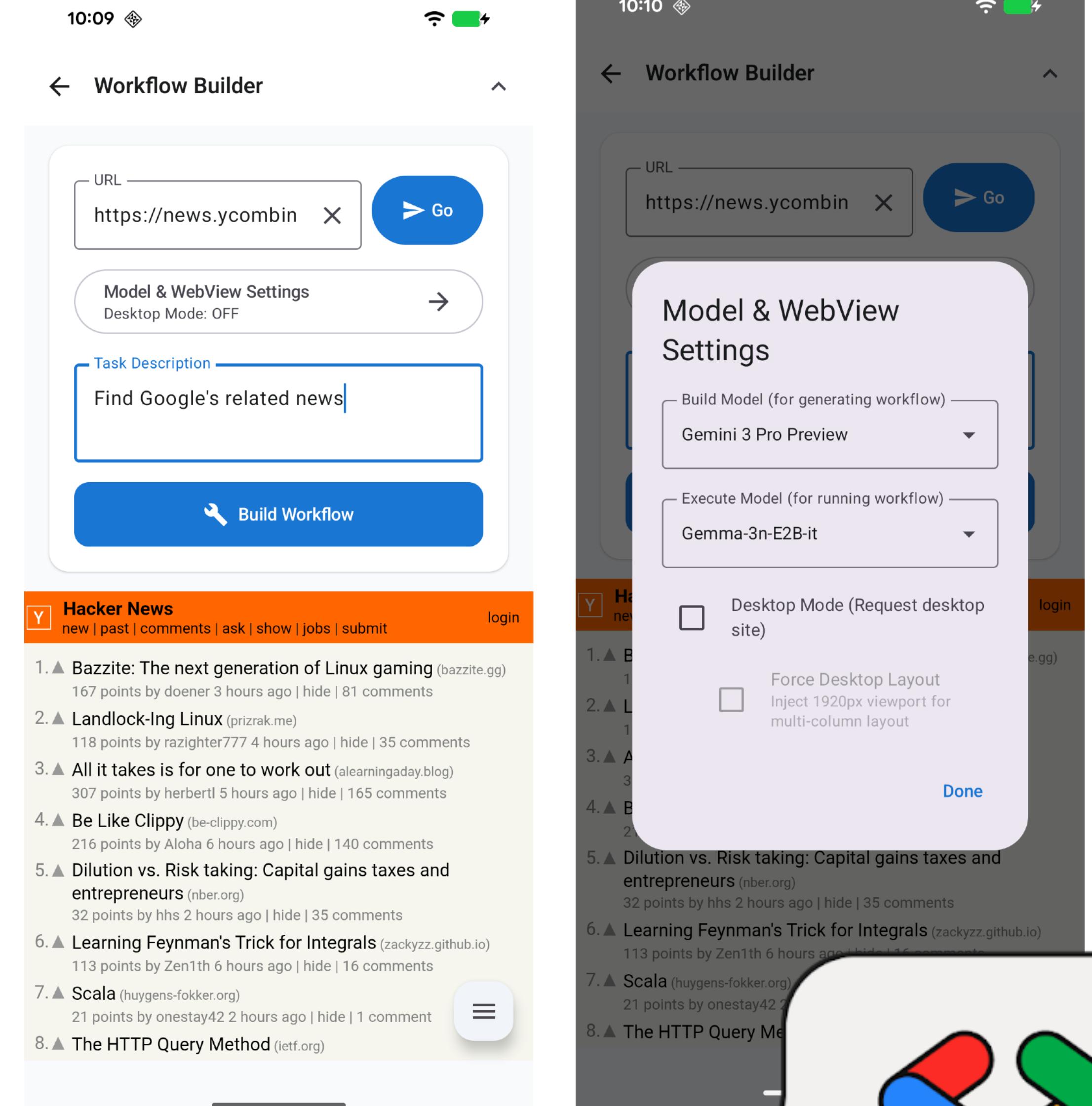
通用訂閱管理員

Agents on Mobile 2025

無料 LLM 雲地協作新玩法

雲地协同 Agent

Gemma 3n 与 Gemini 3



Google
Developer
Groups

Phase 1: Build

工作流程建置階段

Gemini 3 Pro

輸入

使用者目標 + 初始網址

載入頁面

WebView 載入並簡化 HTML

Propose

LLM 分析並提議下一步動作

Verify

在 WebView 中執行驗證

Commit

成功後加入工作流程

↻ 迴圈最多 20 輪

Workflow JSON

完整的工作流程定義
包含所有步驟、類型
與變數 Mapping

Phase 2: Execute

工作流程執行階段

Gemma 3n E2/4B

載入工作流程

讀取步驟清單與設定

變數解析

處理 {{variable}} Placeholder

執行步驟

JS / LLM / Loop / Navigate

儲存結果

存入變數上下文

背景排程處理與推播

定時執行工作流程並推播結果

→ 依序執行所有步驟



Google
Developer
Groups

Google AI Studio

< Dashboard

API keys

Projects

Usage and Billing

Logs and Datasets

Changelog ↗

API Keys

API quickstart

Create API key

Group by

• API key

Project

Filter by

All projects

Key	Project	Created on	Quota tier	⋮
...-8jY Zoe Test 1	ZoeTest1 gen-lang-client-0401173082	Nov 3, 2025	Set up billing Free tier	



Can't find your API keys here?

This list only shows API keys for projects imported into Google AI Studio.
Import other projects to manage their associated API Keys. You can also
create a new API Key above. [Learn more](#)

Import projects



Google
Developer
Groups

Zoe > main

Google Pixel 8a app

Pixel 8a API 36.1

Logcat Logcat

Google Pixel 8a (41071JEKB02904) Android 16, API 36 package:mine

D Installing profile for com.binarytape.zoe
D setTopOnBackInvokedCallback (unwrapped): android.app.Dialog\$\$ExternalSyntheticLambda2@fcfe567
D setTopOnBackInvokedCallback (unwrapped): androidx.activity.OnBackPressedDispatcher\$Api34Impl\$createOnBackPressedAnimationCall
D hide(ime())
I com.binarytape.zoe:8041763d: onCancelled at PHASE_CLIENT_ALREADY_HIDDEN
D setTopOnBackInvokedCallback (unwrapped): androidx.compose.ui.window.Api33Impl\$\$ExternalSyntheticLambda0@22e1e34
I Flattened final assist data: 1248 bytes, containing 3 windows, 10 views
D setTopOnBackInvokedCallback (unwrapped): null
D hide(ime())
I com.binarytape.zoe:c1b2a607: onCancelled at PHASE_CLIENT_ALREADY_HIDDEN
D setTopOnBackInvokedCallback (unwrapped): null
D updatePointerIcon called after view was removed
W Cleared Reference was only reachable from finalizer (only reported once)
W userfaultfd: MOVE ioctl seems unsupported: Try again
D hide(ime())
I com.binarytape.zoe:154a51e5: onCancelled at PHASE_CLIENT_ALREADY_HIDDEN
D show(ime())
I com.binarytape.zoe:bf3a03bb: onRequestShow at ORIGIN_CLIENT reason SHOW_SOFT_INPUT_BY_INSETS_API fromUser false
D Setting requestedVisibleTypes to 511 (was 503)
I Flattened final assist data: 704 bytes, containing 1 windows, 4 views
I com.binarytape.zoe:e625f5dc: onRequestShow at ORIGIN_CLIENT reason SHOW_SOFT_INPUT fromUser false
D show(ime())
I com.binarytape.zoe:e625f5dc: onCancelled at PHASE_CLIENT_REPORT_REQUESTED_VISIBLE_TYPES
D setTopOnBackInvokedCallback (unwrapped): android.view.ImeBackAnimationController@151bfb5
D WindowInsets changed: ime:[0,0,0,958]
W Initializing without READ_DEVICE_CONFIG permission. enabled=false, interval=1, missedFrameThreshold=3, frameTimeThreshold=1000
I com.binarytape.zoe:bf3a03bb: onShown
D hide(ime())
D setTopOnBackInvokedCallback (unwrapped): androidx.activity.OnBackPressedDispatcher\$Api34Impl\$createOnBackPressedAnimationCall
D Setting requestedVisibleTypes to 503 (was 511)
D hide(ime())
I com.google.android.inputmethod.latin:26665f4e: onCancelled at PHASE_CLIENT_ALREADY_HIDDEN
D WindowInsets changed: ime:null
I system_server:33565fa2: onCancelled at PHASE_CLIENT_ON_CONTROLS_CHANGED

Workflow Builder

URL: https://news.ycombinator.com

Model & WebView Settings: Desktop Mode: OFF

Task Description: 抓前 10 篇新聞裡和瀏覽器 Browser 緊密相關的，然後基於前 3 條留言判斷是否樂觀（解釋部分保持在 100 字內）

Build Workflow

Hacker News

new | past | comments | ask | show | jobs | submit | login

- ▲ System 7 natively boots on the Mac mini G4 (macos9lives.com) 237 points by ibobev 11 hours ago | hide | 56 comments
- ▲ Leak confirms OpenAI is preparing ads on ChatGPT for public roll out (bleepingcomputer.com) 246 points by fleahunter 2 hours ago | hide | 191 comments
- ▲ Hachi: An Image Search Engine (eagledot.xyz) 9 points by warangal 31 minutes ago | hide | discuss
- ▲ WinApps: Run Windows apps as if they were a part of the native Linux OS (github.com/winapps-org) 220 points by klausiliveira 10 hours ago | hide | 100 comments
- ▲ How ICE is becoming a secret police force (theconversation.com) 40 points by KnuthIsGod 38 minutes ago | hide | discuss
- ▲ WebR – R in the Browser (webr.sh) 32 points by creat4 4 hours ago | hide | 4 comments

Zoe main

Google Pixel 8a app

Pixel 8a API 36.1

Logcat Logcat

Google Pixel 8a (41071JEKB02904) Android 16, API 36 package:mine

2025-11-29 23:25:55.321 12889-13464 AidlBufferPool com.binarytape.zoe D bufferpool2 0xb40000784
2025-11-29 23:25:55.321 12889-13464 AidlBufferPool com.binarytape.zoe D bufferpool2 0xb40000784
2025-11-29 23:25:55.322 12889-13464 AidlBufferPoolAcc com.binarytape.zoe D evictor expired: 2, ev:

11:25

Workflow Builder

URL https://tixcraft.com > Go

Model & WebView Settings Desktop Mode: OFF

Task Description 帶我搜搜看五月天的最新資訊，總結他們的近期活動，你可以直接使用這個搜尋連結：
<https://tixcraft.com/activity/{{searchKey}}>

Build Workflow

TW 中文

tixcraft ticketmaster Taiwan

搜尋節目

24/7 WITH US TWS TOUR '24/7 WITH US' IN KAOHSIUNG 2026.01.31 (SAT) KAOHSIUNG MUSIC CEN

1:1

Zoe README.md

```
{  
  "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}}"  
    }  
  ]  
}
```

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



Thank You! / Q&A

El Zhang (2BAB)

Google Developer Expert - Android



Google
Developer
Groups