Google AI 产品版图与生态位研究报告

报告日期：2026.5.17 报告人：wang hua

执行摘要

如果把“Google AI”按**对外公开、以 AI 为主要价值主张、可访问/可购买/可下载的产品家族**来统计，而不是把每一个零散功能按钮都当成独立产品，那么Google 已经形成了一套非常完整的“**从芯片到模型、从开发平台到工作流、从消费者入口到科学平台、从闭源旗舰到开源权重**”的 AI 全栈体系。其最重要的几个支点是：TPU 与 AI Hypercomputer 基础设施、Gemini/Gemma/Imagen/Veo/Lyria/Chirp 等模型家族、Google AI Studio 与 Vertex AI（现已演进为 Gemini Enterprise Agent Platform）的开发与企业平台、Gemini app / Search AI Overviews / Workspace / NotebookLM 等高流量入口，以及 SynthID、Model Armor、AI Principles 等安全治理层。官方到二零二五年末披露的数据已经显示出显著规模：AI Overviews 月活达到 20 亿、Gemini app 月活超过 6.5 亿、已有 1300 万开发者在用 Google 生成式模型，超过 70% 的 Google Cloud 客户在使用其 AI 能力；到二零二六年五月初，Gemma 系列累计下载量已超过 5 亿次。[1]

Google 的核心生态位，不是单纯做一个“最强聊天机器人”，也不是只做“企业 API 供应商”，而是试图成为 **AI 时代的基础操作层**：上游用自研芯片和云基础设施控制供给，中游用多模态模型族控制能力分发，下游用 Search、Android、Workspace、YouTube、Cloud 和 NotebookLM 等入口控制用户触点，再通过开放权重、开发工具、协议和合作伙伴网络把第三方生态拉进来。这个位置与 OpenAI 更偏“模型 + 助手”、与 Anthropic 更偏“模型 + 企业 API”、与 Meta 更偏“开源权重”、与 Adobe/Runway 更偏“创意工具”都不同。Google 的独特之处在于：**其 AI 不是一条产品线，而是一个覆盖消费、企业、开发、研究和基础设施的多层复合系统**。[2]

技术上，Google 的最大优势是**模态覆盖面极广**：Gemini 覆盖文本、图像、音频、视频、代码与 agent 能力；Imagen 做图像；Veo 做视频；Lyria 做音乐；Chirp 做语音；AlphaFold 做生命科学；Gemma 把这套技术开放到边缘设备与第三方生态。商业上，Google 又采用了清晰的分层变现：消费者订阅（Google AI Plus / Pro / Ultra）、企业座席与工作套件（Workspace、Code Assist、Agentspace/Enterprise app）、开发者与云按量调用（Gemini API / Vertex AI）、以及通过 Search 和 Android 把 AI 渗透进既有广告与平台收入结构。[3]

但 Google 也有明显短板。第一，**产品命名与分层复杂度偏高**：Vertex AI、Generative AI Studio、Agent Studio、Agent Builder、Agentspace、Gemini Enterprise app、Gemini for Workspace、Gemini app、NotebookLM、Deep Research 等之间存在重叠和迁移成本。第二，**Search 级规模带来了比多数同行更高的错误与合规暴露面**：AI Overviews 在上线后很快因错误答案引发舆论，Google 也公开说明了修正措施；此后又持续面临欧洲出版商和监管机构就 AI Overviews 与内容使用提出的竞争与媒体多元性投诉。第三，Google 虽然在开放权重方面通过 Gemma 迅速追赶，但其前沿闭源模型的完整训练数据与配方披露仍明显少于真正的开源社区期望。[4]

结论上，Google AI 已经构建出一个**高度一体化、分发能力极强、同时兼顾闭源旗舰与开放生态的复合型生态位**。如果说 OpenAI 更像“AI 原生应用与模型公司”，Meta 更像“开源权重发动机”，Anthropic 更像“企业级安全模型公司”，那么 Google 更像是“**AI 的互联网级平台公司**”：它的真正壁垒，来自把模型、工具、工作流、流量入口和算力统一到同一家公司内部。[5]

研究范围与口径

本报告采用以下口径：**将“Google AI 产品”定义为 Google、Google Research、Google DeepMind、Google Cloud、Google Workspace、Google Labs 等体系内，对外公开命名、可公开访问或购买、且以 AI 为主要价值主张的产品家族或服务家族**。因此，本报告纳入模型家族、API 与开发平台、企业 AI 产品、消费者 AI 应用、科学平台、硬件与 AI 基础设施、开放权重/开放工具，以及仍在售的传统云 AI API；**不把**每一个内嵌在 Gmail、Photos、Pixel 或 Android 中的单点 AI 功能都单列为独立产品，而是把它们视为 Gemini 或相关模型在下游产品中的“集成落地”。这是为了保证可审计性与可比性。

另一个必要说明是：Google 在二零二五年后出现了较明显的**命名重构**。例如，Vertex AI 在当前官方目录中已被表述为 **Gemini Enterprise Agent Platform**；Agentspace 的部分能力又被并入当前 Cloud 目录中的 **Gemini Enterprise app** 叙事中。因此，报告在必要时会同时给出“当前名称 / 历史名称”，并以**首次公开发布时的名称**作为时间线锚点，以降低混淆。[6]

本报告优先使用 Google 官方博客、产品页、文档、模型卡与研究平台页面；对市场地位、监管争议与外部采用补充使用 Reuters、Statcounter、Synergy/Statista、Stanford CRFM HELM 与 Artificial Analysis 等二手或第三方来源。对无法通过本次公开检索精确锁定的历史发布日期、定价 SKU 或训练数据细节，统一明确标注为“未公开说明”或“本报告未解析到”。[7]

Google AI 产品版图

核心模型与科学平台

产品家族	首次公开发布	官方定位	核心能力	底层模型/技术	目标用户/市场	定价/许可	主要集成/API/平台	重要更新截至 2026-05-14	主要来源
Gemini	2023-12-06	Google 的旗舰多模态模型家族	文本、图像、音频、视频、代码理解与生成；推理；agentic workflows	1.5 明确引入 MoE；2.0 强调原生图像/音频输出与原生工具使用；2.5 为“thinking model”；3/3.1 强化推理、工具调用、agentic coding	消费者、开发者、企业	Gemini API 按 token 计费并有免费层；消费者通过 Gemini app 与 Google AI 计划；企业通过 Vertex/Workspace/Cloud	Gemini app、Google AI Studio、Gemini API、Vertex/Agent Platform、Search、Workspace、Android 等	1.5（长上下文）→ 1.5 Flash → 2.0（agentic era）→ 2.5（thinking）→ 3 / 3.1（更强推理与音频/图像能力）	官方：[8]
Gemma	2024-02-21	轻量级开放权重模型家族	文本生成、总结、问答、推理；后续扩展	与 Gemini 共享研究技术；Gemma 3n 明确采用 PLE 参数缓存	开发者、研究者、边缘与本地部署用户	开放权重，允许负责任的商业使用	Hugging Face、Kaggle、Vertex AI、	Gemma 2、CodeGemma、PaliGemma、	官方：[9]

产品家族	首次公开发布	官方定位	核心能力	底层模型/技术	目标用户/市场	定价/许可	主要集成/API/平台	重要更新截至 2026-05-14	主要来源
			到多模态、翻译、嵌入、医疗与函数调用	与 MatFormer；Gemma 4 支持文本/图像输入，部分小模型支持音频，最长 256K，上百语言			Google AI Studio、Cloud Run、AI Edge	RecurrentGemma、Gemma 3、Gemma 3n、MedGemma、EmbeddingGemma、TranslateGemma、FunctionGemma、Gemma 4	
Imagen	2024-05-14 作为当前商业化旗舰家族广泛推出	文本到图像生成家族	高质量图像生成、编辑、排版文字、风格与摄影级生成	Imagen 3 在 2024 I/O 公开；Imagen 4 于 2025-05 上 Vertex AI；部分场景后续与 Gemini Image/Nano Banana 产生重叠	创作者、开发者、企业营销与内容团队	Vertex 按量/企业合同；在 Gemini/Workspace/Labs 中随订阅或产品额度提供	Vertex AI、Gemini、ImageFX、Whisk、Workspace/Vids 等	Imagen 3 → Imagen 4；编辑能力持续增强	官方：10
Veo	2024-05-14	Google DeepMind 视频生成模型家族	文本到视频、图像到视频、电影语言控制、后续加入音频视频联合生成与更强物理真实感	2024 首发 1080p 与电影风格控制；后续 Veo 2 / Veo 3 / 3.1 逐步强化真实感、编辑与音频	创作者、营销团队、开发者、企业媒体生产	Vertex 按量/企业合同；Flow / VideoFX / Gemini 依订阅或额度提供	Vertex AI、VideoFX、Flow、Gemini 部分功能	Veo → Veo 2 → Veo 3 → Veo 3.1 / 3.1 Lite	官方：11
Lyria	2025-04-09 进入 Vertex 企业叙事；消费者侧 2026-02 在 Gemini app 更广泛落地	音乐与音频生成模型家族	根据文本、图像等提示生成音乐；可控制情绪、歌词、节奏；支持水印	Lyria 3 Pro/Clip 面向开发者；Lyria RealTime 支持实时音乐生成；产出带 SynthID 水印	创作者、开发者、企业内容团队	Vertex 企业接洽；Gemini/Google AI 计划按订阅/额度	Gemini app、Vertex AI、MusicFX、Music AI Sandbox、Workspace/Vids（后续）	2025 企业预览加入 Vertex；2026 Lyria 3 在 Gemini 更大范围可用	官方：12
Chirp	2025-04-09 以 Chirp 3 企业化推进	语音识别与语音生成模型家族	多语种 ASR、自动语种识别、说话人分离、HD 声音、即时自定义声音	Speech-to-Text 页面称其训练于数百万小时音频与数十亿文本句子；Speech v2 / TTS 当前均接入 Chirp 3	语音产品、联络中心、媒体、企业开发者	STT/TTS 按量计费	Speech-to-Text API v2、Text-to-Speech、Vertex AI 媒体能力	2025 Chirp 3；随后增加 Instant Custom Voice、多说话人改进	官方：13
LearnLM	2024-05-14	基于 Gemini、面向学习与教学优化的模型家族	更符合学习科学原则的教学、辅导、分层提示与个性化学习体验	官方称其"基于 Gemini，并以教育研究为基础进行微调"	教育产品、学习场景开发者、学校与学生	作为模型与产品能力嵌入 Gemini、Classroom 等；独立商业定价未公开	Gemini、Google for Education、AI Studio 预览访问	2024 发布；2024-12 技术报告称在教学表现上优于其他模型，并持续并入 Gemini 2.0 等	官方：14
MedLM / MedGemma	MedLM：2023-12-13；MedGemma：2025-05-20	医疗行业专用模型家族	医疗问答、摘要、临床辅助、医疗影像与多模态医学理解	MedLM 模型卡明确称其由 Med-PaLM 2 等医学调优模型支撑；MedGemma 走开放权重路线	医疗机构、医疗软件商、健康 AI 开发者、研究者	MedLM 在 Vertex AI 上提供并已弃用；MedGemma 为开放权重	Vertex AI（MedLM）、Gemma/Health AI Developer Foundations（MedGemma）	MedLM 2025-09 停用；MedGemma 4B/27B 发布后持续迭代到 1.5	官方：15
AlphaFold 平台	AlphaFold：2020 研究突破；AlphaFold DB 全球大规模发布 2022-07；AlphaFold 3 与 AlphaFold Server：2024-05-08	AI for Science 代表平台	蛋白质结构预测；AlphaFold 3 扩展到蛋白质与配体、核酸等相互作用预测	AlphaFold 3 对蛋白与其他分子交互的预测相对已有方法实现至少 50% 提升；研究者可用 AlphaFold Server 免费提交复杂分子预测	生命科学研究者、药物研发、学术界	AlphaFold DB 免费开放；AlphaFold Server 免费工具；AlphaFold 3 代码/权重学术访问	AlphaFold DB、AlphaFold Server、DeepMind science 页面	2024-11 起 AlphaFold 3 模型代码对学术界开放	官方：16
SynthID	2023-08-29（图像 Beta）	AI 内容水印与识别技术	对图像、文本、视频、音频植入不可感知水印；提供检测入口	由 DeepMind、Google Research 等合作推进；2024 扩展到文本与视频；2025 推出 Detector 门户	Google 自有产品、开发者、创作者、企业与审核场景	技术嵌入 Google 产品；部分文本实现已开源参考实现	Gemini、Imagen、Lyria、Veo、Cloud/Labs、SynthID Detector	2023 图像 Beta → 2024 文本/视频 → 2025 Detector	官方：17

开发者平台、云服务与企业构建层

产品家族	首次公开发布	官方定位	核心能力	底层模型/技术	目标用户/市场	定价/许可	主要集成/API/平台	重要更新截至 2026-05-14	主要来源
Google AI Studio + Gemini Developer API	2024-05 公共开发者推开	Gemini Developer API 的原型设计与快速构建入口	Prompt 试验、模型选择、代码导出、API key、文件/音频/图像/视频与实时能力试用	直接接入 Gemini 系列与部分图像/音频能力	独立开发者、创业团队、原型验证者	API 提供免费层与按量付费层；AI Studio 本身作为开发入口免费使用	Web IDE、REST/API、SDK、Cloud Run 部署、论坛社区	2024 后持续加入 2M context、代码执行、Live API、webhooks、性能分层	官方：18
Vertex AI / Gemini Enterprise Agent Platform	2021-05-19（Vertex AI GA）	Google Cloud 的统一 AI/ML 与生成式 AI / agent 平台	训练、调优、监控、部署、Model Garden、RAG、治理、评测、企业运行	200+ 模型；当前官方表述为 Gemini Enterprise Agent Platform，承接 Vertex AI	企业开发者、数据科学家、平台团队	Cloud 按量计费、容量层或合同价；新客户有试用额度	Cloud、Model Garden、Agent Studio、Workbench、BigQuery、GKE、Cloud Run	2023 引入 Model Garden / Generative AI Studio；2025 后向 agent 平台重构	官方：19
Vertex AI Agent Builder / Agent Search	2024-04-09	企业级生成式应用与 agent 构建平台	搜索、RAG、对话、agent、网站和应用内个性化搜索	以 Gemini、向量搜索、Grounding、企业连接器为底座	企业应用开发者、客服系统、知识管理、站内搜索	Agent Search 公开有按查询计费；Agent Builder 多为 Cloud 用量/合同价	Vertex、网站/应用搜索、企业文档与数据连接器	从 Gen App Builder / Enterprise Search 演进为 2024 的 Agent Builder 与 Agent Search	官方：20
Google Agentspace / Gemini Enterprise app	2024-12-13（Agentspace）	面向企业员工的 AI 代理与 AI 搜索工作平台	跨连接器搜索、No-code agent、内置 NotebookLM Plus、企业数据接地与安全	基于 Google Cloud AI、搜索与代理能力	企业员工、知识工作者、IT/业务团队	以企业销售与合同为主，公开统一标价有限	企业数据源连接器、NotebookLM、Google Cloud	到 2026 当前 Cloud 目录已以 Gemini Enterprise app 的方式呈现更统一的 agent 工作平台叙事	官方：21
Gemini Code Assist	2024-04（由 Duet AI for Developers 演进）	面向编码与云工程的 AI 助手	代码补全、整段函数生成、错误定位、漏洞修复、SQL 辅助、IDE/Cloud Shell 对话	当前已深度接入 Gemini 2.5 及后续模型	开发者、数据工程师、云平台团队	Standard 月付约 22.8 美元/用户；Enterprise 月付约 54 美元/用户；亦有 no-cost tier	VS Code、JetBrains、Cloud Shell、Google Cloud Console	2025 以后增加 agent mode，部分原"tools"功能被 agent mode 取代	官方：22
Firebase Genkit / Vertex AI for Firebase SDKs	2024-05-14（Genkit Beta）	面向应用开发者的开源 AI 应用框架与 Firebase 集成层	工作流、RAG、工具调用、评测、本地调试 UI、部署到 Cloud Run/Firebase；客户端 SDK 直接调 Gemini	统一接模型提供商；支持 Google AI / Vertex	全栈应用开发者、移动与 Web 开发者	开源框架；底层模型调用按相应 API/云计费	Firebase、Cloud Run、Next.js、Node.js、客户端 SDK	持续扩展插件与遥测能力，成为 Google 面向 app 开发者的轻量层	官方：23
Agent Development Kit / A2A	2025-04-09	面向多 agent 系统的开源开发框架与互操作协议	代码优先 agent 开发、调试、部署；agent 间安全通信与协作	ADK 提供多语言支持；A2A 作为开放协议推动互联	高级开发者、企业平台团队、生态伙伴	开源；云部署与配套服务按云资源计费	Google Cloud、GitHub、Linux Foundation 生态、多语言 SDK	2025-06 A2A 捐赠至 Linux Foundation，联合 AWS、Microsoft、Salesforce、SAP、ServiceNow 等	官方：24
Google AI Edge	2024 统一文档入口逐步定型；AI Edge Gallery 2025-09 公测上架	Google 的端侧 AI 栈	MediaPipe Solutions、LiteRT、端侧 LLM/多模态/RAG/LoRA、样例与测试应用	LiteRT 建立在 TensorFlow Lite 基础上；MediaPipe 提供通用视觉/语音/LLM 任务	移动端、Web、IoT、边缘设备开发者	开源/免费工具层为主；底层硬件与部署成本自担	Android、iOS、Web、MediaPipe、LiteRT、AI Edge Gallery	2025 Gemma 3/3n 端侧支持强化；2026 增加 NPU benchmark 与更多端上生成式能力	官方：25
Model Armor	2025-02-03	生成式与 agentic AI 的运行时安全层	Prompt injection 防护、敏感信息泄露防护、有害内容筛查、MCP 场景保护	作为 Google Cloud 服务，独立于具体模型	企业安全团队、平台团队、合规团队	提供免费层，之后按服务使用计费	与 Gemini Enterprise Agent Platform、LangChain、MCP 等集成	2026 已支持更广泛的 inline protection 与代理场景	官方：26
Gemini CLI	2025-06-25	开源终端 AI agent	在终端中使用 Gemini 完成编码、命令、agent 流程与开发辅助	依赖 Gemini 模型；强调个人开发者高配额接入	开发者、CLI 用户、自动化爱好者	免费开源；底层模型访问按计划/配额	终端、本地开发环境、与 Genkit/其他开发工具联动	成为 Google agentic developer story 的关键入口之一	官方：27

消费者、工作流与创作入口

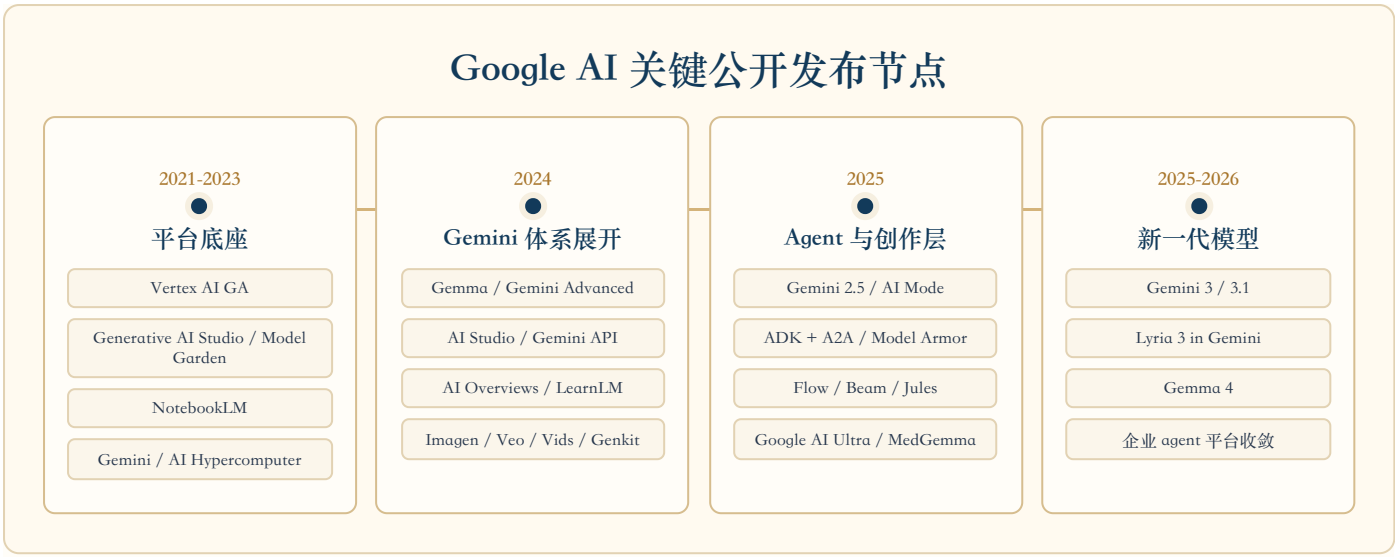
产品家族	首次公开发布	官方定位	核心能力	底层模型/技术	目标用户/市场	定价/许可	主要集成/API/平台	重要更新截至 2026-05-14	主要来源
Gemini app	2024-02-08（Bard 更名并推出移动应用）	Google 面向个人用户的通用 AI 助手	聊天、写作、规划、编码、图像/音频/视频生成能力接入、Live、Deep Research、Gems、跨应用动作	由 Gemini 家族持续升级驱动；后续接入 Lyria、Veo、图像编辑等	大众用户、专业用户、学生、创作者	免费层；Google AI Plus 约 7.99 美元/月，Pro 约 19.99 美元/月；Ultra 于 2025-05 以 249.99 美元/月在美国推出	Web、Android、iOS（Google app）、与 Gmail/Drive/Maps 等 Google 服务联动	从 Ultra 1.0 到 Gemini 2.0 / 2.5 / 3 / 3.1；加入 Deep Research、音乐生成、视频生成等	官方：28
Gemini for Google Workspace	2023-05-11（Duet AI），2024-05-15 以 Gemini 叙事全面扩展	将生成式 AI 融入 Gmail、Docs、Sheets、Slides、Meet、Chat、Drive 等	写作、摘要、表格组织、侧边栏问答、会议记录、文件分析、跨应用协作	2024 年明确升级到 Gemini 1.5 Pro 等模型；后续将 NotebookLM 与 Gemini app 访问纳入工作套件	企业、学校、组织	2025 起多数 Business/Enterprise 计划内含 AI；Business Standard 年付约 14 美元/用户/月等	Workspace 全家桶、Web、移动端	从单点助手发展为嵌入式 AI 层，并增加自动化/agentic 工作流	官方：29
Google Vids	2024-04-09 公布，2024-11 更广泛可用	面向工作的 AI 视频创作应用	自动故事板、草稿生成、素材/脚本/音乐建议、Drive/Photos 集成、模板化视频协作	以 Gemini 与 Google 媒体模型为底层	企业培训、营销、内外部沟通、销售赋能	随部分 Workspace / Gemini 计划提供；个人侧后续纳入 Google AI Pro/Ultra 部分权益	Workspace、vids.new、Google Drive/Photos	从 Labs/早期测试到大多数 Workspace 客户可用	官方：30
NotebookLM	2023-07-12	基于自有资料的 AI 研究助手与“思考伙伴”	基于用户资料做总结、问答、洞察、音频概览、深度研究、表格/视频/学习材料生成	2024 官方明确称其使用 Gemini 1.5 多模态能力；后续与更强 Gemini 能力持续融合	学生、研究者、知识工作者、企业知识管理	免费版 + 付费 Pro/Plus；企业可通过 Workspace 或 Google Cloud 购买	notebooklm.google、Workspace/Agentspace、Google AI 计划部分包含	Audio Overviews、Business/Plus、Deep Research、更多文件类型、Video Overviews、Data Tables	官方：31
Search AI Overviews / AI Mode	AI Overviews：2024-05-14；AI Mode：2025-03-05	把生成式回答与搜索融合到 Google Search	AI 摘要、复杂问题、长问题、多模态查询；AI Mode 进一步强化对话式搜索、后续加入更复杂能力	由 Google 定制版 Gemini 模型驱动；2025 起升级到 Gemini 2.0 / 2.5	亿级搜索用户、广告与搜索生态	消费者免费使用，商业模式仍以搜索广告与平台分发为主	Google Search、Google app、Lens 等	2024 覆盖扩张到 100+ 国家，月度覆盖超 10 亿；到 2025-11 官方称达 20 亿月活	官方：32
Google Beam	2025-05-20（由 Project Starline 演进而来）	AI-first 的 3D 视频通信平台	把 2D 视频流转为逼真的 3D 通话；支持近实时语音翻译方向	使用体积视频模型与光场显示；构建于 Google Cloud 与 AI 能力之上	企业视频会议、远程协作、高端会议室	企业硬件/平台采购，公开统一定价未披露	与 HP、Zoom、渠道合作伙伴共同进入企业市场	2024 与 HP 启动商业化准备，2025 正式更名为 Google Beam	官方：33
Google Labs 创作工具族	ImageFX / MusicFX：2024-02-01；VideoFX：2024-05-14；Whisk：2024-12；Flow：2025-05-20	面向创作者的实验性创作入口	文生图、文生乐、文生视频、图像重混、电影叙事创作	背靠 Imagen、Lyria、Veo、Gemini 等模型	创作者、早期尝鲜用户、营销与内容团队	多为 Labs/订阅额度制；Flow 在发布时向 Google AI Pro/Ultra 美区订阅用户开放	Labs、ImageFX、MusicFX、VideoFX、Whisk、Flow	2024-12 Veo 2/Imagen 3 加入 Labs；2025 Flow 成为更完整的电影工作台	官方：34
Jules	2025-05-20 公测	异步编码 agent	读取代码、修 Bug、写测试、生成变更说明、GitHub 工作流集成	由 Gemini 2.0 时代能力驱动的云端安全执行环境	开发者、个人项目与团队	公测/实验性产品，后续商业化模式未完全定型	GitHub、云端异步执行、开发者工具链	2024-12 试验性亮相；2025-05 进入 public beta	官方：35

基础设施、开放生态与传统专用云 AI 服务

产品家族	首次公开发布	官方定位	核心能力	底层模型/技术	目标用户/市场	定价/许可	主要集成/API/平台	重要更新截至 2026-05-14	主要来源
Cloud TPU / AI Hypercomputer	Cloud TPU v4 对外可用 2021；AI Hypercomputer 2023-12 正式成体系	Google Cloud 的 AI 基础设施底座	大规模训练与推理、TPU/GPU/CPU 一体化、网络/存储/软件协同优化	TPU v4、v5p、Trillium（v6e）、Ironwood（v7）等	大模型训练者、企业 AI 平台、Google 自家产品	云基础设施按量/预留/合同计费	Google Cloud、Vertex/Agent Platform、GKE、Cloud Run	2024 发布 Trillium 并于 2024-12 GA；2025 发布 Ironwood	官方：36
Coral / Edge TPU / Coral NPU	Coral：2019-03-06；Coral NPU：2025-10（超出本报告截止，仅作后续注记）	本地 Edge AI 硬件平台	在低功耗设备上进行本地推理、隐私友好、硬件加速	Edge TPU 单芯片 4 TOPS；支持 LiteRT/TFLite 模型	IoT、嵌入式、工业与边缘设备开发者	硬件售卖，价格因板卡/模组而异	开发板、加速器、SOM、USB/M.2 模组	2019 持续销售；AI Edge 叙事下与 Gemma/MediaPipe 配套更紧密	官方：37
TensorFlow / Keras 生态	TensorFlow：2015 对外开源；Keras 3：2023/2024 官方发布多后端重写	Google 最重要的开源 ML 训练/部署生态之一	训练、推理、部署、模型 API、跨后端深度学习	TensorFlow、Keras 3 支持 JAX / TensorFlow / PyTorch / OpenVINO 等后端	研究者、工程师、教育与工业界	开源	Python 生态、云训练、端侧 LiteRT 等	Keras 3 成为多后端统一接口；LiteRT 继承 TFLite 基座	官方：38
Natural Language AI	精确首发日期未在本次材料中锁定	云端文本理解服务	情感分析、分类、实体抽取等	Google Cloud ML 模型	企业开发者	云 API 按量计费	Cloud API	2026 仍在 Cloud AI 产品目录中	官方：39
Speech-to-Text	精确首发日期未在本次材料中锁定	语音转写 API	语音识别、实时转写、联络中心支持	现可调用 Chirp 3	企业开发者、客服、媒体	按处理音频量计费	STT API v2、Contact Center、Agent Assist	2025 以后以 Chirp 3 为新一代模型	官方：40
Text-to-Speech	精确首发日期未在本次材料中锁定	文本转语音 API	自然语音合成、高清音色、定制声音	现可调用 Chirp 3 HD voices / Instant Custom Voice	企业开发者、音频应用、媒体	按字符计费	Cloud TTS、Vertex Studio 试用	2025-2026 以 Chirp 3 为新一代能力	官方：41
Translation AI	精确首发日期未在本次材料中锁定	机器翻译服务	多语言翻译与本地化	Google 翻译/ML 能力	企业开发者、全球化产品团队	云 API 按量计费	Cloud Translation	2026 仍在 Cloud AI 产品目录中	官方：42
Vision AI	精确首发日期未在本次材料中锁定	图像理解服务	目标检测、OCR、图像理解、自定义视觉模型	预训练 Vision API + AutoML Vision	企业开发者、零售、工业质检等	云 API / AutoML 按量计费	Cloud / Edge / AutoML	2026 仍在产品目录中	官方：43
Video AI	精确首发日期未在本次材料中锁定	视频理解服务	视频层级标签、镜头/帧级元数据、内容发现	Video Intelligence / AutoML Video	企业开发者、媒体与搜索	云 API 按量计费	Video AI	2026 仍在产品目录中	官方：44
Document AI	精确首发日期未在本次材料中锁定	文档理解平台	OCR、表单/票据/合同解析、结构化抽取	文档处理 pipeline 与行业处理器	金融、供应链、政企、后台流程自动化	按页/处理量/处理器计费	Cloud Document AI、RAG/搜索场景	2026 仍为独立核心产品	官方：45
CX Agent Studio	精确首发日期未在本次材料中锁定	对话式 AI 平台	低代码/可视化构建虚拟客服与多轮对话 agent	intent-based + generative AI / LLM mixed stack	客服、电话机器人、网站机器人、智能设备	云服务/企业合同	Web、App、IVR、消息平台	在生成式 AI 时代继续向高级 agent 平台演化	官方：46
Gemini Enterprise for Customer Experience	精确首发日期未在本次材料中锁定	端到端客户体验 AI 应用	高级虚拟代理、人工座席辅助、多渠道客户交互	Google 对话式 AI + 多模态能力	大型客服与联络中心	企业合同价	Contact Center / Cloud AI	2026 仍在 Cloud AI 目录中	官方：47

关键发布时间线

以下时间线选取的是最能代表 Google AI 版图演化的主节点，而非每一个小版本或功能更新；对应事实来自上文各产品的官方发布博客与产品页。[48]



关键对比与产品结构判断

Google AI 内部旗舰矩阵

维度	Gemini	Gemma	Vertex / Agent Platform	Gemini app	Workspace with Gemini	Search AI Overviews / AI Mode	NotebookLM	TPU / AI Hypercomputer
在生态中的角色	前沿通用智力核心	开放权重与边缘/社区扩张器	企业构建与变现主平台	消费者 AI 中枢	办公生产力渗透层	最大规模分发入口	资料驱动研究助手	供给控制与成本/性能护城河
主要用户	开发者、企业、消费者	开发者、研究者、本地部署者	企业与平台团队	个人用户	企业/学校/组织	全球搜索用户	学生、研究者、知识工作者	大模型训练与推理客户
商业模式	API 按量 + 订阅入口	开放许可 + 云部署带动使用	按量 + 合同 + 生态销售	Freemium + Google AI plans	Seat-based 套餐	广告/搜索留存/平台强化	Freemium + Pro/Enterprise	基础设施按量/预留
构筑壁垒方式	多模态 + 长上下文 + agent 能力	打开社区、硬件兼容、本地化	数据、治理、连接器、运维能力	默认助手入口与跨应用动作	日常工作流高频粘性	搜索流量与信息分发垄断优势	用户自有资料接地与高可解释性	自研芯片与系统级优化
主要对手	GPT、Claude、Llama 前沿版	Llama、Mistral、Qwen 系开放模型	Azure/OpenAI、AWS/Anthropic、Databricks	ChatGPT、Claude app、Perplexity	Microsoft 365 Copilot	Perplexity、OpenAI Search、Bing AI	ChatGPT Deep Research、Perplexity Pages	NVIDIA 云 GPU、AWS Trainium/Inferentia
当前最明显的重叠/冲突	与 Imagen/Nano Banana、NotebookLM、Search 发生能力重叠	与 Meta 开源路线正面竞争	与 AI Studio / Agentspace 命名重叠	与 NotebookLM/Deep Research 有边界重叠	与 Agentspace、Gemini app for work 重叠	与 Gemini app "问答"体验重叠	与 Gemini Deep Research 重叠	与第三方 GPU 生态共存又竞争

上表的核心含义是：Google 不是把所有 AI 能力塞进一个产品，而是故意把它们分布到**能力层、平台层、工作流层、分发层、基础设施层**五个不同位置。这样做提高了变现面，但也增加了用户理解成本。其结果是，Google 的 AI 生态明显不是"单产品赢家通吃"，而是"多入口共享底层模型、在不同市场以不同商业模式收割"的组合拳。[49]

按本报告口径划分的产品结构

下图是基于本报告按"产品家族"而非"单个功能点"统计得到的结构分布，用于说明 Google AI 的重心显著集中在**模型 + 平台 + 应用入口**三层，而不是单独押注某一个 chatbot。

生态位与竞争格局

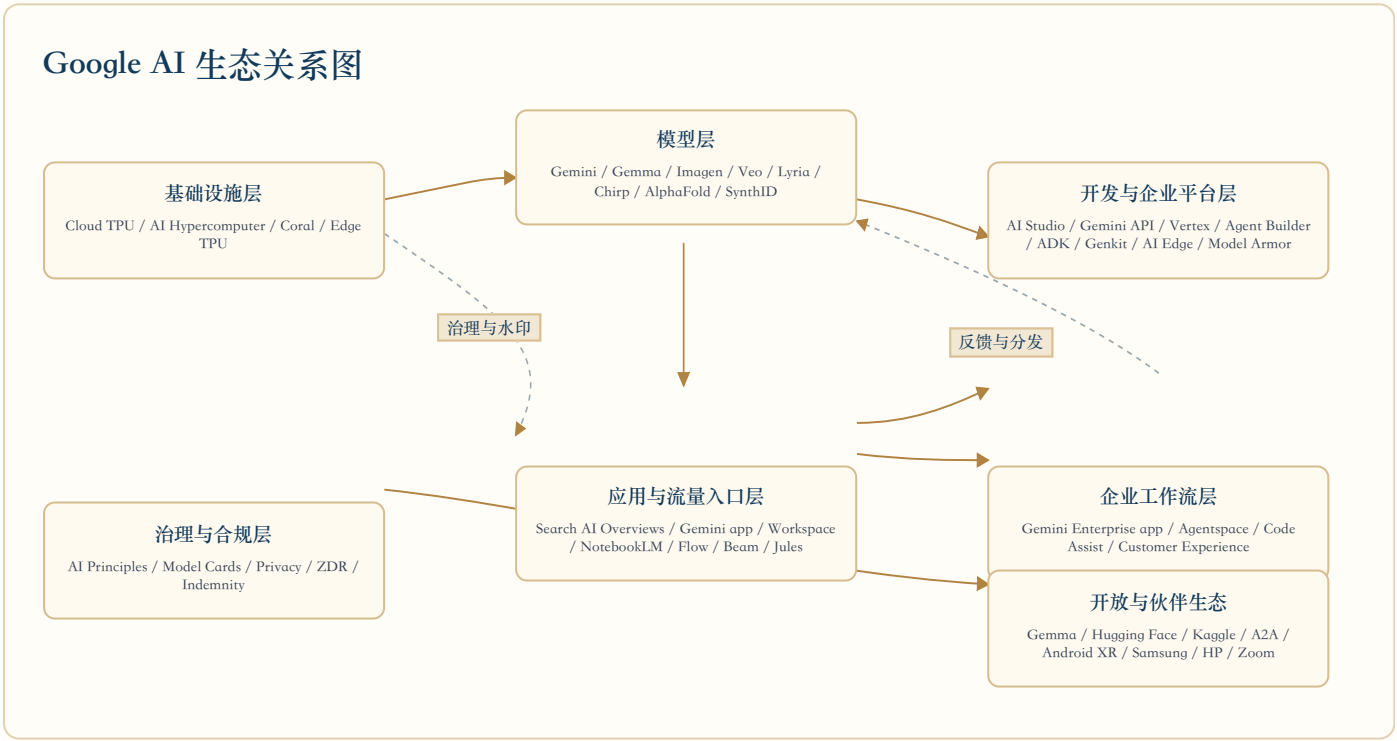
Google AI 的生态位，可以概括为“**互联网级 AI 平台公司**”。它不是只卖模型，也不是只卖应用，而是在同一家公司内部完成了以下闭环：**TPU/AI Hypercomputer 提供供给与成本优势 → Gemini/Gemma/Imagen/Veo/Lyria/Chirp/AlphaFold 提供能力层 → AI Studio/Vertex/ADK/Genkit/AI Edge 把能力变成开发入口 → Search/Workspace/Gemini app/NotebookLM/Android/Beam/Flow 等承担流量与应用入口 → SynthID/Model Armor/AI Principles 提供治理与对外可解释性**。这条闭环在当今主流 AI 厂商中非常少见。OpenAI 没有 Search/Android/Workspace 这类巨大原生入口，Meta 没有同等企业云平台，Anthropic 没有消费级全球分发，Adobe/Runway 没有芯片与通用大模型底座。[50]

Google 的一个关键战略动作，是**同时押注闭源旗舰与开放权重**。Gemini 负责争夺前沿“最好模型”与企业 API 心智，Gemma 则负责吸引研究者、独立开发者、本地部署与边缘设备生态。Gemma 与 Kaggle、Hugging Face、Vertex、Cloud Run、AI Edge 形成天然联动，这使 Google 可以在与 Meta 的开放模型竞争中，不必在“完全开源”与“完全闭源”之间二选一，而是通过开放权重把开发者流量导向自己的工具链和云。官方材料也明确强调了 Gemma 与 Hugging Face、NVIDIA TensorRT-LLM、Kaggle 等的配套关系。[51]

第二个关键动作，是**把企业 AI 从“模型调用”升级为“代理工作平台”**。这条线索从 Gen App Builder / Enterprise Search 演化到 Vertex AI Agent Builder，再到 Agentspace，以及当前 Cloud 产品目录中的 Gemini Enterprise Agent Platform / Gemini Enterprise app，说明 Google 正试图把企业客户从“买 API”转向“买可治理的 agent workforce 平台”。这与 Salesforce、ServiceNow、Microsoft、AWS 都在推动的 agent 平台化方向一致，但 Google 的优势在于它既有搜索、连接器、RAG、模型、芯片，也能把 NotebookLM 和 Workspace 这类工作流入口纳入同一叙事；短板则是命名与产品边界比对手更复杂。[52]

第三个关键动作，是**把消费者入口与企业入口共用底层模型**。AI Overviews、AI Mode、Gemini app、Workspace、NotebookLM 看似是不同产品，实则都在复用 Gemini 家族、搜索接地能力、长上下文与个性化推理能力。官方已经明确：AI Overviews 月活达到二十亿，Gemini app 月活超过六亿五千万，开发者达到一千三百万。与此同时，Google 依然拥有全球搜索约九成份额、移动搜索约九成五份额的分发基础。这意味着即便 Gemini 作为独立 chatbot 的品牌声量常被 ChatGPT 掩盖，Google 依然可以通过 Search、Android、Workspace 把 AI 变成默认层而非“额外应用”。这正是其最深的生态位壁垒。[53]

生态关系图



伙伴网络、对手与空白

Google 在合作上采取的是“**关键入口自持 + 生态协议开放 + 终端和分销合作**”三段式。基础设施端，Google 公开强调 Anthropic、Mistral、AI21 等都在其 AI Hypercomputer 上运行，Anthropic 也被公开点名为在 TPU 上训练模型的领先 AI 公司之一；开放模型端，Gemma 明确与 Hugging Face、Kaggle、NVIDIA TensorRT-LLM 做集成；终端与设备端，Gemini 在三星 Galaxy S24 系列开始大规模合作，Android XR 又与 Samsung 和 Qualcomm 共同推进；会议硬件端，Project Starline/Google Beam 与 HP、Zoom 及渠道伙伴合作；协议端，A2A 被捐赠到 Linux Foundation，与 AWS、Microsoft、Salesforce、SAP、ServiceNow 等共同推进。Google 试图通过这些合作，把“Google AI 不是只能在 Google 自家云里用”的信号放大，同时又把最核心的模型、平台、流量和芯片留在自己手里。[54]

从竞争上看，Google 至少同时面对六类对手。第一类是 **OpenAI / Microsoft**，在前沿模型、代码助手和办公 AI 上与 Gemini、Code Assist、Workspace 直接对打。第二类是 **Anthropic / AWS**，在企业安全叙事与模型 API 领域竞争，同时又在 A2A 等开放生态里合作。第三类是 **Meta / Llama**，在开放权重和开发者心智上与 Gemma 激烈对抗。第四类是 **Adobe / Runway / Midjourney / Pika**，在创意生成与电影工作流上与

Imagen/Veo/Flow 竞争。第五类是 **Perplexity 与新搜索 AI 玩家**，在“AI 原生搜索入口”上针对 Google Search 发起挑战。第六类是 **NVIDIA**，并非在终端应用层竞争，而是在 AI 基础设施控制力和开发者默认心智上与 Google TPU/AI Hypercomputer 形成结构性竞逐。Google 的优势是覆盖面最广；劣势是必须同时在多个战场维护一致体验。[55]

Google 当前最显著的生态重叠点有三组。其一，**Google AI Studio、Agent Studio、Vertex/Agent Platform** 之间存在从个人开发者到企业开发者的连续性，但品牌和界面层级不够清晰。其二，**Gemini app 的 Deep Research / Gemini 对话能力 与 NotebookLM** 都在争夺“研究助手”位置，前者偏开放网络探索，后者偏用户资料接地。其三，**Gemini Image / Nano Banana、Imagen、ImageFX、Whisk、Vids** 正在形成不同层次的图像与视频生产工具，但容易让外部用户搞不清“到底哪个才是官方主线”。这些并不会摧毁生态本身，却会提高获客和教育成本。[56]

技术、商业与采用分析

技术上，Google AI 的第一强项是**模态完整性**。Google 不是只有一个通用 LLM，而是拥有通用大模型 Gemini、开放轻量模型 Gemma、图像 Imagen、视频 Veo、音乐 Lyria、语音 Chirp、水印 SynthID、科学平台 AlphaFold 这一整组横跨文本、视觉、音频、视频、音乐与科学建模的能力栈。其他头部厂商通常在其中两到三类模态上强，但像 Google 这样把“工作流级产品、创作者工具、科学平台、企业 API、移动端侧模型”同时覆盖到位的并不多。[57]

第二强项是**从云到边缘的连续部署能力**。Gemini 负责云端旗舰，Gemma 负责开放权重与本地推理，Google AI Edge 用 LiteRT 与 MediaPipe 把模型拉到手机与设备上，Coral/Edge TPU 则为更小型边缘场景提供硬件支撑。这个结构让 Google 在“端云一体”和“本地隐私”叙事上比只做云端 API 的厂商更完整。特别是 Gemma 3、3n、4 在移动、桌面与工作站上的定位，显然是冲着“把模型从云里释放出来”去的。[58]

第三强项是**供给侧控制力**。Google 既是模型公司，也是基础设施公司。官方一再强调 Gemini 在 TPU 上训练与服务，Trillium 与 Ironwood 又分别强化了训练/推理时代的性能和能效；AI Hypercomputer 则把芯片、网络、软件与消费模式打包成一套系统方案。Google 因而在面对 GPU 紧缺、推理成本飙升时，拥有比纯应用层公司更大的调度与优化空间。这是其最“传统但最硬”的护城河。[59]

弱项同样清楚。首先是**透明度不完全对称**。Google 为 Gemini 3.1 Pro、Gemini 3.1 Flash Image 等提供模型卡，说明安全、限制与评测，但对完整训练语料、参数规模、配方和后训练细节仍保留较多不公开部分；相比开放权重的 Gemma，这种差异更加明显。对前沿闭源模型而言，这当然是行业常态，但在“开发者信任”和“学术可复现性”维度上并非优势。[60]

其次是**大规模搜索产品的错误代价极高**。AI Overviews 在二零二四年五月大范围推出后，Google 自己就专门发文解释其错误来源与修复动作。这与独立聊天产品不同：Search 一旦出错，影响的是全球级流量入口、出版商生态与广告关系，因此 Google 在搜索 AI 上承受着比普通 LLM 产品更高的舆论和监管放大。到二零二五至二零二六年，欧洲出版商和监管机构围绕 AI Overviews 提出的反垄断与媒体多元性投诉，正是这一点的延续。[61]

商业上，Google 的策略并不是单一订阅，而是**四层变现并行**。第一层是消费者订阅：Google AI Plus / Pro / Ultra 为 Gemini、Flow、NotebookLM 等提供更高额度和更强模型。第二层是办公与开发座席：Workspace、Gemini Code Assist、企业代理平台按用户或企业合同销售。第三层是开发与云：Gemini API、Vertex/Agent Platform、Agent Search 等按 token、查询量或云资源计费。第四层是“AI 化既有平台收入”：例如 AI Overviews 并未切断搜索商业模式，而是在广告/搜索框架中嵌入 AI 回答。换言之，Google AI 的营收结构并不会像纯 AI 公司那样集中在“一个订阅按钮”，而是更接近“把 AI 变成所有 Google 主要业务的上层加速器”。[62]

采用侧，Google 的优点在于**开发者、消费者和企业三个漏斗都已经打开**。截至二零二五年十一月，官方称 1300 万开发者已使用 Google 生成式模型、Gemini app 月活超 6.5 亿、AI Overviews 月活 20 亿、超过 70% 的 Google Cloud 客户在使用 AI；Gemma 到二零二六年五月初累计下载量超 5 亿次。仅从这些官方数据看，Google 已经不是“追赶者”，而是在流量分发和生态渗透上进入第一梯队。此外，Google 仍掌握全球约 90% 的搜索市场、移动搜索约 95.5% 的市场，这意味着其 AI 不是在“零起点获客”，而是在海量既有入口之上升级。[63]

治理、隐私与监管态势

Google 在治理层的优势，在于它比许多竞争对手更早把**AI Principles、模型卡、年度进展报告、产品级安全模型和企业合同保护**制度化。官方 AI 页面明确把“大胆创新、负责任开发与协作进步”作为 AI Principles 的框架，Google 也持续发布与安全生命周期、模型卡和责任实践有关的文档。对企业客户，Google Cloud 还提供生成式 AI indemnification（包括训练数据和输出层面的责任承担）、Code Assist 的安全合规模块、以及 Vertex AI 的零数据保留选项。就“把 AI 真正卖给大公司”这件事看，Google 的制度化程度是其重要优势。[64]

在消费者隐私上，Google 的姿态相对清楚，但并不意味着没有争议。Gemini Apps Privacy Hub 明确表示：Gemini 聊天当前**不用于投放广告**；同时，它也说明了**人类审阅、活动保留与设置开关**等机制，这意味着普通消费者使用 Gemini app 时仍需理解“产品改进用数据”和“聊天保留/临时聊天”的差别。与此形成对比的是企业侧：Gemini for Google Cloud 与 Vertex AI 文档强调按 Cloud 术语、DPA、项目隔离和可选零数据保留来处理数据。这说明 Google 实际上在运行两套明显不同的隐私叙事：**消费产品强调可控但默认更便利，企业产品强调合同和隔离**。[65]

在内容透明度上，Google 选择把 **SynthID** 做成横跨图像、文本、视频、音频的统一水印方案，并在二零二五年推出 Detector 门户。与只做“AI 生成内容标识声明”相比，Google 明显更想把“可识别性”做成平台能力，再嵌入 Gemini、Imagen、Veo、Lyria 等各类产品。这既服务于监管趋势，也服务于企业客户的品牌安全要求。[66]

不过，Google 也是当前最容易受到监管和内容产业压力的 AI 公司之一，原因恰恰来自其生态位。首先，AI Overviews 直接冲击新闻与内容网站流量，Reuters 报道显示，欧洲出版商已就 AI Overviews 和 Google 使用在线内容/摘要问题提出正式投诉或推动监管调查。其次，在欧盟 DMA 和更广泛的数据共享/竞争规则讨论中，Google 又因为其搜索与 Android 地位，被同时要求“开放更多接口”和“保证隐私与安全”。Google 公开主张某些数据共享要求会增加隐私风险，说明它在“平台开放”和“平台控制”之间面临典型的大平台两难。换言之，Google 在治理上做得越制度化，反而越容易成为监管样板间。[67]

结论与开放问题

综合来看，Google AI 已经形成了一个**不是单点爆款，而是分层联动的 AI 生态系统**。其生态位不是“最好用的一个模型”，也不是“最受欢迎的一个助手”，而是一个更具平台属性的位置：它既能控制供给侧（芯片、云、系统软件），又能控制能力层（Gemini 等模型），还能控制需求侧入口（搜索、Android、Workspace、NotebookLM、Gemini app）。因此，Google AI 的真正竞争逻辑不是与某一个对手做一一对位，而是把任何外部对手都拉回自己更擅长的层面：把 chatbot 竞争拉回 Search 和 Android，把模型竞争拉回 TPU 和 Cloud，把创作工具竞争拉回 YouTube/Workspace/Labs，把企业 agent 竞争拉回连接器、治理与工作套件。[68]

从战略上看，Google 最值得关注的并不是“下一个单模型基准谁第一”，而是它能否继续把当前分散的命名与产品边界收敛成更清晰的三层结构：**消费者层（Gemini / Search / NotebookLM）**、**开发者层（AI Studio / Gemini API / Genkit / AI Edge）**、**企业层（Agent Platform / Agentspace / Workspace）**。一旦这三层叙事变得更清晰，Google 的分发、算力和生态协同优势会被更充分地感知；反之，如果命名与定位持续复杂化，其优势将持续被“单品心智更强”的对手稀释。[69]

开放问题与局限

本报告有四点需要明确保留。其一，“**全部产品**”**采用的是产品家族口径**，而非把每个细分功能与每个模型 SKU 都拆开统计。其二，部分**传统云 AI 服务**的最早公开发布日期，在本次公开材料中未能全部精确回溯，因此在表格中已明确标注“未精确锁定”。其三，Google 在二零二五至二零二六年对部分 Cloud AI 产品进行了较明显的**品牌重组和名称迁移**，这会影响一一映射。其四，许多前沿模型的**完整训练数据、参数与配方仍未公开**，因此“底层技术”部分只能基于官方已披露内容与模型卡概括，不能等同于完整技术审计。