

海外市场公有云 MaaS 服务运营模式与横向对比分析报告

分析范围：海外市场（不含中国大陆） 时间基准：2026 年 8 月

对象：AWS、Google Cloud、Microsoft Azure、阿里云、腾讯云、字节跳动（火山引擎）、百度智能云

一、分析大纲

- MaaS 定义与范畴**——在公有云上的含义、核心特征，与 IaaS / PaaS / SaaS 的本质区别
- 主流交付与运营模式归纳**——预训练模型 API、多模型市场、微调定制、托管推理、Agent/应用平台、行业垂直、开源+云托管等模式及其商业逻辑
- 逐厂商分析**（七家，聚焦海外市场）——核心产品、服务模式、目标客户、定价、技术优势、区域部署
- 横向对比**——模型与开放度、性能成本、生态、合规、海外策略五个维度
- 总结与选型建议**——趋势、格局、差异化，以及对潜在使用者的实操建议

市场背景（用于校准量级）：Gartner 2026 年 7 月预测，全球 AI 平台与模型市场 2026 年约 **642 亿美元**，**同比 +63%**，其中基础生成式模型细分市场 234 亿美元、**同比 +104%**；AI-as-a-Service 整体市场预计从 2026 年约 235 亿美元增至 2033 年 1890 亿美元（CAGR 34.7%）。MaaS 已从“新概念”转为云厂商收入增长的核心引擎。

二、MaaS 定义：它是什么，不是什么

MaaS (Model-as-a-Service) 指云厂商把“已训练好的模型能力”作为一项**托管服务**对外提供：用户通过 API / SDK 按调用量（通常按 token）付费，即插即用，**无需自建 GPU 集群、自行部署推理栈或训练模型**。其典型构成是“模型 + 推理运行时 + 工具链”三位一体：

- 模型目录**：自研模型（如 Gemini、Nova、Doubao、Qwen、ERNIE）+ 第三方托管模型；
- 托管推理**：多租户弹性推理、GPU/专用芯片调度、缓存、批处理，均由厂商负责；
- 配套工具链**：微调/蒸馏、RAG、Agent 编排、Guardrails（护栏）、评估与可观测、数据驻留选项。

与 IaaS / PaaS / SaaS 的本质区别：

层面	IaaS	PaaS	SaaS	
交付单元	裸计算/存储/网络（如 GPU 虚拟机）	应用运行平台（运行时、DB、消息）	模型能力 (API)	完整业务应用（含 UI 与 流程）
用户自己做 的事	搭全部软件栈	写并部署应用代码	集成模型到自己的 应用	直接使用功能
智能/推理	完全自理	自理或调用第三方	厂商全托管	厂商内置
计费	资源时长/容量	资源+服务	按 token / 吞吐	订阅/坐席

关键判断：MaaS 在技术栈上常被归入 "AI PaaS" 的一个子类，但它的**服务单元是"智能能力"而非"运行平台"**——用户消费的是推理结果，而不是一个可以自由部署代码的环境。与 SaaS 的区别在于 MaaS 交付的是**组件**，仍需开发者集成，不是一个开箱即用的成品应用（云厂商的 Copilot/Workspace 类产品才是 SaaS 层）。

三、主流 MaaS 交付与运营模式归纳

综合七家厂商的实践，可归纳为 **六种模式**，多数厂商是 "多模式叠加"：

模式 1：托管模型 API（预训练模型即服务）——最主流

将自研或第三方模型以标准化 API 开放，按 token 计费。

- **特点**：接入成本最低、迭代最快；模型作为 "商品" 明码标价。
- **适用**：所有需要快速上线的 LLM 应用、开发者、ISV。
- **商业价值**：模型本身成为高频流量入口，带动周边云服务消费（Gartner 指出生成式模型市场一年翻倍，正源于此）。
- **代表**：Azure OpenAI Service、Google Gemini API、AWS Bedrock（按模型计费）、阿里云 Model Studio、腾讯 Hunyuan API、火山方舟、百度千帆。

模式 2：多模型市场 / 模型目录（Model Garden / Marketplace / Catalog）

在同一平台上聚合自研与数十家第三方模型，统一 API 与安全管控。

- **特点**：厂商中立性、模型可替换、避免供应商锁定。
- **适用**：对模型选择有策略性考量的大中型企业、跨国多区域部署。
- **商业价值**：把 "模型竞争" 转化为 "平台粘性" —— 客户换模型不换云。
- **代表**：AWS Bedrock Marketplace（100+ 模型、18+ 供应商）、Google Model Garden（400+）、Azure AI Foundry Model Catalog、阿里云 Model Studio 第三方模型。

模式 3：模型微调与定制服务

在原厂模型基础上提供 LoRA/全参微调、蒸馏、持续预训练、知识库注入（RAG 自定义）。

- **特点**：将通用模型转为领域模型，兼顾 "能力" 与 "数据不出域"。
- **适用**：行业专属场景、私有知识密集型企业。
- **商业价值**：从 "卖 token" 升级为 "卖数据工程+模型工程"，客单价显著提升。
- **代表**：AWS Bedrock Fine-tuning/Distillation（Nova 定制与基础模型同价）、Vertex AI 微调、Azure Foundry 微调、火山方舟 "企业专有模型托管方案"。

模式 4：模型托管与推理平台（BYOM——自带模型）

让用户把自己训练/微调的模型（或开源模型）一键托管上线，由云厂商管推理扩缩。

- **特点：**打通"训练→部署→上线"闭环，是 MaaS 向传统 ML 平台（MLaaS）的延伸。
- **适用：**有自研模型能力、或对开源模型有深度定制需求的组织。
- **商业价值：**锁定高级开发者，拉动算力（GPU）消费。
- **代表：**AWS SageMaker + Bedrock 自定义模型部署、Vertex AI（Model Garden 开源模型一键部署）、Azure Foundry 模型目录部署、火山"企业自持模型托管"。

模式 5：Agent / 应用开发平台（MaaS 向上延伸）

在模型之上提供 Agent 编排、RAG、多工具调用、工作流与可观测，从"模型 API"走向"应用平台"。

- **特点：**是 2025–2026 年所有厂商的主战场；模型调用量由 Agent 生态指数级放大。
- **适用：**企业流程自动化、智能客服、Copilot 类产品。
- **商业价值：**把计费从"按 token"延伸到"按 Agent 会话/业务结果"，并带动数据、向量库、工作流等套餐消费。
- **代表：**Google Agent Platform / Gemini Enterprise Agent Platform、Azure AI Foundry + Copilot、AWS Bedrock Agents、腾讯 Agent 开发平台 3.0、百度千帆（已定位为 Agent 中心平台）。

模式 6：开源模型引流 + 云上付费托管

以开源权重建立生态与心智，再通过云上托管、企业版、托管服务变现。

- **特点：**这是中国厂商（阿里 Qwen、腾讯 Hunyuan 部分、字节 Seed 部分）与 Meta（Llama）的共性打法，也是 DeepSeek 走红后国际三大厂纷纷将其纳入目录的原因。
- **商业价值：**开源降低信任门槛与获客成本，把社区流量沉淀为云上付费推理。
- **代表：**Qwen（下载量与第三方部署量全球领先）→ 阿里云 Model Studio/Qwen Cloud；Llama → 三大厂目录均有。

四、逐厂商分析（海外市场）

4.1 AWS（Amazon Web Services）

核心产品与模式：

- **Amazon Bedrock**（核心 MaaS）：聚合 100+ 基础模型、30+ 区域、18+ 供应商，含 **Bedrock Marketplace** 扩展长尾与垂直模型；支持按需、批处理（约五折）、Provisioned Throughput（预留吞吐）三种计费。
- **自研模型：**Amazon **Nova**（Premier/Pro/Lite/Micro 文本与多模态）、Titan（已逐渐让位 Nova）；Nova 定制（微调/蒸馏）与基础模型同价，是 AWS 差异化卖点。
- **SageMaker**（MLaaS 底座）：训练、托管、推理，覆盖"自带模型"路线。

- **上层延伸**：Bedrock Agents、Knowledge Bases、Guardrails、Bedrock Studio，以及面向终端用户的 **Amazon Q**（SaaS 层）。

模型目录现状（截至 2026 年中）：Anthropic Claude（Opus 4.x/5、Sonnet 4.6/5）、OpenAI GPT 系列、Meta Llama 4、Mistral 全系、Cohere、NVIDIA Nemotron、DeepSeek、Qwen、GLM、MiniMax、Moonshot Kimi 等——**开放度在三大厂中最高、最"中立"**。

定价：按模型差异巨大（约 \$0.035-\$10/百万 token）。例：Claude Sonnet 5 上市促销 \$2/\$10，2026 年 8 月底后转标准价 \$3/\$15（输入/输出每百万 token）。提供缓存、批处理、预留吞吐等多档降本手段。

目标客户：中大型企业、ISV、有合规与数据主权诉求的行业客户（金融/医疗/政府）。

技术优势：基础设施深度（Nitro 虚拟化、Trainium/Inferentia 自研芯片、SageMaker 全栈）、多模型可替换性与成本优化工具链最成熟、安全管控（Guardrails/数据驻留/主权区域）完善。

区域部署：30+ 区域，覆盖北美、欧洲（法兰克福/伦敦/巴黎/斯德哥尔摩/苏黎世）、亚太（东京/首尔/新加坡/孟买/悉尼/雅加达）、中东与主权云（欧洲主权区、以色列主权区）。

海外策略判断：AWS 明确走"**多模型中立平台 + 算力与平台粘性**"路线，不与任何单一模型绑定（虽有约 80 亿美元投资 Anthropic 并作为其首选云），把模型竞争转化为对客户更有利的可替换性，同时以 Nova 自研模型与芯片摊薄推理成本。

4.2 Google Cloud

核心产品与模式：

- **Vertex AI**（核心 MaaS/AI PaaS）：内含 **Model Garden**（400+ 模型，自研 Gemini + 开源模型 Llama/Mistral/DeepSeek 等）、Gemini API（AI Studio 面向开发者）、Agent Builder / Agent Platform、AI Hypercomputer（TPU+GPU 一体算力）。
- **自研模型**：**Gemini** 家族（2.5 Flash-Lite → 3.1 Pro Preview 全谱系），以及**开源 Gemma** 作为开源心智补充。
- **特点**：模型与 TPU 算力、Google 搜索/工作区生态深度协同（Grounding、代码、Agent 能力原生整合）。

定价：分层定价（Standard/Priority/Flex），全球端点与区域端点不同价。典型价：Gemini 2.5 Flash-Lite \$0.10/\$0.40（业内最低一档），2.5 Flash \$0.30/\$2.50，2.5 Pro \$1.25/\$10（200K 上下文内，超长上下文有附加费），3.1 Pro Preview 约 \$2/\$12。批处理/Flex 另有折扣。

目标客户：重视多模态、搜索增强、Agent 能力的开发者与企业；对"Google 生态协同"（Workspace、Android、Search）有依赖的客户；价格敏感的创业团队（Flash-Lite 极低价获客）。

技术优势：**原生多模态与长上下文（1M）**、TPU 一体化算力（成本结构优）、Agent Platform 企业级成熟度、全球端点低延迟路由。

区域部署：全球多区域，提供数据驻留选项；但**部分新一代模型仅通过全球端点提供**，非全球端点（区域端点）在 Gemini 3 及以后家族 2026 年 7 月起采用差异化定价——这是其数据驻留与成本策略的关键细节。

海外策略判断：Google 走"**自研 Gemini 为核心 + 开源 Gemma + 开放模型目录**"路线，策略介于 AWS 的纯中立与微软的绑定之间；以**多模态/搜索/Agent** 的"**智能差异化**"而非单纯模型供给取胜，性价比定位明显（Flash-Lite 对标 GPT-nano）。

4.3 Microsoft Azure

核心产品与模式：

- **Azure AI Foundry**（统一平台，前 Azure AI Studio）+ **Azure OpenAI Service**（OpenAI 模型的云上独家渠道）+ **Model Catalog**（含 OpenAI、Meta、Mistral、Cohere、DeepSeek 及微软 **Phi** 开源模型）。
- **自研模型**：**Phi** 系列（开源小模型）、GPT-Chat 类入口；真正的旗舰能力来自与 OpenAI 的战略绑定。
- **上层延伸**：**Microsoft Copilot** 全家桶（M365/Windows/Developer）是 MaaS 之上最完整的 SaaS 变现层。

定价：OpenAI 模型价格体系即 Azure 主价格锚。GPT-5 \$1.25/\$10、GPT-5-mini \$0.25/\$2、GPT-5-nano \$0.05/\$0.40、GPT-5 Pro \$15/\$120（每百万 token）。提供 Global/Data Zone/Regional 三种部署形态，**PTU（预留吞吐单元）持续负载最高降约 70%**，另有 Batch API 折扣。

目标客户：企业客户与政府机构（合规认证数量行业第一）、依赖 Copilot 的微软存量客户、全球性跨国企业。

技术优势：**OpenAI 模型的云上独占/优先渠道**（GPT-5 等首发即上 Azure）、企业级安全与合规体系最全（欧盟数据边界、主权云、政府云）、Foundry 的 Model Router/Realtime/评估可观测工具链、Maia 自研芯片逐步落地。

区域部署：**60+ 区域，全球覆盖最广**，含欧盟数据边界、美国政府云、主权云选项；Data Zone 与 Regional 部署为数据驻留强需求客户设计。

海外策略判断：微软走"**OpenAI 能力 + Azure 合规底座**"的绑定路线，是三者中模型集中度最高（OpenAI 占绝对主导）、但合规与 Copilot 生态差异化最强的厂商；对"模型即入口、平台即粘性"的依赖最深。

4.4 阿里云（Alibaba Cloud）

核心产品与模式：

- **Model Studio（国际版"百炼"）**：一站式模型服务平台，提供 Qwen 全系 + 第三方模型，OpenAI 兼容 API，多模态覆盖文本/图像/音频/视频（含 Wan 视频模型）。
- **Qwen Cloud**：2025 年 5 月于新加坡全球发布，主打"免信用卡、一个网页从免费试玩到一键取 Key"的低摩擦开发者体验，是阿里 MaaS 出海的核心动作。
- **开源策略**：Qwen 开源家族（Qwen2.5/3、Qwen3.7 等）是全球下载与第三方部署量最大的中国模型，走典型的"开源引流 + 云上托管变现"路线。

定价：明显低于国际同级。当前主线：Qwen3.7-Max \$2.5/\$7.5、Qwen3.7-Plus 更低（每百万 token），多模态与视频按秒/按 token 计费；配合免费额度与新用户优惠。

目标客户：出海中国企业、东南亚与中东开发者、对价格敏感且需 OpenAI 兼容低成本迁移的全球化用户。

技术优势：Qwen 系列的中英双语与长上下文能力、Wan 视频生成、**OpenAI 兼容 API 的零迁移成本**（改 Key 和 base URL 即可）、多区域就近接入。

区域部署：Model Studio 已覆盖 **新加坡、美国（弗吉尼亚）、中国香港、日本（东京）、德国（法兰克福）及中国大陆（北京）**；以亚太为中心向外辐射。

海外策略判断：以"性价比 + 开源生态 + 迁移兼容"三件套切入，优先深耕东南亚；与腾讯、字节在海外最大的区别是其自研开源模型生态（Qwen）已具备全球品牌，可依托模型心智倒灌云服务收入。

4.5 腾讯云（Tencent Cloud）

核心产品与模式：

- **Hunyuan（混元）系列 API：**覆盖文本（Hunyuan-Large/Turbo/A13B 等）、翻译（30+ 语言）、图像、视频、**3D 生成（Hunyuan 3D Engine，2025 年 11 月全球发布）**；一年内开源/发布 30+ 模型。
- **Agent 开发平台 3.0：**2025 年 9 月随全球数字生态峰会发布，面向国际市场，主打"场景化 AI + SaaS"组合。
- **Hunyuan 3D 是差异化亮点：**从文本/图像/草图生成商业级 3D 资产，面向游戏、电商、影视特效、3D 打印；3D 模型 API 已上腾讯云国际站，并给创作者每日 20 次免费生成额度。

定价：整体低价策略 + 免费额度；文本模型价格与国际头部对齐、多模态（尤其 3D/视频）按量计费。

目标客户：出海中国企业、亚太企业（游戏/社交/电商/广告营销）、内容创作者与开发者。

技术优势：3D 生成与多模态创作、多语言翻译（30+ 语言）、依托腾讯游戏/社交/微信生态的场景理解与数据。

区域部署：腾讯云国际站 Hunyuan API 覆盖 **新加坡、雅加达、曼谷、首尔、东京、香港、弗吉尼亚、硅谷、圣保罗、法兰克福**——海外区域面是中国厂商中最宽的之一；官方披露海外客户数一年翻倍、连续三年高双位数增长。

海外策略判断：以"场景化（游戏/社交/媒体）AI + 开放开源 + 亚太深耕"切入，打法上比阿里更"垂直场景导向"；海外增长主要靠亚太而非欧美。

4.6 字节跳动 · 火山引擎（Volcano Engine）

核心产品与模式：

- **Doubao（豆包）模型家族：**2025 年迭代至 1.6，2026 年 2 月发布 2.0 重大升级，2026 年 6 月发布 **2.1 Pro**（Coding/Agent/VLM 三大方向强化），已通过火山引擎 API 开放；另有 **Seedance** 视频生成（2.0/2.5）、**Seedream** 图像（5.0 Pro）、Seed Speech 语音（含克隆）等。
- **火山方舟（Volcano Ark）：**国内核心 MaaS/Agent 平台；配套推出"企业自持模型托管方案"、Responses API、Doubao 搜索开放服务（面向 Agent 长任务）。
- **海外路径：**主要通过 **BytePlus**（字节国际 B2B 品牌，新加坡等区域）提供 Seed 系列 API；火山引擎本身海外公有云部署有限，更多是"能力出海"而非"云平台出海"。

定价：字节是国产大模型价格战发起者（曾大幅降价并给高频免费额度），海外经 BytePlus 的定价同样激进，TTS/图像/视频按调用与时长计费。

目标客户：出海企业（尤其短视频/直播/电商/内容创作赛道）、在 TikTok/抖音生态内做 AI 应用的开发者。

技术优势：视频生成（Seedance）与语音克隆在同类中领先、Agent 型模型的编码与工具调用能力、字节内容生态的海量场景数据。

区域部署：海外经 BytePlus（以新加坡 ap-southeast-1 为核心）与全球边缘节点；火山引擎国际数据中心覆盖有限，尚未形成与阿里/腾讯可比的多区域云网。

海外策略判断：字节的出海策略是**"能力借道生态出海"**——不主打"中国云"，而是通过 BytePlus 与 TikTok 生态把模型能力嵌入全球内容产业；对**内容生成场景**的竞争力最强，对"通用企业云 MaaS"的威胁目前弱于阿里/腾讯。

4.7 百度智能云（Baidu AI Cloud）

核心产品与模式：

- **千帆（Qianfan）平台：**百度官方 MaaS 平台，2026 年已明确定位为 **Agent 中心的企业级模型与应用平台**（多智能体协作、RAG、工作流、模型服务、MCP 组件）。
- **ERNIE 系列：**ERNIE 4.5 / X1（推理模型）、**ERNIE 5.0**（2.4T 参数统一多模态底座）、ERNIE 5.1（Agent/推理/创作强化）等持续迭代；ERNIE Bot（文小言）为 C 端入口。
- **开源：**ERNIE 4.5 等部分开源，依托 PaddlePaddle 生态。

定价：极具攻击性的低价。ERNIE 4.5 输入约 **\$0.55/百万 token、输出 \$2.2**（即人民币 ¥0.004/千 token 级别）起，ERNIE X1 更低。

目标客户：主要为中国大陆企业与开发者；海外几乎无独立公有云部署。

技术优势：ERNIE 中文理解与搜索增强、PaddlePaddle 全栈、Agent 平台化；但模型性能在国际基准上已不具备对 OpenAI/Gemini 的代差优势。

区域部署：**海外无公有云区域。**海外可用性仅停留在英文文档/博客叙事与零散 API 接入，不构成体系化的 MaaS 出海。

海外策略判断：百度是七家中**海外 MaaS 缺位最明显**的厂商——其核心市场与合规重心均在中国大陆，海外策略实质是**"跟随式 API 供给"**而非**"平台出海"**。对海外使用者而言，其价值主要体现在**低成本中文能力**与 PaddlePaddle 生态，而不是**"在海外用 ERNIE"**。

五、横向对比

5.1 平台与产品维度

维度	AWS	Google Cloud	Microsoft Azure	阿里云	腾讯云	火山引擎	百度
核心 MaaS 平台	Amazon Bedrock	Vertex AI	Azure AI Foundry	Model Studio/ Qwen Cloud	Hunyuan API	火山方舟（海外走 BytePlus）	千帆
自研旗舰模型	Nova	Gemini / Gemma	Phi	Qwen	Hunyuan	Doubao / Seed	ERNIE
模型目录开放度	最高（100+/18 供应商）	高（400+, Model Garden）	高（但 OpenAI 主导）	中（Qwen 为主 + 少量第三方）	中（自研为主）	低（自研为主）	低（自研为主）

维度	AWS	Google Cloud	Microsoft Azure	阿里云	腾讯云	火山引擎	百度
第三方/开源模型	Claude、GPT、Llama、Mistral、DeepSeek、Qwen、Kimi...	Llama、Mistral、DeepSeek、Gemma	GPT、Phi、Llama、Mistral、DeepSeek	MiniMax 等少数	少量	少	少
旗舰价格参考 (\$/百万 token，入/出)	Claude Sonnet 5: \$3/\$15	Gemini 3.1 Pro Preview: \$2/\$12	GPT-5: \$1.25/\$10	Qwen3.7-Max: \$2.5/\$7.5	低价对齐	低价+免费额度	ERNIE 4.5: \$0.55/\$2.2
最低端模型价格	\$0.035 起	Flash-Lite \$0.10/\$0.40	GPT-5-nano \$0.05/\$0.40	Qwen 轻量版更低	低价	低价	极低
特色能力	微调/蒸馏、Guardrails、主权区	长上下文 1M、多模态、Agent	Copilot 全家桶、合规最强	开源生态、Wan 视频	3D 生成 、翻译	视频/语音生成	中文、Paddle

5.2 生态、合规与海外策略维度

维度	AWS	Google Cloud	Microsoft Azure	阿里云	腾讯云	火山引擎	百度
开源策略	不主导开源，托管开源	Gemma 开源	Phi 开源	Qwen 开源（全球最大中国模型生态）	Hunyuan 部分开源	Seed 部分开源	ERNIE 部分开源
海外区域覆盖	30+ 区域	全球多区域（部分模型仅全球端点）	60+ 区域最广	新加坡/美东/香港/东京/法兰克福	新加坡/东京/首尔/雅加达/美欧等 10 区	有限（BytePlus 为主）	无海外区域
安全合规	认证全、主权云	认证全、数据驻留	认证数量最多、数据边界	出海合规逐步补齐	亚太合规为主	内容合规为主	中国大陆合规为主
海外市场策略	多模型中立平台	自研智能差异化	OpenAI 绑定+合规底座	开源引流+性价比	场景化（游戏/社交/3D）+亚太	生态借道（TikTok/BytePlus）	基本缺位
对出海客户的适配度	全球原生	全球原生	全球原生	高（迁移兼容）	高（亚太）	高（内容场景）	低
主要风险点	模型价格战挤压利润	区域端点/定价复杂性	OpenAI 依赖集中	欧美合规与数据信任	欧美市场渗透浅	平台出海未成体系	海外不可用

六、总结与选型建议

6.1 五个关键趋势

- MaaS 竞争重心从"模型"转向"平台与 Agent 生态"。**模型能力趋同、价格快速下探，2026 年的主战场是 Agent 平台、工具链、可观测与数据协同——这决定了各厂商的差异化不是"谁的模型强"而是"谁的生态转化率高"。
- 价格两极分化：**国际大厂"旗舰高价 + 端侧小模型低价"（GPT-nano、Flash-Lite 低至 \$0.05-\$0.10/百万 token），中国厂商以系统性低价+免费额度切入，成本结构优势明显（自研芯片、开源生态摊薄）。
- 开放度成为竞争维度：**AWS 的中立目录、Google 的开放 Garden、以及 Qwen/DeepSeek/Llama 的开源路线，都在对冲"模型绑定"焦虑；"模型可替换"成为企业采购的默认要求。
- 中国厂商出海分化：**阿里（开源心智+平台化）与腾讯（场景化+区域广度）已形成体系化出海；字节以内容生成能力借道生态；百度基本未出海。中国厂商的共同短板是欧美市场的数据信任与合规信任。
- 合规与数据驻留成为高端市场准入门槛：**Azure 数据边界、AWS 主权区、Google 区域端点、阿里/腾讯的新加坡/法兰克福节点，都在服务同一诉求——"模型可以国际化，数据必须本地化"。

6.2 竞争格局

- **第一梯队（全球原生）：**Azure（OpenAI 独占+合规）、AWS（中立+基础设施最深）、Google（智能+生态协同），三者垄断欧美高端市场。
- **第二梯队（亚太/出海）：**阿里（Qwen 生态+性价比）、腾讯（场景化+区域广）。
- **第三梯队（能力型）：**火山引擎（内容生成垂直最强）、百度（海外缺位）。

6.3 选型建议（按使用者视角）

使用者画像	推荐	理由
欧美合规要求高的中大型企业	Azure（数据边界/认证）或 AWS（主权区）	合规即入场券，其次是 Copilot/Agent 生态
需要多模态、搜索、长上下文、Agent 深度的开发团队	Google Vertex AI	Gemini 原生多模态 + 1M 上下文 + Agent Platform
追求模型可替换、规避绑定的架构决策者	AWS Bedrock	目录最全、Guardrails/微调工具链最成熟
出海东南亚/中东、价格敏感、需 OpenAI 兼容零迁移	阿里云 Model Studio	Qwen 性能/价格比优，Key 与 base URL 即迁
游戏/电商/3D/营销内容的创作团队	腾讯 Hunyuan 3D	3D 与多语言创作差异化最强
短视频/直播/音频内容出海	字节 BytePlus（Seed 系列）	视频/语音生成领先，紧贴内容产业
预算极低的国内中文场景	百度千帆	ERNIE 价格最低、中文可用

最后一句实操提醒：2026 年的 MaaS 采购已不再是"选一个模型"，而是"选一个能让我低成本换模型、数据合规落地、且 Agent 化后可规模化的平台"——把 **模型可替换性、数据驻留能力、Agent 工具链成熟度** 三者作为评分卡的核心维度，比单比某一家模型跑分更有决策价值。

