智谱、MiniMax20260903

智谱与 MiniMax ARR 爆发式增长，国产算力进入经济性验证期

摘要

- 智谱 AI 与 MiniMax 的 MaaS 业务均呈爆发式增长，智谱 8 月底 ARR 达 16-20 亿美金，MiniMax 8 月 ARR 达 8 亿美金且 B 端贡献超 80%。
- 智谱 AI 毛利率由负转正至 26.5%，得益于规模效应及模型迭代使推理成本降低 80%，预计未来 1-1.5 年毛利率有望翻倍。
- MiniMax 收入结构向 MaaS 转型，占比升至 63%，虽因模型切换及补偿计划导致毛利率降至 18%，但随 M3Pro 定价提升预计将回升。
- 国产算力进入经济性验证期，智谱“算力乘数”同比提升 14 倍，计划年底实现 1 元算力投入产出 1 元 MaaS 收入。
- 模型演进从 Coding 向 Agent 及 Cowork 模式进化，智谱 GLM-6 与 MiniMax M3Pro 均瞄准 GPT-4 水平，强化自进化与长任务能力。
- 投资侧看好国产算力产业链（超节点、GPU、交换芯片）及 AI 应用（网安、办公软件如金蝶、迈富时等）的商业化落地。

Q&A

2026 年上半年智谱的收入构成是怎样的，其中 MaaS 业务的占比和增速如何？

2026 年上半年，智谱的收入结构已完全以 MaaS 业务为主。MaaS 业务，即开放平台及 API 服务，涵盖了 API 调用付费和 Coding Plan 订阅模式，其收入占比达到 86.5%，约为 8.3 亿人民币。另一部分独立部署业务的收入占比为 13.5%，与 2026 年年初相比，该业务在战略和收入增速上均呈现收缩趋势。因此，MaaS 业务已成为公司当前主要的收入来源，并预计将保持高速增长。

智谱近期的 ARR（年度经常性收入）增长趋势如何，尤其是在 2026 年 7 月至 8 月期间，其增长驱动因素是什么？

智谱的 ARR 呈现出显著的加速增长态势。公司于 2026 年 7 月 10 日左右披露的 ARR 为 10 亿美金，而至 8 月底，该数字已增至 16 亿美金，即在一个半月内增长

了 6 亿美金。此 ARR 计算口径为月均值乘以 12。若采用更激进的周均值乘以 52 的口径（类似 MiniMax 的计算方式），8 月底的 ARR 估算可达 20 亿美金。2026 年 7 月至 8 月 ARR 的大幅增长主要由以下几个因素驱动：首先，6 月 15 日发布的 GLM-5.2 模型和 8 月初发布的 GLM-5.3 模型开始放量；其次，公司在 8 月初开放了 Coding Plan 的限售，并同步大幅提升了订阅价格，这对 ARR 增长贡献显著。

智谱 2026 年上半年的收入与其高额的 ARR 之间存在较大差距，应如何理解这一现象？

2026 年上半年 8 亿多的收入与 7、8 月份高达 10 亿至 16 亿美金的 ARR 之间的差距，主要是由于 ARR 的爆发性增长集中在 6 月底之后。截至 2026 年 6 月底，公司的 ARR 水平尚不高，若参考 Anthropic 约 4:1 的 ARR 与收入比例进行倒推，当时智谱的 ARR 可能约为 5 亿多美金。进入 7 月份后，ARR 迅速增长至 10 亿美金，并在 8 月进一步攀升，这种增长的陡峭曲线导致了上半年收入与期末 ARR 数值上的显著差异。

从 Token 调用量、用户增长和付费结构来看，智谱的业务增长呈现出哪些具体特征？

智谱的业务增长呈现出量价齐升和显著的头部客户效应。在 Token 调用量方面，相较于 2026 年初，总调用量增长了 40 倍，其中 Coding Plan 的调用量增长了 23 倍。在价格方面，API 调用模式下的 Token 售价同比增长了 100%，Coding Plan 的订阅价格也在 8 月初有可观提升。在用户增长方面，截至 8 月，平台注册用户数从年初的 300 万增至 740 万，增长 144%，而付费日活用户数则增长了 6 倍，显示付费转化率大幅提升。头部客户效应尤为突出，前十大用户的日均 Token 调用量相比年初增长了 98 倍。

智谱的高质量用户群体对 ARR 的贡献情况如何，其客户结构呈现怎样的分布？

智谱的高质量用户群体（定义为 ARR 大于 10 万美元的用户）呈现出典型的头部集中效应。截至 2026 年 8 月，在总计 740 万注册用户中，仅有 189 个高质量客户，但这部分客户贡献了当月 16 亿美金 ARR 中的 7 亿美金。具体来看，这 189 个客户的分布如下：ARR 大于 10 万美元的用户有 115 个，大于 50 万美元的有 25 个，大于 100 万美元的有 37 个，大于 1,000 万美元的有 8 个，大于 2,500 万美元的有 2 个，以及 ARR 达到 2.5 亿美元的客户有 2 个。这种 1%的用户贡献超过 80% ARR 的现象在行业内是常见的，显示出深度使用 AI 的头部企业具有极高的价值贡献。

人在草木间

智谱 2026 上半年的毛利率水平如何，其提升的驱动因素是什么，以及未来的改善空间在哪里？

2026 年上半年，智谱的整体毛利率为 26.5%。其中，占收入 86.5%的开放平台及 API 业务毛利率为 24.6%，相较于 2025 年同期的负毛利实现了显著提升，对比 2026 年初则提升了约 6 个百分点。毛利率提升主要得益于三方面：一是规模效应摊薄了 Token 调用成本；二是模型能力提升带动了产品定价的上涨；三是各类工程优化手段降低了推理成本。 未来毛利率仍有持续提升空间，预计在未来一到一年半内有望在现有基础上翻倍。主要的优化方向包括：第一，提升集群利用率和规模效应，特别是期待年底如升腾 950 DT 等国产算力形成规模化大集群后，能进一步提高效率；第二，优化新旧模型迭代带来的资源冗余和迁移成本，目前 GLM-5 系列模型的成本相比 2 月份发布时已降低 80%，例如通过优化 harness 将命中率从 90%提升至 98%；第三，通过更智能的路由策略，将更多任务分配给性价比更高的轻量化模型执行，从而优化整体成本结构。

智谱规划的商业模式演进路径是怎样的？

公司的商业模式规划了一个从 Chatbot 到自主 AI 的演进路径。在经历了初期的 Chatbot 阶段后，当前正处于从 Coding 向 Agent，再到 Cowork 的模式进化过程中。这一系列模式的逐步演进，最终目标是实现能够自主运行的 AI 模式。

智谱 AI 在商业模式、算力部署和模型迭代方面有哪些最新的战略规划和进展？

智谱 AI 的商业模式正从以 API 调用为主的 Coding 模式，逐步向订阅制为主的 Agent 模式，以及未来可能以结果付费的 Cowork 模式演进。其中，Coding 是基础能力，而 Cowork 模式被视为价值量最高的形态，它涉及多种 Agent 交互以交付工程成果。公司正选择在网络安全、法律、金融和数据分析等高价值、成果可验证且具备高智力门槛的场景进行探索。目前，网络安全被认为是最清晰的验证方向，因为其任务流程，如漏洞发现、工具调用验证和多步攻防等，能充分利用模型在代码理解、异常发现和环境反馈调整等方面的复杂底层能力。近期 Anthropic 的 Claude 3.5 Sonnet 和 OpenAI 的 GPT-4o 均显著增强了网络安全能力，OpenAI 还推出了 GPT-4o Cyber 产品，这都印证了该趋势。在地域扩张方面，智谱 AI 正积极与海外 CSP 平台合作，计划通过开源模型或收入分成的方式进入海外市场，预计短期内将有具体进展。 在算力层面，公司采用自有集群、租用云服务和算力租赁三种方式，以租赁为主。为保障大规模模型训练所需的同构大集群稳定性，公司计划加大自有大集群的建设投入，这可能与 7 月份的定向增发有关。算力资源的使用上，训练的优先级高于推理，因为基础模型的智力上限决定了商业化的天花板。自 2026 年初，公司全面实施“All in infra”战略，将国产芯片纳入主力推理算力。近期，GLM-5.3 及 Oxsyrhynchus 模型在国产芯片集群上成功承载了 62 万亿 token 的超大规模流量，并在 6 天内实现了约 10 万卡级的国产芯片规模化低成本推理，标志着已进入验证国产卡经济性与性价比的阶段。公司提出了“算力乘数”这一新指标，即每投入一元算力（含训练与推理）

所产生的 MaaS 收入。该指标在 2026 年年中相较 2025 年同期提升了 14 倍，主要得益于单位 token 承载的智能提升和单位任务消耗的 token 成本下降。公司期望在一年左右时间内，实现每投入 1 元算力即可获得 1 元 MaaS 收入。 关于模型迭代，公司基本保持两到两个半月一次的升级周期，例如 GLM-5 系列从 2 月的 5.0 版本迭代至 8 月的 5.3 版本，预计 GLM-6 大版本将在 2026 年内发布。下一代模型的核心思路将围绕“自进化”能力，即模型能自主判断停止时机并进行自我纠正。在 GLM-5 这个 700B 参数级别的模型上，公司已通过极致的后训练技术显著提升了 coding 效果。未来，GLM-6 将继续提升基础模型的参数上限，这需要实现数据规模的扩展（data scaling），并结合更优的后训练技术，针对网络安全、法律、金融等高价值场景进行特定强化学习和优化，目标是使 GLM-6 的性能逼近 GPT-4 的水平。

MiniMax 在 2026 年的收入结构、ARR 增长驱动因素以及毛利率变化情况是怎样的？

MiniMax 的收入结构发生了显著变化，从过去以 AI 原生产品（如星野、Talkie、海螺等应用）为主要收入来源，转变为以 MaaS 服务为主。2026 年，其 MaaS 服务收入达到 7,400 万美元，同比增长 700%，在总收入中的占比从 2025 年的 30%提升至 63%。由于其 AI 原生产品在海外市场表现出色，公司海外市场贡献的收入占比超过六成。AI 原生产品目前的收入占比约为 30%多，同比也实现了翻倍增长。 公司的 ARR（年度经常性收入）增长迅速，于 2026 年 8 月达到 8 亿美元，相较于 5 月份的 4 亿美元实现了翻倍。这一增长主要由多个因素共同驱动：首先，6 月 1 日发布的新一代大语言模型 M3，在经历了初期的爬升期后，于 7 月和 8 月显著加速了 ARR 的增长；其次，8 月 1 日开始贡献增量的多模态模型 H3，因其优异性能（上线三周下载量超 2,400 万次，衍生出 300 个公开模型）和实时视频生成等创新应用，成为新的增长点；再次，办公场景应用的整体提升也带动了 ARR 的增长；最后，8 月份 DeepSeek 宣布涨价，使得 MiniMax 模型的性价比优势凸显，进一步提升了调用量。 业务重心的转移也体现在客户结构上，目前 B 端客户贡献的 ARR 已超过 80%，而 2025 年同期仅为 30%。其企业客户与开发者数量已超过 200 万，相较 2025 年底增长了 10 倍，这主要得益于 M3 模型升级后 coding 能力的提升。2026 年 7 月的 token 消耗量达到了同年 1 月份的 20 倍。 在毛利率方面，MiniMax 的整体毛利率从年初的 25.4%下降至 18%。这主要是由收入结构变化导致的：一方面，MaaS 服务收入占比从 30%大幅提升至 63%，而在 MaaS 服务中，增速更快的大语言模型相较于多模态模型毛利率更低；另一方面，M3 模型推出初期因准备不足进行了一些 token 补偿计划，也对毛利率产生了一定的负面影响。

MiniMax 在 2026 年下半年及 2027 年的毛利率趋势如何？其研发支出与毛利增长之间呈现何种关系？

预计公司在 2026 年下半年及 2027 年的毛利率将持续提升。此前毛利率受到 M3 模型推出的一次性影响，但在七八月份公司运营回归正轨后，毛利率已有所回升。

未来的毛利率增长驱动因素包括：下一代模型（特别是 M3 Pro 版本）将采用更高的 token 定价；同时，预计下半年研发支出的增速将低于毛利的增速。

MiniMax 的算力构成是怎样的？其训练效率和现有算力储备能否支持未来的模型规划？

MiniMax 的算力主要来源于三个方面：自建算力、云服务以及通过“token 工厂”模式进行的算力租赁。公司在算力运用方面的一个显著优势是其高达 97%的有效训练时间占比，这在行业内处于第一梯队水平。根据现有评估，公司目前掌握的算力资源足以支撑后续规划的三款新模型的研发与训练需求。

MiniMax 在国产芯片的应用现状及未来规划是怎样的？公司如何评估国产芯片的效用？

MiniMax 已经部分采用了国产芯片。公司预计到 2026 年第四季度，国产芯片的使用占比将进一步提升，并期望届时国产芯片能够在降低成本方面发挥作用，即提升性价比。根据公司的观察，采用国产芯片有助于提高整体算力利用率并降低单位 token 成本，因此被视为一个良好的选择。

MiniMax 接下来的新模型发布计划是怎样的？各款模型的定位和技术特点分别是什么？

MiniMax 计划推出三款新模型。第一款是 M3.1，可视为 M3 的后训练版本，其主要目标是验证在 M3 上构建的全链路基础设施的成熟度，并提升模型的整体稳定性与推理效率。第二款是参数量接近 3T 的大尺寸旗舰模型 M3 Pro，该模型在预训练数据量和长任务后训练能力上将力求达到当前技术可实现的上限。第三款是 H3.1，它是在 H3 多模态模型基础上的体验优化版本。目前，M3.1 和 H3.1 的研发已接近完成，而 M3 Pro 仍在研发过程中。

MiniMax 在文本模型和多模态模型之间的战略优先级是如何定位的？公司对开源策略和成本效益有何考量？

尽管公司自成立之初便定位于全模态发展，但在当前资源有限的情况下，文本模型的研发优先级高于多模态模型。从长远来看，公司认为两类模型的发展是协同而非资源竞争关系，未来的跨界融合能力将催生更优的智能。公司始终将追求极致性价比作为核心目标之一，认为降低 token 成本能为后续的研发迭代（例如用更少资金进行更多轮次的训练）创造更大空间。同时，公司将延续开源开放策略，认为这是中国模型发展的首选路径。

综合来看，当前大模型行业在商业化、算力部署和模型迭代方面呈现出哪些核心趋势？

行业呈现出几个核心趋势。在商业化层面，MaaS 服务的毛利率普遍存在较大提升空间，这主要依赖于 infra 能力的增强；同时，下一代大尺寸模型的 token 价格预计将继续提升。办公场景正成为带动 ARR 和 token 消耗量增长的重要领域，这得益于模型在长任务和自治智能体方向的迭代，满足了协同办公场景的需求。在算力部署层面，自建算力的重要性日益凸显，尤其对于训练更大参数量的模型而言，上市融资为模型公司加大自建算力投入提供了条件；同时，国产算力已进入经济性验证阶段，预计到 2026 年年底其性价比将有显著提升，有助于国产大模型降本增效。在模型迭代层面，行业普遍坚持开源策略，并持续聚焦于长程任务、自治智能体系统的研发，最终方向是实现基于 RLHF 的自我训练与自我进化。现阶段，大语言模型因其明确的商业化路径，仍是各方投入的重点。

基于当前大模型行业的发展趋势，下半年的主要投资机会分布在哪些领域？

2026 年下半年，中国的开源模型在全球范围内具备显著的投资价值，其在 to B 和 to C、多模态和大语言模型等领域的商业化发展均呈现出多元且充分的亮点。商业化的蓬勃发展将催生对算力的巨大需求，使国产算力扮演至关重要的角色。因此，国产算力产业链值得高度关注，包括超节点、国产 GPU，以及预计在 2027 年随着超节点使用量提升而需求增加的国产交换芯片。相关标的包括能提供超节点整机的厂商，如浪潮、华勤等。在应用侧，随着模型长程任务能力的提升，网络安全公司的能力也得到增强，构成利好。办公应用的市场接受度正迅速提高，云厂商及互联网公司的推广投入将进一步激发市场关注。此外，企业自身的 Agent 应用也为 A 股市场的企业办公软件公司（如迈富时、金蝶、鼎捷等）提供了良好的发展机遇，其 AI 产品的功能和易用性正不断改善。