智谱 20260831

智谱 AI 2026H1 研报：ARR 达 16 亿美元，算力乘数与毛利双升

摘要

- 2026 年上半年营收 9.54 亿元，API 业务占比超 86%，同比增 27 倍；毛利率 26.4%，较 2025 年提升 6 个百分点，目标 12-18 个月内升至 50%以上。
- ARR（年度经常性收入）实现脉冲式增长，从 2026 年初的 2.5 亿美元增至 8 月的 16 亿美元，预计年底指引达 24 亿美元，规模化渗透速度显著优于海外对标公司。
- 算力乘数（每元算力投入产出比）年初至今提升 14 倍，单位 Token 推理成本下降 80%，预计未来 6-12 个月内达到 1.0 的盈亏平衡临界点。
- GLM5.3 系列实现“双帕累托最优”：5.3 模型主攻高价值长程任务，5.3Flash 通过稀疏+线性注意力架构优化成本，AA 指数达 57.5 分，卡位普惠场景。
- MaaS 平台注册用户超 740 万，付费日活增长 6 倍，前十大客户日均调用量较年初增长 98 倍，显示模型在复杂工作流中的用户粘性与使用深度增强。
- 商业模式从 DtoB 转向 BtoD/BtoC，发力 Cowork 数字员工场景，测算全球潜在市场空间达 5 万亿，模型能力已跨越“可用线”进入真实办公流。
- 研发投入 21.3 亿元，超过净亏损额，亏损率同比收窄 3.5 倍；通过自建超大规模国产算力集群与 Infra 全栈优化，实现算力供给与模型迭代的良性循环。

Q&A

请介绍一下公司 GLM 模型家族在 2026 年上半年的迭代历程、各版本的核心特点以及其在产品版图中的战略定位？

2026 年以来，GLM 模型家族在 8 个月内推进了 5 代。GLM 是公司第一代从 Vibe Coding 到 Agentic Engineering 实现跨范式意义的基座模型，其 4.5、4.
系列将模型参数规模从约 3,000 亿扩展至 744B，为上半年奠定了基础。
公司在 4 月和 6 月分别推出了主攻长程任务（long horizon task）的 GL

人在草木间

和 GLM 5.2。其中，GLM 5.2 是首个进入全球前三的国产模型，并在特定背景下取得了可用模型的全球第一。GLM 5.3 则是在 744B 基座上进行的一次受控实验，通过极致的后训练，其前沿编程能力相比 5.2 实现了代际提升，并涌现出网络安全防御能力。近期推出的 GLM 5.3 Flash 是一款约 300B 参数的模型，它不仅具备前沿编程能力，其架构也是面向下一代模型的一次实验。在产品版图中，GLM 5.3 和 5.3 Flash 分别占据了高价值任务模型和兼顾智能与成本的帕累托最优前沿两个关键位置，使公司整体模型与产品架构更为稳健。

公司 MaaS 平台的业务表现和市场进入策略在 2026 年上半年有何变化？

2026 年上半年，MaaS 平台业务实现了超出预期的历史性放量。在 API 调用价格提升约 101%的基础上，tokens 调用量较年初增长了 40 倍，其中编程套餐产品（coding plan）的增长超过 23 倍。市场进入策略（go to market strategy）可以概括为“更深、更宽、更均衡”。首先，在产品范式上，公司在维持编程领域技术领先的同时，开始发力协同工作，目标是面向真实环境和复杂工作流的数字员工。其次，在客户拓展模式上，从早期由开发者推动企业采购的“D to B”（developer to business）模式，转向企业采购后为软件开发、研发、数据、产品、网络安全乃至运维等多个条线部署的“B to D”或“B to C”模式。最后，在价值定位上，公司从专注前沿智能，开始通过推出 GLM 5.3 Flash 等产品进入普惠场景，以期在更广泛的场景中进行尝试。

公司编程套餐产品（coding plan）自上线以来的增长情况如何？

公司的 coding plan 产品自上线至今恰好一年（365 天），实现了 234 倍的同比增长，已成为全球规模最大、增速最快的编程服务之一。

2026 年上半年公司云端平台的收入结构和年度化收入增长情况是怎样的？

公司的云端平台收入结构日益均衡。在传统 To B 企业级 API 调用业务持续贡献收入的同时，coding plan 业务也成为重要增长点。2026 年 2 月推出 GLM 5 时，coding plan 因算力限制曾一度售罄。随着算力基础设施的改善，该产品于 7 月 31 日后开放限购，产品布局趋于均衡。从年度化收入增长来看，以 2025 年 12 月 31 日为基准，推出 GLM 5.1 后（2026 年 4 月），整体收入增长了 4 倍；推出 GLM 5.3 后增长了 10 倍；至 2026 年 8 月推出 5.3 Flash 时，整体 AR 增长已达到 46 倍。

公司在开发者生态建设方面取得了哪些进展？

在生态层面，除了 GLM 模型家族，公司于 2026 年 6 月推出了再扣的 3.0。其 Hannes 工具和 IDE 工具在 Code 平台推出后，仅一个月用户数就突破了 100 万。这一系列布局体现了公司在行业竞争深化背景下，通过从 Coding 到 Cowork、B to B 到 B to C、前沿智能到普惠场景、API 到 API 加 coding plan、再到 LM 加 harness

的矩阵式扩张，实现了更均衡的生态卡位和产品配置。

公司 MaaS 平台的注册用户规模、付费用户活跃度以及头部客户的调用量情况如何？

目前，智谱 MaaS 平台已成为中国最大的开发者平台之一，注册用户数超过 740 万，2026 年至今增速达到 144%。平台更看重的付费日活跃用户数较年初整体增长了 6 倍，显示出高质量用户占比和增速均在提升。从用户使用深度来看，以收入规模计算的前十大用户，其最新的日均调用量较年初增长了 98 倍（准确值为 97.98%），反映出模型在长程任务时代的用户粘性与使用深度显著增强。

公司在头部客户和高价值客户（Premium Customer）方面有哪些新的突破？

在头部客户方面，继 2026 年 3 月披露中国 Top 10 互联网巨头中有 9 家是公司客户后，在过去的半年里，这 9 家客户中有 4 家已将 GLM 模型作为其全球首要选择。在高价值客户方面，截至 2026 年 8 月末，按年化收入口径统计，超过 10 万、50 万、100 万、1,000 万、2,500 万美元的客户数量分别为 115 个、25 个、37 个、8 个和 2 个。与年初相比，各档位的客户数量分别增加了 94 个、20 个、35 个、7 个和 2 个。

公司 2026 年上半年的整体财务表现如何，特别是在收入、毛利率及业务结构方面？

2026 年上半年，公司整体营收为 9.54 亿元，毛利率为 26.4%。其中，开放平台和 API 业务收入达到 8.25 亿元，同比增长 2,736%（约 27 倍），占总收入比例超过 86%，标志着收入结构的彻底转变。该业务板块的毛利率为 24.6%，较 2025 年全年的 18.9%提升了近 6 个百分点。

公司如何衡量算力投入的效率，其表现如何？

公司使用“算力乘数”这一核心数据来衡量算力投入效率，即每投入 1 元算力（包括训练与推理）所对应的 MaaS 平台业务收入。2026 年至今，该乘数整体提升了 14 倍。

公司在 2026 年上半年的财务表现如何，特别是在研发投入与净亏损方面有哪些关键变化？

2026 年上半年，研发投入的绝对金额扩大至 21.3 亿人民币。经调整净亏损额为 19.64 亿，较 2025 年同期上升约 12%。然而，从经调整亏损率的角度看，相较于 2025 年同期实现了约 3.5 倍的大幅收窄。一个显著的变化是，研发投入已超过净亏损额，这表明核心业务产生的毛利在覆盖销售及管理费用后，已开始部分弥补研发投入。而在 2025 年年报披露时，研发投入与损益金额均为 31.8 亿人民币

左右，当时的情况是所有亏损基本等同于研发投入。

从 2026 年 3 月至今，大模型在长时程任务（长城任务）上的性能演进呈现出怎样的趋势？

在过去约五个月的时间里，大模型在长时程任务上的性能取得了显著进展。以 METR 独立实验室的报告为例，2026 年 3 月时，最新模型 Cloud Open4.6 在对标人类成功率 80%的基准上，将任务时长推进至接近 2 小时，当时国内模型普遍处于其三分之二的水平。而截至目前，长时程任务的模型已能处理 8 小时以上，甚至超过一天的任务长度。除了任务时长的延伸，任务的复杂度也出现了阶梯式提升，从早期的单步任务，发展到多步代理工程、常识推理的系统工程、前沿复杂工程，直至当前正在攻坚的真实环境推理，共经历了五个复杂度的台阶。

公司对于大模型技术选型的核心理念是什么，以及如何在预训练、后训练和推理三个维度上实现扩展？

公司技术选型的核心理念是“Right Scaling”，即在 scaling law 持续有效的认知基础上，并非简单地追求参数规模的扩大，而是在更广泛的范围内进行多维度扩展。具体而言，在有限算力的前提下，选择在特定阶段将算力投向最合适的维度。这主要体现在预训练、后训练和推理三个层面。 在预训练层面，公司凭借早期在 1.75 万亿参数“悟道”模型和首个千亿级商业化模型 GLM-130B 上的积累，建立了强大的数据与环境管线，涵盖了从 GitHub PR、网页数据抓取到知识图谱合成和智能体轨迹验证等。公司认为，总参数量决定了模型的知识量和记忆力上限，而模型的有效深度、训练深度和任务环境等维度的 scaling 则决定了其推理深度。基于“TPP”（Token-per-Parameter）法则，在数据总量固定的情况下，过度增加总参数反而会损害推理能力，而扩大激活专家的维度则能提升模型稳定性。 在后训练层面，公司投入了大量资源。算法上，从 DeepSeek V3 开始采用 GRPO 组内优化策略，后转向基于 critic based 的 FAO，为每个长轨迹进行独立标注，以期通过数据飞轮效应提升强化学习效果。在 GLM5.3 模型上，公司进一步在环境合成方面取得突破，在一个月内基于几 PB 级的环境进行长时程环境的强化学习，实现了环境与奖励信号的合成化扩展。 在推理层面，公司在过去半年进行了大量优化工作，包括在交换机层面采用 ZQ 组网技术，在千卡集群上实现成本下降至十分之一和 15%的吞吐量提升；应用 Index Share 等工程优化技术；以及在静态持续修复方面解决 Scaling Pack 问题。这些努力使得公司在国产算力储备规模处于国内前沿的情况下，能够高效优化基础设施，为独立产品（如 Coding Plan）和大规模企业级 API 调用提供有力支撑。

最新发布的 GLM5.3 系列模型在技术架构和市场定位上有何特点，如何体现其“双帕累托最优”的策略？

GLM5.3 系列模型体现了公司在技术与市场上的“双帕累托最优”（Double Pareto）策略，旨在同时覆盖高价值任务的深度和普惠级场景的宽度。 GLM5.3 Flash 模

型是面向下一代大基座模型的实验方向。该模型是一个原生多模态模型，使用了接近 30T 的原生多模态数据进行训练。在架构上，首次采用了稀疏加线性注意力混合架构。相较于参数规模相近的 2025 年模型 GLM4.5，其注意力乘数减少了近一半，从而在 KV Cache 缓存和算力需求上实现了质的下降。这种面向推理极致优化的设计，使其能够在 10 万卡级别以上的国产大集群上运行，支持大规模调用。 在市场定位上，GLM5.3 系列形成了双卡位布局。GLM5.3 模型定位于高价值任务场景，其在智能表现和性价比的综合考量下，性能优于市场上绝大多数模型，目前仅有海外的 OPUS5、FABLE5 和 5.6SO 三款模型处于其覆盖范围之外。而 GLM5.3 Flash 模型则凭借其低成本架构和与国产芯片的深度融合，在普惠层级实现了最佳性价比。尽管参数相对较小，但其 AA 指数仍达到 57.5 分，具备一线前沿模型的编程能力，在超过 50 个主流模型的帕累托前沿中占据了领先位置。

公司在后训练管线上的具体投入和成果体现在哪些方面，尤其是在 GLM5.3 模型上取得了怎样的性能提升？

公司在后训练管线上进行了深度投入，并取得了显著成果。首先，在算法层面，从早期的 DeepSeek V3 开始就采用了 GRPO（组内优化策略），以解决长轨迹数据难以进行组内归一化的问题。随后，行业内较早地转向了 critic based 的 FAO（critic-based FAO），即对每个长轨迹进行独立的 critic 标注。这种方法旨在通过数据飞轮效应，利用不断深化的训练数据和奖励信号来提升强化学习效果，反过来又能标注出更深层次的数据。 在 GLM5.3 模型上，公司在数据的基础上，进一步在环境合成方面取得了突破。在一个月的时间内，集中力量在几 PB 级的环境基础上进行长时程环境的强化学习，实现了环境本身乃至奖励信号的合成化扩展。这一系列努力带来了巨大的性能代差：与 GLM5.2 相比，GLM5.3 在 Terminal Bench 3.0 上的得分从 4.6 分大幅提升至 28.4 分；在 DeepSeek 上的得分也从 46.2 分提升至 66.9 分。这些数据清晰地展示了在后训练环境上进行极致投入和扩展所带来的成果。

如何理解“算力乘数”这一概念，以及它在评估算力投入回报方面的应用？目前公司的算力乘数表现如何，未来的发展趋势是怎样的？

算力乘数是指每投入一元算力所带来的整体收入提升，该投入不仅包括推理侧，也涵盖了训练侧的成本和费用。这是一个用于评估 Capex 投入在何种节奏、规模和速度上改善收入的基准。公司在初期阶段的算力乘数约为 0.02，作为对比，Anthropic 在 12 个月前（约 2025 年第一季度）的数据为 0.25。尽管起点较低，但公司的算力乘数每季度均有大幅提升，尤其在 2026 年第二季度，随着基础设施取得大规模进展，其上升曲线实现了显著增长。从年初至今，整体算力乘数提升了 14 倍。公司有信心在未来 6 至 12 个月内达到算力乘数为 1 的临界点，即每投入一元算力能带来一元收入，届时主要问题将转向毛利率。此外，在一次基于超大规模国产卡集群的受控实验中，GM5.3 模型成功经受住了 Open Code 和 Open Router 双平台每日几十 T 调用量的考验，证明了国产加速卡在推理层面完全能

够支持一线模型的全球调用需求。

公司历史上的 ARR（年度经常性收入）披露情况是怎样的？如何理解 ARR 与财务确认收入之间的关系，以及公司当前的 ARR 水平和未来的指引？

公司官方共进行过三次 ARR 披露。第一次是在 2026 年 3 月 31 日发布年报时，披露截至当日的 ARR 为 2.5 亿美元，并给出了全年 10 亿美元的首次指引。第二次是在 2026 年 7 月的市场交流中，公布 ARR 达到 10 亿美元。第三次是本次交流，披露 ARR 已达到 16 亿美元。ARR 与财务确认收入存在差异：ARR 是管理层口径，反映的是订单收入，其中包含了部分在财务上需向后分摊的收入，例如包年、包季或包月的订阅服务费在财务确认时需要按天分摊。同时，月均 ARR 是基于当月月均数据乘以 12 计算得出，因此对于高速增长的公司，其与确认收入的差距会更大。以海外头部模型公司为例，Anthropic 在 2025 年年底 90 亿美元的 ARR 对应 47 亿美元的收入，比例关系约为 2:1；而在 2026 年 7 月，其 640 亿美元的 ARR 对应上半年 165 亿美元的确认收入，呈现约 4:1 的关系。公司的 ARR 增长呈现脉冲式，主要受算力供给的阶段性错配影响。截至 2026 年 6 月底，公司的 ARR 约为 5.3 至 5.4 亿美元，对应上半年确认收入为 9.5 亿人民币（约 1.4 亿美元），符合约 4:1 的比例关系。2026 年 7 月和 8 月的 ARR 分别为 10 亿美元和 16 亿美元，这两个月的收入相较上半年有非常高比例的攀升，这得益于 7 月份 5.2 版本和 8 月份 5.3 及 5.3 flash 版本的密集放量。基于上半年的情况和近两个月的真实评估，公司将年底的 ARR 指引设定在 24 亿美元。

与行业对标公司相比，公司在 ARR 增速方面表现如何？

公司的 ARR 增长速度显著。从 2026 年年初至今（1-8 月），公司的 ARR 相较 2025 年 12 月 31 日增长了 46 倍，而同期对标的 Anthropic 增长约 7 倍，尽管其基数较大。目前公司 16 亿美元的 ARR 水平，相当于 Anthropic 在 2025 年 3 月的水平，这意味着两者在 ARR 规模上的差距已从最初商业化起步时的 24 个月缩短至 17 个月。在规模化渗透速度方面，公司将 ARR 从 1 亿美元提升至 16 亿美元用时约 6 个月（从 2026 年 2 月至今），而 Anthropic 完成同等规模的爬坡则花费了 19 个月。

驱动公司收入和 ARR 增长的核心因素是什么？MaaS 平台和高频应用场景的具体运营数据表现如何？

增长主要由两大因素驱动。首先是前沿智能 MaaS 平台的历史性放量。从年初至今，平台注册用户增长了 144%，覆盖全球 233 个国家和 2 万多个城市，已成为中国模型公司中最大的独立 MaaS 平台。平台整体调用量翻了 40 倍，最新日均调用量达到 40 万亿。其中，排名前十的客户调用量同比增长 98 倍，其日均调用量超过 15 万亿，占平台总量的近 40%。其次是高频场景快速驱动了用户使用深度的提升。以 Coding Plan 为例，该产品于 2025 年 9 月推出时，首周日均调用额仅为 2.5 万人民币。尽管在 2026 年上半年因算力瓶颈限售了半年，直到 7 月份

才恢复销售，但目前的日均调用额已接近 600 万人民币，增长超过 200 倍。Coding Plan 是面向高质量开发者的专业编程订阅产品，其运营尤其考验厂商的推理优化能力。

请介绍公司上半年不同业务模式（如 Coding Plan、通过 Work Body 转售 Token 等）的收入占比情况，以及在会计处理上是否存在 Gross 和 Net 的差异？此外，上半年研发支出低于预期的原因是什么，下半年的研发支出预期如何，以及毛利持续覆盖研发开支是否需要以高频的模型更新为前提？

关于收入结构，特别是开放平台和 API 业务的构成，目前高质量用户群组中包含规模非常大的客户。由于公司提供的是平台级服务，用户统计口径与 ToC 产品不同。例如，一个互联网巨头的 IDE 产品，其背后可能代表数十万开发者用户，但在公司层面仅计为一个客户。因此，公司披露的 740 万注册用户主要是基于 B 端客户的统计。 在收入来源上，除了公司官方 Mars 平台的 API 调用和 Coding Plan，确实有相当一部分收入来自与大型互联网公司 IDE 产品的分发合作。不过，公司的合作模式与海外模型厂商不同。国内市场并未出现需要反向利用 CSP 大规模算力的情况，因此会计上 Gross 和 Net 的问题在公司业务中不显著。大部分与 IDE 工具的分发合作更接近于一种分发分层，且许多 IDE 厂商选择将折扣让利给最终用户，而非作为渠道分成。这部分合作带来的收入占比不高，在 20%以内，多数情况下为个位数百分比，不会出现海外市场中模型厂商与算力公司可能出现的三七开或二八开的分成模式。 关于研发费用，当前模型公司超过三分之二的研发投入主要集中在算力侧。上半年研发投入低于预期，主要有两方面原因。首先，公司自身的 Infra 能力有显著提升。算力侧的投入不仅包括算力采购或租赁的成本，也取决于 Infra 能力，即如何用同等规模的算力进行更高效的训练，并提供更高价值的 Tokens 来服务用户。Infra 能力的提升是关键因素之一。其次，这也与整体算力的供给节奏有关，公司上一轮融资是在今年 7 月份完成的。

公司下半年的资本支出和研发费用投入规划是怎样的？在算力配置和模型迭代节奏方面有何具体考量？

公司拥有完善的算力规划与纪律，能够满足以 GM5.3 Flash 为代表的、对算力优化要求极致的模型需求，从而推动智能与成本的优化。在此基础上，算力配置和使用的灵活性与纪律性至关重要。 关于下半年的资本支出和研发费用投入，可以明确以下几点：首先，作为基础模型公司，相关投入会继续增长，但增长将遵循严格的纪律，例如资本支出会与公司的 ARR（年度经常性收入）及整体商业化增长按一定比例挂钩，但具体比例细节不便披露。因此，从全年来看，下半年的资本支出肯定会高于上半年，但不会失去纪律性。其次，模型的迭代节奏不会放缓。在当前行业环境下，保持领先地位的关键在于迭代速度、智能表现和成本控制这三个因素的乘积效应。迭代速度是其中非常重要的一环，公司将按既定规划和节奏持续推进。

公司开放平台在中期报告中披露的毛利率约为 24-25%，与海外头部模型公司存在差距，请问公司如何看待未来的提升空间以及国内 SOTA（State-of-the-Art）模型公司的竞争格局？

上半年公司给出的指引是毛利率将以个位数（single-digit）的幅度理性攀升，最终对比 2025 年年底实现了接近 6 个百分点的增长。全年来看，预计毛利率仍将稳步攀升。与海外模型公司从去年约 40-50%提升至当前 70-80%的毛利率指引相比，国内模型公司面临着几点核心差异，这些也是未来毛利率的提升空间所在：第一，基础设施的爬坡周期。公司较早获得了大规模国产算力集群的支持，但新集群的利用率尚未完全成熟，仍处于快速扩容阶段。新的算力基础设施上线后，需要经历模型部署、流量迁移和利用率爬坡的完整过程。在此期间，成本先行，而收入和 Token 调用量是逐步填充的，因此前期毛利率会受到拖累。随着调用规模持续扩大，固定成本被更多 Token 分摊后，规模效应将得到充分释放。 第二，模型快速迭代带来的切换成本。国内模型技术迭代速度极快，公司基本保持每 2 到 2.5 个月推出一代新模型的节奏。新旧模型切换期间，旧版本模型（尤其是较早期的版本）的推理部署、缓存和资源配置无法在第一时间完全退出，技术迭代越快，早期产生的资源冗余和迁移成本就越高。公司正通过基础设施优化和更有纪律的规划来尽可能规避这部分损耗。 第三，全栈优化能力的持续提升。推理成本的降低是一个系统工程。在过去半年，公司单位 Token 的推理成本已下降接近 80%，这是模型层、系统层和芯片层三层共同优化的结果。具体包括：模型架构设计（如更低的激活参数、混合稀疏架构优化）、系统层优化（如 KV Cache 优化、PagedAttention、量化调度、网络优化）等。这是一个精细化的全栈优化过程，任何一个环节的微小改进，经过乘数效应都能带来系统效率的显著跃升。公司在此方向上已取得成效，未来仍有较大提升空间。 第四，市场进入（Go-to-Market）策略的精细化。未来，大规模模型厂商的服务将是矩阵式的，而非用单一模型或单一价格承接所有流量。公司的策略是：旗舰模型服务于对能力极度敏感的高价值任务，占据高价格带，这是公司一直坚持的“打塔尖”路线。同时，随着算力基础（尤其是国产算力）的升级，公司从 Flash 模型开始，也面向高频、成本敏感型需求提供服务。这类模型通过优化的架构设计和国产算力支持，在保持前沿智能水平的同时，实现了更低的售价和底层计算成本。未来，公司将通过更成熟的“智能路由”机制，将用户的不同任务精准匹配到最合适的模型上，减少当前可能存在的“轻量级任务由贵模型承接”或“高价值任务由轻量级模型承接”的错配情况。产品和定价策略的持续优化将大幅提升毛利率。 综合以上四个方面的改善，公司有信心在未来 12 到 18 个月内，将整体毛利率提升至 50%以上的水平。

公司计划在年底推出的模型将对标海外的先进水平，请问具体将在哪些层面（如多模态、长文本等）实现对标甚至超越？同时，如此大规模的模型对算力提出了怎样的要求？

公司计划在年底推出的新模型，目标是在多个维度上对标甚至超越业界最前沿的水平，例如在多模态和长文本（long context）等领域。

关于公司下一代模型与 Miso 模型（音）的对标情况，以及到 2026 年底公司的算力储备预计能支撑多大规模的商业化收入？

公司的下一代模型将继续遵循 Scaling Law，在数据、基座模型、后训练以及基础设施四个维度进行扩展，目标是在 2026 年底模型层面达到接近 Miso 模型（音）的水平。 首先，在数据层面，将利用更多、更高质量的数据来提升模型的智能上限。 其次，在基座模型层面，将在 GEM5.2 和 5.3 模型探索后训练上限的基础上，进一步提升基座模型的规模上限。 再次，在后训练层面，将基于更大规模的数据和更多的交互环境，通过强化学习和反馈来进一步提升模型效果。 最后，在基础设施层面，将通过扩展更高效的 Infra，在有限的算力条件下部署更大的模型并提供更大范围的服务。 通过这四个维度的扩展，公司有信心在 2026 年底使模型能力接近 Miso 模型（音）的水平，并将在 Coding Agent、安全、法律、金融等高价值场景上持续逼近其效果。

公司在 Coding 业务之外，下一个商业化场景 Cowork 所对应的 5 万亿市场规模是如何测算的？以及判断该市场迎来加速发展的时间点和关键标志是什么？

Cowork 场景的市场规模测算并非沿用传统企业级软件渗透率的逻辑，而是基于对高级知识工作者市场的任务池化潜力评估。具体测算逻辑如下：以 2025 年全球约 120 万亿 GDP 为基数，根据世界银行估算，其中 40%暴露于 AI。在这部分中，知识密集型服务行业（如法律、金融、网络安全等）约占 63%。由此计算出的全球知识工作者市场（包含工资、外包成本、软件及运营成本）规模约为 20 万亿至 30 万亿。考虑到当前模型发展阶段尚处于长程任务能力，可被 AI 转化的任务值比例约为 20%至 40%，因此最终得出约 5 万亿的市场空间。 判断 Cowork 市场加速的时间点，关键在于模型能力是否越过“可用线”。这个门槛的标志是，模型在复杂工作流中完成的工作成果达到了 60 至 70 分的及格水平，用户愿意在此基础上进行完善，而非推倒重来。以个人工作流为例，过去在 Vibe Coding 阶段，非技术背景的用户难以使用模型。但当模型具备强长程任务能力后，情况发生了改变。例如，在进行市场分析时，通过向编排工具提出模糊需求，基于 GEM5.3 模型的系统不仅能抓取新闻、为个股建立可视化窗口，还能主动提出创新性解决方案，比如通过整合分钟级和秒级的交易快照数据，解决了港交所 T+2 交割系统导致的数据延迟问题，从而实现了对港股交易行为的实时监控。这个由模型自主生成的可视化系统，尽管在 UI 等方面尚有不足，但其核心功能已达到七八十分的水平，并已完全融入个人日常工作流，替代了传统的金融终端工具。这种体验表明，模型已经从一个工具转变为工作流中的“同事”角色，标志着 Cowork 的应用正在真实发生。

在模型后续的竞争力构建中，单位成本下的智力水平是关键，公司在 Infra 侧通过哪些优化来提升算力乘数？

公司在 Infra 侧已实现了显著的算力乘数提升。

公司在 Infra（基础设施）层面未来核心的技术优化方向和侧重点是什么？

公司在 Infra 层面的优化分为几个阶段。2026 年上半年初，在 GM5 模型架构已确定的前提下，我们首先致力于在有限算力下提升推理效率。具体措施包括：第一，深度优化算子并对模型权重进行量化操作，以提高算力利用效率，使其更接近理论计算上限。第二，针对 GM5 这类大模型推理涉及的跨节点场景，我们通过优化计算与网络通信的重叠来提升效率。此外，我们还将传统的 Clos 组网优化为 ZQ 组网，在 PD 分离部署中能极大减少网络拥塞，确保 P 和 D 节点间 KV Cache 传输的高效性。第三，我们定位到 KV Cache 是主要的性能瓶颈。对于支持 100 万上下文长度的模型，单个请求的 KV Cache 可能占用高达 50GB 的空间。为此，我们采用 Index Share 技术将 KV Cache 占用降低至原来的 1/4，并利用 Layer Split 技术减少每张显卡上存储的 KV Cache 量。这些措施不仅提升了并发处理能力，也减少了网络传输的数据量，从而提高了系统整体吞吐。 基于在优化过程中积累的对模型架构与硬件协同的理解，我们在最新的 Flash 模型上应用了这些成果，设计了新的模型架构，例如稀疏和线性混合注意力机制，以进一步减少 KV Cache 占用。未来，这些已有的成果将被应用于更大规模的模型，确保其在更大规模下依然保持高效。总的来说，我们的优化策略是从硬件层到模型架构的协同优化，以实现更高的效率。目前我们尚未达到算力的上限，在异构算力和模型架构方面仍有持续优化的空间。

考虑到国内算力获取成本日益增高，且公司 API 定价相比海外模型具有性价比，开放平台毛利率在未来 12-18 个月提升至 50%左右的目标，其背后的主要驱动因素是什么？

开放平台毛利率的提升主要基于以下几点： 首先，从市场结构上看，存在一个结构性机会，即算力供给将从过去的多元异构状态，逐步转向由国产大规模集群主导，从而带来规模效应。 其次，在算力获取策略上，公司正从过去以租用算力为主，转向自建超大规模算力中心。自建集群将大幅提升公司在生态层面的算力能力。同时，公司自 2026 年初便启动了算力合伙人计划，与海内外众多生态伙伴合作。鉴于大模型公司是产业链中算力需求的主要驱动方，凭借当前的业务体量和需求，公司能够在生态中有效整合资源，进行错峰调度。 最后，在技术层面，Infra 的持续优化是关键。自 2026 年 GM5 模型发布至今，推理成本已降低了 80%，并且优化速度仍在加快。除了推理引擎本身，公司在调度和负载均衡方面也进行了大量优化，以减少损耗，提高集群利用率。随着客户数量的增长，集群的整体负载和利用率也得到提升。此外，产品与 Infra 的协同优化也带来了显著效果，例如在 Z code 产品中，缓存命中率已从 90%提升至 98%以上。更高的缓存命中率意味着在同等算力下可以服务更多的 Token，实现更高的 Token 吞吐量，这对降低推理成本有极大帮助。

公司在海外市场的拓展方式是怎样的，以及未来的模型在闭源和开源策略上会有何变化？

在海外业务拓展方面，我们正与海外主流的 CSP 进行紧密合作。在 2026 年 6 月之前，我们已与 AWS 在分发层面启动了初步合作。虽然该合作中算力主要由我方提供，AWS 作为渠道方，但它为我们与终端用户之间建立了重要的桥梁，为后续与 CSP 及新兴云服务商的合作打开了空间。近期，我们观察到与欧洲和美国的两家大型云服务商及独立模型公司存在合作机会，即使是开源模型，也有可能采取类似 OpenAI 的合作模式。我们对即将公布的合作进展感到乐观。 关于开源与闭源的选择，我们认为当前阶段闭门造车并不可取，因为与海外前沿模型相比仍有差距需要追赶。市场空间足够大，无需通过闭源来获取市场份额。对我们而言，开源是我们的初衷和使命，我们相信最前沿的智能技术应当服务于整个社会。只要我们能持续提升模型的智能上限，商业模式自然会随之涌现。此外，开源是中国模型发展至今并形成友好团结生态的重要原因，我们受益于此，也不可能脱离这个生态。我们希望与更多生态伙伴一起，在更高的智能水平上，让前沿智能普惠大众。