

智谱 Z.AI 02513.HK 2026 中期业绩公告（中文翻译版）

目录

北京智谱华章科技股份有限公司 (Z.AI Co., Ltd.)	2
截至 2026 年 6 月 30 日止六个月中期业绩公告	2
一、财务摘要 (Financial Summary)	2
收入按业务条线分解 (Revenue Breakdown)	2
二、管理层讨论与分析 (Management Discussion and Analysis)	3
2.1 业务回顾 (Business Review)	3
2.2 战略重点 (Strategic Focus)	3
2.3 技术路线与模型进展 (Technology Roadmap & Model Progress)	5
2.4 商业化进展：对的模型天生会走向市场	6
2.5 生态与安全：为能力赢得现实世界的信任	8
2.6 未来展望 (Future Outlook)	9
三、财务回顾 (Financial Review)	11
3.1 关键经营表现项目分析	11
3.2 关键资产负债项目分析	14
四、简明合并财务报表 (Consolidated Financial Statements)	15
4.1 合并损益及其他全面收益表	15
4.2 合并财务状况表 (Consolidated Statement of Financial Position)	16
五、合并财务报表附注 (Notes to the Condensed Consolidated Interim Financial Statements)	17
附注 1：编制基础	17
附注 2：会计政策变更	18
附注 3：收入与分部报告	18
附注 4：合并损益及其他全面收益表中的所得税	19
附注 5：每股亏损	20
附注 6：物业、厂房及设备	21
附注 7：应收账款及应收票据	21
附注 8：预付款项及其他应收款	22
附注 9：应付账款及其他应付款	22
附注 10：股息	23
附注 11：企业合并 (XCore Sigma 收购)	23
附注 12：报告期后非调整事项	24
六、其他信息 (Other Information)	24
6.1 中期股息	24
6.2 上市证券的购买、出售或赎回	24
6.3 报告期后重大事项	24
6.4 企业管治守则合规	25

6.5 标准守则合规	25
6.6 《上市规则》下的持续披露义务	25
6.7 公持量充足性	25
6.8 审核委员会	25
6.9 中期财务资料审阅	25
6.10 中期业绩与中期报告的刊发	25
七、术语表 (Definitions & Glossary)	25
7.1 定义 (Definitions)	25
7.2 技术术语表 (Glossary of Technical Terms)	26

北京智谱华章科技股份有限公司 (Z.AI Co., Ltd.)

(于中华人民共和国成立的股份有限公司)

股份代号: 2513

截至 2026 年 6 月 30 日止六个月中期业绩公告

香港交易及结算所有限公司及香港联合交易所有限公司对本公告的内容概不负责，对其准确性或完整性亦不发表任何声明，并明确表示概不对因本公告全部或任何部分内容而产生或因倚赖该内容而引致的任何损失承担任何责任。

董事会欣然提呈本集团截至 2026 年 6 月 30 日止六个月的未经审核简明合并中期业绩，连同 2025 年同期的比较数字。

本公告中所载的若干金额及百分比数字已进行四舍五入调整，或已四舍五入至小数点后一或两位。任何表格、图表或其他数据中所列总额与所列各项金额之和之间的任何差异，均因四舍五入所致。

一、财务摘要 (Financial Summary)

项目	截至 6 月 30 日止六个月 (人民币千元, 未经审核)		同比变动
	2026 年	2025 年	%
收入 Revenue	953,892	190,877	+399.7%
毛利 Gross profit	251,611	95,424	+163.7%
研发开支 R&D expenses	(2,131,171)	(1,594,661)	+33.6%
期内亏损 Loss for the period	(2,071,992)	(2,357,852)	-12.1%
经调整期内净亏损 (非 IFRS)	(1,964,138)	(1,751,974)	+12.1%

收入按业务条线分解 (Revenue Breakdown)

业务条线	2026 H1 (人民币千元)	占比	2025 H1 (人民币千元)
开放平台及 API (Open Platform & API)	825,176	86.5%	29,100
企业级智能体 (Enterprise-level Agents)	55,561	5.8%	13,739
企业级通用大模型 (Enterprise-level General-Purpose LLMs)	67,037	7.0%	147,555
技术服务及其他 (Technical Services & Others)	6,118	0.7%	483
合计	953,892	100.0%	190,877

关键观察：开放平台及 API 业务收入占比由 15.2% 跃升至 86.5%，公司收入结构完成由”本地化部署”向”云端 MaaS 订阅”的根本性切换。

二、管理层讨论与分析 (Management Discussion and Analysis)

2.1 业务回顾 (Business Review)

2026 年上半年，公司在大模型编程 (Coding) 领域验证了从早期简单代码补全，到整文件代码生成，再到完整工程项目交付 (Agentic Engineering, 即”智能体工程”) 的全过程实现。模型已不再是单纯的工具，而是具备了独立完成整项任务的能力。由此带来的收入结构变化是：**开放平台及 API 服务（按调用量计费）收入占比由上一全年的 26.3% 跃升至 86.5%，本地化部署收入占比则由 73.7% 降至 13.5%**。2026 年上半年这一核心收入结构的重构，本质上是模型能力代际跃迁的必然结果。

网络安全 Co-work 是同一能力曲线的延伸。软件漏洞发现不依赖记忆，而依赖模型的推理深度和长程自主性——这正是编程能力跨越”工程交付”门槛后自然延伸的方向。报告期内，公司模型在 CyberGym 上取得 84.5% 的得分（超越 Fable 5 和 GPT-5.6 Sol），在 ExploitGym 上完成 130 个任务（仅次于 Fable 5 和 GPT-5.6 Sol，排名第二）。

支撑这两条曲线的是基础设施。自 2026 年初起，公司全面实施”全栈基础设施 (All-in-Infra)”战略，在**训练与推理两侧同时投入，并将国产芯片纳入主力推理算力**。公司已实现 10 万级国产芯片的规模化低成本推理，单位 Token 推理成本较年初下降 80%。在全球算力供给持续紧张的背景下，这条路径决定了公司的成本曲线掌握在自己手里。

报告期内，GLM MaaS 平台企业及开发者用户累计超过 580 万。公司总收入约人民币 9.54 亿元（约 1.42 亿美元），同比增长近 400%；其中开放平台及 API 收入约人民币 8.25 亿元（约 1.23 亿美元），同比增长 2,735.7%。

2026 年 8 月，公司发布 GLM-5.3 与 GLM-5.3-Flash（内部代号 Ox-Alpha），并同步将 GLM Coding Plan 订阅面向全部用户开放，覆盖 Lite、Pro、Max 与 Team 四个版本，使其成为中国最早以订阅模式提供”编程智能体”能力的大模型服务之一。截至本公告日期，GLM MaaS 平台企业及开发者用户已超过 740 万。

2.2 战略重点 (Strategic Focus)

智能上界定任务边界，任务价值定商业空间。

2026 年初，我们提出一个公式：

AGI 商业价值 = 智能上界 × Token 消耗规模

智能上界决定模型能处理哪些任务、每个 Token 能创造多少用户价值；智能上界越高，模型可承接任务的价值密度越大，每单位智能对应的经济产出越高。Token 消耗规模体现智能利用的广度与深度——智能上界首先解锁任务边界，任务边界再决定 Token 在何处按价值密度分配。同样数量的 Token 用在闲聊与用在修复生产环境漏洞上，产出相差可达数个数量级。

能力阶梯 (Capability Ladder) 模型能力的演进不是一组平行的产品线，而是一段只能逐级攀升的阶梯：

对话 (Chat) → 编程 (Coding) → 智能体 (Agent) → 协作 (Co-work) → 自主 AI (Autonomous AI)

- **对话**：单轮响应，任务在一轮内结束，价值密度最低但覆盖面最广。
- **编程**：交付可运行代码，引入首个客观验收标准——能跑通就跑通，跑不通就跑不通。
- **智能体**：交付完成的多步任务链，模型开始自行调用工具、处理中间失败。
- **协作**：交付“专业人员愿意审阅而非推翻重做”的工作成果，进入真实工作流并按完成的任务计价。
- **自主 AI**：交付持续运行能力，模型在无人监督下完成长期目标。

每一级向上都被一项具体的技术门槛所卡：

- 对话 → 编程的门槛是结果可验证性；
- 编程 → 智能体的门槛是长程规划与错误恢复；
- 智能体 → 协作的门槛是可靠性达到“专业人员愿意审阅”的水平；
- 协作 → 自主 AI 的门槛是模型能判断自身输出的正确性。

跨不过这道门槛，对应层级的商业模式就不成立。

市场空间 (Market Space) 按粗略的“数量级”估算：

- **对话** ≈ 约 1,000 亿美元搜索市场——是互联网时代给出的答案；
- **编程** ≈ 约 5,000 亿美元软件研发支出；
- **协作** ≈ 约 5 万亿美元知识工作支出；
- **长期** (AI 重构操作系统与人机交互、行业级升级) ≈ 约 30 万亿美元可触达空间。

(上述市场空间数字为公司基于公开行业数据的粗略估算，仅用于说明任务层级与市场量级的对应关系，未经独立核实，亦不构成公司对其业务、财务状况或经营业绩的预测或保证。公司股东及投资者不应过度倚赖该等数字。)

相应的，行业评价标准也在切换：从单轮响应质量与 Token 消耗量，转向长程任务成功率、单位智能成本、结果可验证性，以及为客户交付的实际经营结果。

报告期内，公司聚焦于阶梯的**第二档 (编程)**与**第四档 (协作)**之间。编程方向已走完整链路；网络安全已进入协作验证阶段；法律等其他方向仍处早期。2026 年上半年公司研发开支约人民币 21.31 亿元 (约 3.17 亿美元)，体现了对“攀登新高”的坚定投入。

为什么是这个顺序 我们对 AGI 持朴素理解：它不是某个人的智能，而是人类智能的总和，并最终应具备生成相对论级别原创知识的能力。决定这条路径推进速度的，是模型能力的代际水平。这是公司一切产品与商业逻辑的起点。

公司是中国最早投入大模型研发的团队之一。从 GLM 预训练架构和“悟道”系列到今天的 GLM-5 系列，我们本质上专注一件事：**训练更强的模型**。行业热时，我们不散资源去铺多条产品线；行业冷时，我们不停下模型迭代。回顾来看，每一次收入陡升都发生在模型能力提升、新任务场景解锁之后——而不是之前。

2.3 技术路线与模型进展 (Technology Roadmap & Model Progress)

模型公司能力的核心在于将单位资源转化为智能的效率。算力、数据、人才都有成本；公司将这三类输入转化为模型智能的效率，决定了它能走多远。

我们如何理解 SOTA 过去一年，行业对”SOTA”的理解被改写。在某一时刻领先榜单已不再是最重要的事。真正重要的是谁能以更快的节奏、更好的性价比，把更强的模型持续推向市场。我们将 2025 年 9 月至当前 5 家厂商的 27 个旗舰模型放入同一坐标系：横轴为范式进步、纵轴为智能指数、深度轴为单位任务成本。在此框架下：单点冠军只是转瞬即逝的点；**持续最优才是不断向上的轨迹**。所有玩家都在前进，但节奏和成本不同。

GLM 系列在约 11 个月内完成 6 次迭代——4.6、4.7、5、5.1、5.2、5.3，智能指数由 32 提升至 60，单任务成本稳定在约 0.20 美元，覆盖图中所示全部三种范式阶段。**GLM-5.3-Flash 以单任务成本 0.045 美元进入”智能-成本”前沿**。（智能指数与单任务成本来源：Artificial Analysis Intelligence Index，截至 2026 年 8 月 27 日；任务指标：Intelligence Index 任务的单位成本。）

报告期内，公司技术路线横跨三代模型：

(1) GLM-5：以编程为入口，解锁长程智能 我们之所以选择编程作为能力突破的入口，是因为软件工程是离”可验证智能”最近的真实环境：任务目标清晰、工具链完整，代码能否运行、测试能否通过、问题能否修复都构成清晰的反馈信号。没有可验证的反馈，就没有可供优化的目标。

GLM-5 的能力重心从代码生成扩展到系统理解、任务规划、工具调用、持续调试与工程交付，**推动模型从 Vibe Coding（氛围编程）迈向 Agentic Engineering（智能体工程）**。

编程对公司而言的价值远不止是帮开发者写代码。在更长的任务链上，模型需要持续理解目标、调用工具、处理异常并根据环境反馈调整策略。通过编程这一”高反馈密度”的真实环境，所积累的长程任务能力可迁移到网络安全、数据分析、复杂自动化等其他专业领域。报告期内，编程方向已成为公司的首要收入来源，上述能力迁移已在网络安全 Co-work 中得到验证。

(2) GLM-5.2 与 GLM-5.3：边规模扩展，边向纵深 scaling

- **扩展规模 (Scaling Up)**：扩大模型规模、上下文长度、算力与系统容量；
- **向纵深 scaling (Scaling Deep)**：在后期训练 (post-training)、强化学习、长程轨迹与任务反馈上加大投入。

在当前阶段，我们把资源优先投向后者。历史上早有先例：Kaplan 缩放律让参数规模远超数据增长；Chinchilla 以”20 Token / 参数”的比例修正了这条路径——70B 模型在 1.4T Token 上训练，全面跑赢 280B 的 Gopher。在推理时代，模型每天被调用数十亿次，推理成本超过训练成本，最优解转向”小而训得久”。MoE 进一步将参数划分为两个维度：总参数决定能记住多少，激活参数与深度决定推理能走多深。最新研究也指出，最优 Token-参数比随任务而异：偏记忆任务偏好更大参数，偏推理任务偏好更多数据。网络安全这类任务不需要记住更多 CVE，而是要求模型把一条 20 步的因果链推到底而不中断——这种能力不在总参数里。

GLM-5.3 是这一判断的一次验证。它与 GLM-5.2 采用**相同的架构、相同的总参数、相同的激活参数**，唯一的变量是后期训练规模——用一个月时间放大长程任务环境与强化学习的规模。在公司自建真实编程评测上，GLM-5.3 端到端任务完成率较 GLM-5.2 提升超过 50%（评测方法与样本量详见 GLM-5.3 技术博客）。这近似一次”对照实验”，说明规模不只是参数，后期训练在当前对性能提升的边际贡献最大。

我们的判断是：强大的模型只能在合适的时机用恰当的数据、恰到好处的训练规模来获得。目标不是单纯追求最大规模，而是在当前数据与算力约束、足够训练深度下，交付”智能上界”最高的那个模型。

这一思路沿四个维度展开：

1. **数据**：持续提高知识密度与任务复杂度，并通过开发者生态、行业合作、模型间自博弈 (Model-with-Model Self-Play) 积累高质量长程智能体轨迹，经筛选、验证、合成后反馈进训练；
2. **中训练 (Mid-training)**：介于预训练与后期训练之间，公司单独划拨算力用于中训练阶段，以提升知识密度、延长有效上下文长度，并为工具调用、长程任务打基础模式；三阶段边际回报不同，算力按阶段分项预算而非一次性投入；
3. **算法**：面对奖励分配、反馈利用、长程持续学习的新挑战，我们正由 GRPO 转向探索基于评论器的 SAO (Critic-based SAO)，并由 SFT 扩展至 OPD；自 GLM-5.2 起，公司已对超长轨迹引入基于评论器的单轨迹优化；
4. **工程与环境**：Slime 提升大规模强化学习训练效率；Index Share 降低长上下文稀疏注意力开销；MTP 优化推理解码；LayerSplit 与 ZCube 改进 KV cache 管理与集群网络效率；同时持续扩展 Coding Agent 和更多真实交互环境，让模型在“执行-失败-重规划”的循环中持续学习。

四维互相咬合：**环境产出轨迹，算法提升轨迹利用率，工程支撑更大规模的训练推理，数据决定起点。**

(3) GLM-5.3-Flash (测试代号 Ox-Alpha)：新一代架构与系统能力 如果说 GLM-5.3 代表当前的能力上界，GLM-5.3-Flash 代表成本前沿。两者面向不同任务场景：前者面向需要最高智能的长程任务；后者面向高频、海量、对成本敏感的调用。

GLM-5.3-Flash 采用全新架构、为超低成本而生：**总参数 320B、激活参数 18B、层数 45 层**，采用稀疏注意力与线性注意力的混合架构，显著降低长上下文服务成本的同时保持长上下文精度，并以流形约束超连接 (mHC) 增强可扩展性。结合 30T 多模态预训练语料，GLM-5.3-Flash 在 benchmark 与真实应用上超越 GLM-5.2，且价格仅为后者的十分之一。

正式发布前，我们将匿名模型 Ox-Alpha 部署到 OpenCode 与 OpenRouter 做真实调用测试：上线当日登顶 OpenRouter 榜单并刷新该平台单日 Token 用量纪录；一周内成为两平台被调用最多的模型，6 天累计 Token 量超过 62 万亿，带动平台整体流量增长超过 20%。

(4) 国产芯片支撑国产模型 GLM-5.3-Flash 的推理，是公司首次以国产芯片集群全栈承载超大真实流量，芯片间通过自研高带宽网络互联。这件事的完成方式本身就值得记录：**整个推理引擎由 GLM-5.3 驱动的 Infra Agent 显著加速，协助工程师开发并优化算子、诊断性能瓶颈、改进部署服务栈。**模型优化系统，系统承载模型——这一闭环在公司内部首次跑通。算子开发周期缩短一半。这也是公司自身 Co-work 路径的一次内部验证。

主要技术约束在国产芯片的内存容量与带宽上，对 1M 上下文的支持尤为吃力。为此我们在 SGLang 之上构建专用推理引擎，自定义优化“以算力换带宽、以通信换内存”，并在集群层采用生产级的 **Encode –Prefill –Decode 分离架构**，把多模态编码、Prompt 预填、单 Token 解码拆分为可独立调度与扩展的工作池。**相比同硬件的初始基线，端到端服务性能提升 3 倍；在编程场景下，国产芯片上的单 Token 推理成本已达到与主流进口 GPU 相当的性价比水平。**国产芯片可在经济上支撑前沿模型的大规模推理。

2.4 商业化进展：对的模型天生会走向市场

(一) 收入结构翻转 公司收入结构的演进，本质上是一场由模型能力提升驱动的交易模式演进：

卖模型 → 卖调用 → 卖订阅 → 卖端到端任务成果

这四种交易模式不是平行的产品线，而是随模型能力提升依次解锁的形式。模型能独立完成任务越完整，客户购买的就越接近“成果本身”，收入计价单位也随之变化。

1. 本地化部署阶段：完成深度行业积累与验证。模型早期尚不能独立完成完整任务时，客户购买的是需要在其自身环境中集成、定制、部署的工具。再加上数据安全、合规与自主可控的要求，本地化部署成为企业采用大模型的天然起点。本阶段交易模式为“卖模型”——一次性授权、项目制交付。这一阶段公司完成了对重点行业的渗透，形成了对复杂场景的理解，并构建了企业级交付能力。这些积累今天仍有价值——变的是交付形式，不是行业理解。

2. 编程阶段：从一次性交付走向持续调用。GLM 编程能力的提升带来 API 与 Coding Plan 的快速渗透。本阶段公司主动收缩本地化模型软件授权业务，转向持续提供“可调用、可扩展、可计量”的智能服务。截至 2025 年末，API 调用量较 2025 年 6 月末增长约 7 倍，开放平台及 API 收入占比约 26.3%，毛利率提升至 18.9%。**编程成为 Token 消耗与云端收入的第一个规模化入口。**这一步的意义不止于增速：收入确认由一次性变为经常性，公司首次拥有可预测的经常性收入。

3. 智能体与协作阶段：从模型调用走向任务交付。2026 年上半年，GLM 在智能体规划、工具调用、持续执行与长程任务上的能力持续提升，业务由单点的编程应用扩展到软件工程、网络安全、数据分析、复杂自动化等高价值场景。交易模式进一步由“卖调用”演进到“卖订阅与端到端任务成果”。报告期内，开放平台及 API 收入达人民币 8.25 亿元，同比增长超过 27 倍，占总收入约 86.5%（去年同期仅 15.2%，提升 71.3 个百分点）。**协作 (Co-work) 是当前正在解锁的部分：**模型接入企业数据、工具与业务系统，深入采购、财务、客服、研发、IT 运维等职能，从理解需求走向执行任务；反过来，企业的真实工作流也成为模型可学习的环境与数据。报告期内，公司在网络安全、法律、金融、教育等行业开展试点，其中网络安全进展最快，其他方向仍处早期，尚未形成规模收入。

回顾这三个阶段，形成一条主线：**模型能力每跃迁一档，客户购买的就更接近真实成果，公司的收入结构也随之重塑。**

(二) 量价齐升、毛利改善、亏损收窄 上述增长并不是以价格为代价的。截至本公告日期，MaaS 平台 Token 调用量较 2026 年初增长超过 40 倍，Coding Plan 调用量增长超过 23 倍；同期 API 平均售价上升约 101%，Coding Plan 订阅价格亦上调。

报告期内，开放平台及 API 业务毛利率提升至 24.6%，同比改善约 25 个百分点，较 2025 年提升约 6 个百分点。公司期内亏损同比下降 12.1%，亏损率（亏损/收入）较去年同期收窄 4.7 倍；期内经调整净亏损率（经调整亏损/期内收入）较去年同期收窄 3.5 倍。**在规模扩张的同时，单经济模型也在改善——这是我们更看重的一套指标。**

(三) 从编程到协作 编程是当前消耗量最大、验证最完整的场景。借助代码生成、工程理解、工具调用与长程任务能力，公司从代码辅助延伸至完整软件开发全生命周期。通过 Harness（智能体外壳），ZCode 承担上下文管理、工具调用、任务调度、缓存与结果验证的工作，提升模型在真实工程中的表现。随着长程能力推进，编程任务的复杂度上限也在持续抬高。

公司用三个维度来决定切入哪些协作方向：

- **智力门槛**——对推理、规划、工具调用与长程执行的要求；
- **单位智能消耗**——完成任务所需的有效 Token、推理时长、算力与人工介入；
- **单位产出价值**——单次交付带来的收入、节约的人工时间、降低的风险。

三者共同决定今天是否值得把任务交给模型。

面向大型企业与公共机构，公司继续提供本地化部署和组织级智能基础设施，满足数据安全、权限管理、系统集成与国产算力适配等需求。这些需求不会消失，但其在收入中的角色已由“主贡献者”转为“补充贡献者”。

(四) 用户规模与参与深度 截至本公告日期，公司 MaaS 平台用户超过 740 万，较 2026 年初增长 144%，付费日活用户增长 603%。**收入前十大客户的日均调用量较 2026 年初飙升 98 倍**。三个数字指向同一方向、量级递进：用户更多，付费用户涨得更快，重度付费用户用得更深。第三个数字尤为重要——它表明模型在客户组织内部正由“试用”走向“生产”。

(五) 算力乘数 (Compute Multiplier) 报告期内，公司“算力乘数”——即每 1 元人民币算力投入（含训练与推理）所产生的开放平台及 API 收入——**同比增长约 14 倍**。

这一曲线的改善来自两端：每个 Token 的智能在提升，每个任务消耗的 Token 与成本在下降。前者由模型能力决定，后者由基础设施决定。**只有两端同时改善，“智能-成本”前沿才能出现 Pareto 推进**。每 1 元算力投入回收的收入在增加，意味着模型的自给自足能力在增强。我们打算继续把这条曲线向上推。

2.5 生态与安全：为能力赢得现实世界的信任

在模型能力与“可依赖服务”之间，差一件事：现实世界是否信任它。公司同时在两条战线推进：开源开放，让安全能力与模型能力同步成长；通过权限控制、流程监管、风险评估、漏洞披露与外部验证，建立能力的可验证性。两者互相强化：**没有开源开放，安全就无法被检验；没有安全，开源开放也无法持续**。

安全不是单纯的合规成本，也可以成为一项可被验证的公共品，帮助市场判断模型能否在现实环境中安全、稳定、长期地部署。

(一) 从能力到公共品 自 GLM-5.2 发布以来，公司联合多家安全团队持续开展红队测试与安全评测。**截至报告期末，累计发现 2,436 个漏洞（经初筛与去重），其中中、高危 1,097 个，覆盖 269 个项目，包括系统内核、操作系统、浏览器引擎、开源基础组件、互联网应用与网络协议**。其中相当部分是长期未被发现的历史漏洞，最久的已存在超过 20 年。

这些发现已演变为持续性的漏洞发现与披露机制，相关漏洞陆续报送 CNNVD/CNVD 等国家级漏洞库。为提升透明度，公司建立**安全披露账本 (Security Disclosure Ledger)** (<https://cvd.z.ai/>)，记录从发现、确认到修复的全过程。账本持续更新，仅披露%22已修复且厂商已披露%22的漏洞；对尚处协同披露阶段

这一机制的意义在于，将模型能力置于第三方可验证的位置：基准跑分由我们自己报；漏洞披露由国家级漏洞库与厂商确认，外部信任的源头是后者。

(二) 开源与开放 随着 GLM-5.3 的发布，公司同步启动“开源护盾”行动，包含三项持续承诺：

1. **对关键开源项目做持续安全审计**，免费帮助维护者识别并修复风险，审计结果通过 OpenVuln (<https://huggingface.co/spaces/zai-org/OpenVuln>) 公开；
2. **为开源社区提供免费模型配额与防御接入**，允许开源项目维护者将这些资源用于安全审计与防御任务；
3. **在 ZCode 中提供代码审计功能**，把安全检查嵌入日常研发流程。

三者共通之处在于，把公司最先进的能力交到“最需要但最买不起”的人手中。开源基础组件支撑着几乎所有现代软件，而维护者常常只是几个不拿薪的个人。当攻击者使用最尖端模型的成本快速下降时，防御者反而可能用不起。如果最强的矛被锁在少数人手里，**最好的盾必须属于所有人**。

开源之前，公司与国内顶尖安全团队一道做了持续的红队测试与自动化越狱评估，用新生成的样本迭代加固审查系统，并邀请专业团队独立评估模型的高风险任务能力。**能力释放与风险缓释之间的平衡由外部判断，不由公司自己判断。**

从商业角度看，开源与安全不是成本项：开源带来的分发规模，是模型获得真实任务反馈的前提；安全披露带来的外部确认，是企业客户决定是否将模型接入生产系统的依据。

(三) 能力边界与受控访问 模型在网络安全上的能力具有双重用途。对于最敏感的能力，公司采用受控访问策略：仅向经过验证的安全研究人员与机构开放，其他用户遵循标准安全策略。开源前，公司完成安全评估、模型加固与使用边界设计。

2.6 未来展望 (Future Outlook)

2026 年上半年，我们已验证模型能独立完成一整项任务。下一步是让它承担一整条业务流程。

(一) 任务：从长程任务到自主系统 更长时间跨度。我们会让模型在更长时间内保持目标一致，持续进行任务分解、工具调用、环境交互、错误修复与结果验证。当下的前沿是小时级工程任务，下一步将是按天级别的完整交付。

多智能体协作。我们将从单一智能体走向多智能体协作，让模型承担完整业务流程，从”辅助工作”走向”执行工作”。这一步的门槛不完全在智能本身，更多在于可靠性与可审计性——模型必须能在无人逐步审视的情况下运行，同时留下可验证的痕迹。

每向外扩展一次任务边界，对模型能力的要求呈数量级提升。随之而来的问题是：能力还可在哪些维度上继续提升？

(二) 能力：Scaling 永不停步 Scaling 仍是通向更高智能的主要路径。变化的是我们 Scaling 的对象。

过去几年，Scaling 几乎等同于参数规模的扩大。然而，当参数跨过”装下世界”的门槛后，进一步扩宽带来的边际回报迅速递减，其他维度的潜力远未挖尽。今天决定模型上界的是四个因素的组合：**基座模型规模、有效深度、训练深度与任务环境**。它们的边际回报不同、在不同阶段轮换；公司的资源分配是判断在每个阶段哪个因素当前边际回报最高。

基座模型规模：智能上界由预训练决定。后期训练是当前边际回报最高的旋钮，但智能的天花板由基座模型决定。模型能练到多深，取决于预训练时见过多少世界、留下多少可激活的结构。当后期训练被推到极致，下一阶段的增长必须回到基座模型。公司从未停止基座模型的迭代。GLM-5.3-Flash 已经使用新一代架构与 30T 多模态预训练语料，验证了新架构在中等参数规模上的效率。在此基础上，公司正推进下一代基座模型的训练，目标包括更大有效规模、更长原生上下文、统一原生多模态建模。

有效深度：让模型多想几步。参数决定能记住多少，深度决定能想多远。Loop Transformer 通过层级回收在不大幅增加参数的情况下提高有效计算深度，把算力由”记忆”转向”思考”（公司已在早期实验中初步验证了这一思路）。这条路径与我们对任务的判断一致：网络安全这类工作需要的是”把 20 步因果链推到底而不中断”的能力，来自深度而非参数。

训练深度：当前最大的未挖尽潜力。GLM-5.3 是这一判断的直接证据。它与 GLM-5.2 采用相同架构、相同总参数与激活参数，唯一变量是后期训练规模，端到端完成率提升超过 50%。这条曲线尚未触顶。公司将继续扩大长程任务的强化学习规模，延长有效训练轨迹，提升轨迹利用效率。

任务环境：新出现的一个变量。当算法与算力不再是唯一瓶颈时，环境的数量、类型与复杂度开始直接决定模型能学到什么。这是过去两年才浮现的一个维度，是公司下一阶段最集中的投入方向之一。

(三) 系统：全栈自训练 四个旋钮里，**环境构建的速度正成为新的瓶颈**。更深一层，已有早期迹象表明 AI 正在接管训练系统的全部三块：**数据、环境与基础设施**。

数据开始自生产。人类产生的高质量数据正在见顶，公域数据几近耗尽，而下一代基座模型需要数万亿 Token 的、质量不断提升的数据。我们的方案是模型间自博弈 (model-vs-model self-play)：一模型生成解题与内容，另一模型评审、测试、执行验证——只有通过对抗性检验的数据才能进入下一阶段。再叠加四层过滤：规则校验确认可编译性、推理一致性、格式合规；执行验证在沙箱中真跑，跑出来的结果本身就是最高质量的训练信号；模型评审再加一层判断；最后人工抽检校验。沉淀下来的是可验证、可执行、可调试的数据，反哺预训练、中训练、后期训练三个阶段。

由此，**数据质量的上限不再由人类标注速度决定，而由模型自身的验证能力决定**。模型越强，它生成的题目越难、验证越严、数据越好；更好的数据再训练更强的下一代模型。数据由一次性消耗品，转化为可跨代复用的资本资产。

环境开始自构建。一个有用的训练环境必须“可执行、可验证、接近真实专业工作”，而且必须有几千个——手工做不来。在 GLM-5.3 技术博客中，我们分享了部分做法：研究智能体从真实工作里捕获任务模式，把原本需要资深工程师几天的工作转化为包含多步依赖与隐藏状态的可运行长程环境；评判智能体逐一试跑，确认可解；验证器在无标准答案的情况下合成，再用解题轨迹检测并封堵奖励漏洞；通过 oracle、no-op、unsolvable 检查的验证器，最终输出可用于强化学习的二元奖励。Slime 架构把数学、代码、沙箱与验证器变成可插拔的数据组件，让环境像流水线上的标品一样生产。

需要说明的是，这套流水线仍涉及相当多的人工介入。我们的下一步是让环境构建与验证更加自主。

基础设施开始自优化。推理系统优化——算子、调度、量化、调度器——本质上是“代码工程”任务，而那恰恰是模型最擅长的领域。GLM-5.3-Flash 的推理引擎是这样搭起来的：GLM-5.3 驱动的基础设施智能体协助工程师开发并优化算子、诊断性能瓶颈、改进服务栈，最终在国产芯片上把端到端服务性能提升 300%。同样的逻辑在训练侧也在运转：Slime 让训练与大规模推理 rollout 共用同一数据流，强化学习积累的配置经验与调度策略可直接迁移到生产服务。

这三条线都指向同一个终局：**模型参与到自身的改进中，形成递归自我提升**。这不是个遥远概念——其早期形态已在 2026 年上半年显现：GLM-5.3 参与了为 GLM 自身服务的推理引擎优化，GLM 生成的环境在训练下一代 GLM。唯一的差别是，当下还有大量工程师站在这个回路里。每一次回路收紧，智能进步的速度都越来越由环境生长的速度决定。

这就是我们的技术愿景——**全栈自训练 (Fully Self Training)**。下一代 GLM，将在前一代 GLM 构建出的环境中训练。

(四) 边界与治理 公司的技术志向是模型能自动参与到训练系统的构建。但我们必须确保：能力评估与风险判断最终保持由人把关，邀请外部机构独立评估。2026 年上半年我们已经在网络安全领域这样做：**最敏感的能力受控访问；开源前经过第三方专业团队独立评估；漏洞经国家级漏洞库披露**。基准跑分由我们自己报；风险判断由外部确认。这种分工不会随能力提升而改变，只会随能力上行而被强化。

公司将持续在能力上限、推理效率、任务价值密度与安全治理之间寻求新的平衡，使更高级别的智能能稳定、可控地进入现实，让每个人都能受益于普惠的智能。

(以下财务回顾与报表部分继续在文件后部。)

三、财务回顾 (Financial Review)

以下讨论基于本中期业绩公告其他章节列示的财务数据及附注，应结合一并阅读。

3.1 关键经营表现项目分析

收入 报告期内，公司实现显著的收入增长。2025 年上半年收入为人民币 1.909 亿元，2026 年上半年收入为人民币 9.539 亿元，同比增长 **399.7%**。收入规模的扩张主要由云端部署业务的爆发式增长所驱动，标志着本集团的主要收入增长引擎由本地化部署业务切换至云端部署业务。

按部署方式划分：

云端部署 (Cloud-based Deployment)

云端部署收入由 2025 年上半年的人民币 0.291 亿元增长至 2026 年上半年的人民币 8.252 亿元，同比增长 **2,735.7%**。增长主要反映了公司持续提升面向高价值任务场景的模型能力，模型在长程任务能力上持续进步，使客户能在复杂场景中持续使用模型，从而带动单用户日均 Token 调用量大幅提升。同时，MaaS 平台的用户数持续扩大，”量价齐升”的势头延续。云端部署收入占总收入的比例由上年同期的 15.2% 提升至 86.5%，已成为本集团的主要收入来源。

本地化部署 (On-premise Deployment)

本地化部署收入由 2025 年上半年的人民币 1.618 亿元下降至 2026 年上半年的人民币 1.287 亿元，同比下降 **20.5%**。下降主要由于云端部署业务爆发式增长的同时，公司更注重本地化业务的质量与行业选择，为未来模型能力由通用场景向垂直行业的高价值、专业化任务场景发展打下坚实基础。

按部署方式划分的收入一览 (人民币千元)：

部署方式	2026 H1 收入	占比	2025 H1 收入	占比
云端部署	825,176	86.5%	29,100	15.2%
本地化部署	128,716	13.5%	161,777	84.8%
合计	953,892	100.0%	190,877	100.0%

按业务形态与核心产品条线划分：

1) 开放平台及 API (Open Platform and API) — 基于通用大模型提供云服务

收入由 2025 年上半年的人民币 0.291 亿元增长至 2026 年上半年的人民币 8.252 亿元，同比增长 **2,735.7%**。增长主要源于公司逐步提升面向高价值任务场景的模型能力，以及 GLM 系列模型多轮迭代所带来的长程任务能力持续提升与质量升级，使用户在复杂场景中持续调用模型，从而带动单用户日均 Token 调用量大幅提升。同时 MaaS 平台用户基数持续扩大，”量价齐升”势头延续。

2) 企业级智能体 (Enterprise-level Agents) — 为企业复杂场景开发的自主智能系统

收入由 2025 年上半年的人民币 0.137 亿元增长至 2026 年上半年的人民币 0.556 亿元，同比增长 **304.4%**。一方面，GLM 系列模型在长程任务与复杂场景下的执行能力持续提升，使面向企业工作流的智能体产品能覆盖更多高价值场景；另一方面，报告期内企业客户对智能体场景化部署需求快速增长，共同推动本条线收入提升。

3) 企业级通用大模型 (Enterprise-level General-purpose Large Models) — 主要通过私有化本地部署交付

收入由 2025 年上半年的人民币 1.476 亿元下降至 2026 年上半年的人民币 0.670 亿元，同比下降 **54.6%**。下降主要由于公司更注重企业级通用大模型业务的质量与行业选择，为未来模型能力由通用场景向垂直行业的高价值、专业化任务场景发展打下基础。与此同时，部分企业级通用大模型客户逐步迁移至开放平台及 API 服务，形成正飞轮效应。

4) 技术服务及其他 (Technical Services and Others)

收入由 2025 年上半年的人民币 48.3 万元增长至 2026 年上半年的人民币 611.8 万元，同比增长约 **1,120.0%**，主要由于公司探索将自研大模型能力延伸至更多应用场景。本条线收入基数较小，其变化对本集团整体收入结构影响有限。

毛利与毛利率 本集团毛利由 2025 年上半年的人民币 0.954 亿元增至 2026 年上半年的人民币 2.516 亿元，同比增长 **163.7%**。毛利率由 2025 年上半年的 50.0% 下降至 2026 年上半年的 26.4%，主要由于云端部署业务占比迅速扩大——该业务尚处高速增长早期，毛利率仍处爬坡阶段，结构性稀释了整体毛利率。但云端业务自身的毛利率由负转正，同比迅速提升。报告期内本集团毛利增长超过 1.6 倍，反映收入结构变化对毛利的正向贡献。

按部署方式 (人民币千元):

部署方式	2026 H1 毛利	毛利率	2025 H1 毛利	毛利率
云端部署	202,871	24.6%	(116)	-0.4%
本地化部署	48,740	37.9%	95,540	59.1%
合计	251,611	26.4%	95,424	50.0%

- **云端部署**：毛利率由 -0.4% 提升至 24.6%。主要驱动：① 收入爆发式增长带来的规模经济，Token 用量指数级增长显著摊薄固定成本；② 随模型智能上界提升而上调定价，Token 用量持续增长，量价齐升的势头延续；③ 持续的全栈自研工程优化降低了模型推理成本。
- **本地化部署**：毛利由 0.955 亿元降至 0.487 亿元，同比下降 49.0%；毛利率由 59.1% 降至 37.9%。主要由于公司主动推进业务模式调整——本地化部署业务产品组合发生变化：高毛利率的企业级通用大模型私有交付规模收缩，处于大规模扩张阶段的企业级智能体产品占比提升；同时部分客户由本地化部署迁移至云端服务。

销售成本 销售成本由 2025 年上半年的人民币 0.955 亿元增至 2026 年上半年的人民币 7.023 亿元，同比增长 **635.4%**。增长主要源于业务扩张与收入增长带来的算力服务费增加，与收入增长趋势基本一致。

资本性支出 资本性支出由 2025 年上半年的人民币 0.537 亿元增至 2026 年上半年的人民币 3.906 亿元，同比增长 **627.4%**。增长主要由于公司完成对北京红钻科技发展有限公司 100% 股权的收购，新增大额物业、厂房及设备。

其他收入 其他收入由 2025 年上半年的人民币 460 万元增至 2026 年上半年的人民币 4,140 万元，同比增长约 **800.0%**，主要由于公司于 2026 年 1 月上市后银行存款规模大幅增加，相应带来利息收入增加。

研发开支 研发开支由 2025 年上半年的人民币 15.95 亿元增至 2026 年上半年的人民币 21.31 亿元，同比增长 **33.6%**，主要由于公司持续加大对模型训练的研发投入，以支持 GLM 系列模型智能上界的持续突破。

销售及营销开支 销售及营销开支由 2025 年上半年的人民币 2.086 亿元降至 2026 年上半年的人民币 1.777 亿元，同比下降 **14.8%**，主要由于公司优化营销资源配置、提升营销投入的转化效率。

一般及行政开支 一般及行政开支由 2025 年上半年的人民币 1.852 亿元降至 2026 年上半年的人民币 1.033 亿元，同比下降 **44.2%**，主要由于报告期内专业服务费及权益结算的股份支付费用较去年同期减少。

金融资产减值损失 金融资产减值损失由 2025 年上半年的人民币 1,090 万元增至 2026 年上半年的人民币 2,740 万元，同比增长 **151.4%**，主要由于业务规模扩张使应收账款及应收票据总额增加。

财务成本 财务成本由 2025 年上半年的人民币 5,330 万元增至 2026 年上半年的人民币 9,590 万元，同比增长 **79.9%**，主要由于报告期内确认汇兑净损失及银行借款利息支出增加。

应占联营公司损益 应占联营公司损益由 2025 年上半年的人民币 1,410 万元增至 2026 年上半年的人民币 1.442 亿元，同比增长 **922.7%**，主要由于报告期内联营公司经营及财务业绩改善。

按 FVPL 计量的金融工具公允价值变动 按公允价值计入损益 (FVPL) 的金融工具公允价值变动由 2025 年上半年的人民币 980 万元增至 2026 年上半年的人民币 5,040 万元，同比增长 **414.3%**，主要由于报告期内公司持有的金融工具估值上升，确认公允价值收益。

向投资者发行的金融工具账面金额变动 向投资者发行的金融工具账面金额变动在 2025 年上半年为损失人民币 4.293 亿元，2026 年上半年减至损失人民币 0.221 亿元，同比减少 **94.9%**。变动主要由于 2026 年 1 月上市时终止向投资者发行的金融工具及相关赎回权并予以终止确认，报告期内仅就上市前较短期间确认账面金额变动。

净亏损 本集团于 2026 年上半年录得净亏损人民币 20.72 亿元 (2025 年上半年：净亏损人民币 23.58 亿元)，同比减少。减少主要由于报告期内毛利增加，以及上市时向投资者发行的金融工具转换为普通股、相关损失大幅减少，研发开支的增加部分抵消了上述减亏。

非 IFRS 指标：经调整期内净亏损 (人民币千元)

项目	2026 H1	2025 H1
期内亏损	(2,071,992)	(2,357,852)
加：权益结算的股份支付费用	85,764	158,852
加：向投资者发行的金融工具账面金额变动	22,090	429,295
加：上市费用	–	17,731
经调整期内净亏损 (非 IFRS)	(1,964,138)	(1,751,974)

3.2 关键资产负债项目分析

财务状况 股东权益由截至 2025 年 12 月 31 日的人民币 81.11 亿元增至截至 2026 年 6 月 30 日的人民币 44.30 亿元，主要由于报告期内本集团成功上市，相应终止向投资者发行的金融工具及相关赎回权，相关金融负债转入本公司权益。

流动性与财务资源 截至 2026 年 6 月 30 日，本集团持有的现金及现金等价物为人民币 39.94 亿元，较 2025 年 12 月 31 日的人民币 22.59 亿元增加 17.35 亿元。该波动主要归因于报告期内收到上市所得款项及银行借款增加，同时运营开支消耗了部分资金。

债务/借款 截至 2026 年 6 月 30 日，本集团银行借款总额为人民币 22.25 亿元，较 2025 年 12 月 31 日的人民币 6.90 亿元增加 15.35 亿元，主要因满足运营需要而借入资金。所有相关借款均为计息贷款，主要以人民币计价且无担保。

租赁负债 截至 2026 年 6 月 30 日，本集团确认的租赁负债总额（含流动与非流动部分）为人民币 3.794 亿元，较 2025 年 12 月 31 日的人民币 5.451 亿元减少 1.657 亿元。该波动主要归因于报告期内支付租赁款项及租赁安排发生变化。

资产负债率 本集团的资产负债率（按银行借款及租赁负债除以期末总权益再乘以 100% 计算）截至 2026 年 6 月 30 日为 **58.8%**，截至 2025 年 12 月 31 日为 **-15.2%**。变化主要由于上市时终止向投资者发行的金融工具及相关赎回权，相关金融负债转入本公司权益，使总权益由负转正。

资产抵押、资本承担、或有负债 截至 2026 年 6 月 30 日，本集团无任何资产抵押、未偿付的资本承担或重大或有负债。

重大投资、子公司/联营/合营企业的重大收购及出售 根据 2026 年 3 月 31 日的董事会决议及 2026 年 4 月 14 日的收购框架协议，公司以不超过人民币 3.6052 亿元的总对价收购北京红钻科技发展有限公司的 100% 股权。收购完成后，北京红钻已作为公司的子公司入账。

除上述披露外，截至 2026 年 6 月 30 日，本集团未进行任何其他重大投资，未参与任何子公司、联营公司或合营企业的其他重大收购或出售。截至本公告日期，本集团并无重大资本资产收购的具体未来计划。

员工与薪酬政策 截至 2026 年 6 月 30 日，本集团共有 **981 名** 永久员工（2025 年 6 月 30 日：883 名），人员配置视运营需要而定。报告期内本集团承担的总薪酬成本（含股份支付）为人民币 4.821 亿元（2025 年上半年：人民币 4.996 亿元）。

人才是公司最核心的战略资源。结合人工智能行业特点，公司构建了能力导向、激励驱动的人才管理体系，以支撑公司在 AI 领域的持续技术创新与业务拓展。

- **薪酬策略：**采用“固定薪酬 + 绩效奖金 + 中长期激励”的三层结构，确保奖励与公司业绩及个人贡献紧密关联。通过员工激励计划与员工持股平台，将核心管理人员、关键技术骨干与业务专才的利益与公司长期发展绑定。
- **员工关怀：**除法定社保与公积金外，额外提供补充商业保险、年度健康体检、餐补、节日福利等多重福利；支持各类文体活动，关注员工身心健康，营造积极向上的工作氛围。

- **AI 原生工作模式：**作为大模型行业的领军企业，公司把 AI 原生工作模式确立为组织运营的根本方式。依托自研模型与智能体系统，将自身定位为大模型能力在真实生产环境中大规模验证的前沿环境。在日常研发中，代码生成、缺陷修复、测试编写等标准化流程大多由 AI 自主完成，员工聚焦问题本质、技术方案设计与创新方向探索。”AI 干活、人做决策”的协同范式，正在重塑公司的生产力与创新速度。

汇率波动风险 本集团的外汇风险源于相关集团实体功能货币以外货币计价的未来商业交易及已确认的资产与负债。本公司功能货币为人民币。绝大部分非人民币资产负债为以美元和港元计价的现金及现金等价物。

本集团主要面临人民币兑美元、人民币兑港元的汇率波动风险。目前，公司未开展旨在管理外汇风险的套期活动，但通过密切监控汇率走势管理外汇风险，必要时会采取相应措施加以缓释。

主要风险与不确定因素 公司经营涉及若干风险与不确定因素，包括：① 与研发相关的风险；② 与商业化相关的风险；③ 与运营相关的风险；④ 与知识产权相关的风险；⑤ 与财务状况及额外资金需求相关的风险；⑥ 与公司开展业务所在的司法辖区相关的风险。重点包括但不限于：

- 人工智能行业是瞬息万变的前沿技术领域，公司未来的业务运营、经营业绩、财务状况与竞争地位取决于其模型与技术持续升级迭代的能力。
- 公司已并预计将继续在研发上投入大量资金，以持续升级迭代其大模型。
- 公司面临客户变动与新增客户发展慢于预期的风险。
- AGI 的发展仍处早期，其未来实现存在重大不确定性。
- 公司需长期维持一支充足、稳定的技术团队，及时响应市场变化。任何与技术骨干相关的不利变动，都可能影响公司重大研发项目的开发、推进和商业化，进而对公司经营产生不利影响。
- 公司面临业务扩张带来的管理风险。资产与收入规模快速增长，同时人员持续增加，对公司管理能力与风险控制提出更高要求，公司可能一定程度上面临业务扩张带来的管理风险。
- 公司依赖第三方提供算力资源，其服务的任何中断或价格波动都可能对公司业务、经营业绩及财务状况产生不利影响。
- 中国 AI 行业所处的监管环境在持续演变并趋严，未来法律法规可能提出额外要求与义务，可能对公司业务、经营业绩、财务状况及前景产生不利影响。
- 公司可能无法为其业务获得或维持充分的知识产权保护，或该等保护的覆盖范围不够宽。
- 公司面临未来市场竞争加剧的风险。若公司技术与产品迭代无法保持在行业领先水平，公司的竞争地位、市场份额与经营业绩可能受到不利影响。
- 受多种因素影响，可触达市场规模与解决方案的需求增长速度可能不如预期，进而对公司业务、经营业绩、财务状况及前景产生不利影响。
- 若公司 MaaS 平台未能按要求运行，将损害公司的业务、经营业绩、财务状况及前景。
- 公司商业化业绩记录有限；大模型市场同时受技术迭代与监管政策变化的影响，难以准确预测发展机会与未来市场规模。

上述列示并不详尽，更多详情请参见招股书”风险因素”一节。

四、简明合并财务报表 (Consolidated Financial Statements)

4.1 合并损益及其他全面收益表

截至 2026 年 6 月 30 日止六个月 (未经审核)·单位：人民币千元

项目	附注	2026 H1	2025 H1
收入 Revenue	3	953,892	190,877
销售成本 Cost of revenue		(702,281)	(95,453)
毛利 Gross profit		251,611	95,424
其他收入 Other income	3(b)	41,436	4,614
销售及营销开支		(177,666)	(208,570)
一般及行政开支		(103,346)	(185,165)
研发开支		(2,131,171)	(1,594,661)
金融资产减值损失		(27,416)	(10,867)
经营亏损 Loss from operations		(2,146,552)	(1,899,225)
财务成本 Finance costs		(95,925)	(53,270)
应占联营公司损益		144,152	14,147
按 FVPL 计量的金融工具公允价值变动		50,385	9,791
向投资者发行的金融工具账面金额变动		(22,090)	(429,295)
除税前亏损 Loss before taxation		(2,070,030)	(2,357,852)
所得税开支 Income tax	4	(1,962)	–
期内亏损 Loss for the period		(2,071,992)	(2,357,852)

期内其他全面收益（税后）：

- 将重分类至损益的项目：境外财务报表折算差额 (3,049) (2025 H1: 414)

期内全面收益总额： (2,075,041) (2025 H1: (2,357,438))

期内亏损归属：

- 本公司权益股东：(2,071,326) (2025 H1: (2,351,173))
- 非控制权益：(666) (2025 H1: (6,679))

期内全面收益总额归属：

- 本公司权益股东：(2,074,375) (2025 H1: (2,350,759))
- 非控制权益：(666) (2025 H1: (6,679))

每股亏损（基本与摊薄）： 人民币 (4.67) (2025 H1: (6.23))

4.2 合并财务状况表（Consolidated Statement of Financial Position）

截至 2026 年 6 月 30 日（未经审核）·单位：人民币千元

项目	附注	2026/06/30	2025/12/31
非流动资产			
物业、厂房及设备	6	855,060	655,825
无形资产		67,801	50,963
商誉		283,815	39,379
联营公司权益		532,133	338,081
其他非流动资产		912,325	198,202

项目	附注	2026/06/30	2025/12/31
非流动资产合计		2,651,134	1,282,450
流动资产			
按 FVPL 计量的短期投资		506,139	377,287
存货及合同成本		87,100	125,817
应收账款及应收票据	7	782,944	303,208
预付款项及其他应收款	8	1,621,311	395,734
合同资产		2,480	1,625
定期存款		–	108,593
银行存款及现金		3,993,722	2,259,147
流动资产合计		6,993,696	3,571,411
流动负债			
应付账款及其他应付款	9	2,135,759	1,328,631
合同负债		302,325	148,644
银行借款		789,831	604,836
租赁负债		161,711	251,316
向投资者发行的金融工具		–	10,072,827
流动负债合计		3,389,626	12,406,254
流动资产净额／（净流动负债）		3,604,070	(8,834,843)
资产总额减流动负债		6,255,204	(7,552,393)
非流动负债			
银行借款		1,434,958	84,616
租赁负债		217,699	293,827
递延收入		172,991	180,146
非流动负债合计		1,825,648	558,589
净资产／（净负债）		4,429,556	(8,110,982)
权益			
股本		44,584	40,281
储备		4,372,464	(8,132,783)
本公司权益股东应占总权益／（赤字）		4,417,048	(8,092,502)
非控制权益		12,508	(18,480)
权益总额／（赤字总额）		4,429,556	(8,110,982)

五、合并财务报表附注（Notes to the Condensed Consolidated Interim Financial Statements）

附注 1：编制基础

Z.AI Co., Ltd.（原名“Knowledge Atlas Technology Joint Stock Company Limited”）于 2019 年 6 月 11 日在中华人民共和国成立为有限责任公司。2025 年 3 月 26 日，公司由有限责任公司变更为股份有限公司。公司股份于 2026 年 1 月 8 日在香港联合交易所有限公司主板上市。

公司及其子公司（合称“本集团”）主要从事提供大模型相关服务。

中期财务报表根据联交所证券上市规则的适用披露规定编制，包括遵守国际会计准则理事会（IASB）颁布的《国际会计准则第 34 号——中期财务报告》（“IAS 34”）。报告于 **2026 年 8 月 31 日** 经授权刊发。

中期财务报表按与 2025 年度财务报表所采纳的相同会计政策编制，惟不包括预期将于 2026 年度财务报表内反映的会计政策变更。任何会计政策变更的详情载于附注 2。

按 IAS 34 编制中期财务报表要求管理层作出判断、估计和假设，这些判断、估计和假设会影响政策的应用与本年止当日的资产负债、收入与费用的报告金额。实际结果可能与这些估计存在差异。

中期财务报表包含简明合并财务报表与若干解释性附注。附注包括对理解自 2025 年度财务报表以来本集团财务状况与业绩变化有重要影响的事件与交易的说明。简明合并中期财务报表及其附注并不包含根据 IFRS 会计准则编制完整财务报表所要求的全部信息。

附注 2：会计政策变更

IASB 已颁布多项新订及经修订的 IFRS 会计准则，于本会计期间首次生效。其中，仅《IFRS 9—金融工具》及《IFRS 7—金融工具：披露》的修订（金融工具分类与计量的修订）与本集团财务报表相关。应用该等修订对本集团于列报期间的合并财务报表并无重大影响。

本集团未应用任何尚未于本会计期间生效的新准则或诠释。

该等修订亦会影响本集团截至 2026 年 12 月 31 日止年度的年度财务报表中关于指定为 FVOCI 的权益工具投资及不按 FVPL 计量且具备特定或有特征金融工具的披露。中期财务报表中未载列额外披露。

附注 3：收入与分部报告

3(a) 收入 本集团主要业务为提供大模型相关服务。

(i) 收入分解

按主要服务类型及收入确认时点划分（人民币千元）：

项目	2026 H1	2025 H1
IFRS 15 范围内与客户的合同收入		
按主要服务类型：		
–本地化部署	128,716	161,777
–云端部署	825,176	29,100
小计	953,892	190,877
按收入确认时点：		
–在某一时点确认	854,270	155,636
–在某一时段内确认	99,622	35,241
合计	953,892	190,877

3(b) 分部报告 本集团按服务类型管理业务。根据向本集团最高级执行管理层内部报告信息的方式（用于资源分配与业绩评估），本集团呈列以下两个可呈报分部。报告期内没有为形成以下可呈报分部而汇总经营分部。

- **本地化部署**：根据客户的具体指示与需求，在客户基础设施上开发并提供定制化大模型相关服务。
- **云端部署**：通过云基础设施向客户开发并提供云端大模型相关服务。

(i) 分部业绩

为评估分部业绩和分配资源，本集团最高级执行管理层按以下基础监督各可呈报分部的业绩：收入与费用按各分部产生的收入与各分部发生的直接费用分配至各可呈报分部；分部业绩计量的指标为毛利。报告期内并无分部间销售。一个分部对另一个分部提供的协助（包括资产与技术诀窍的共享）不计量。

本集团的其他经营收入与费用（如其他收入、销售及营销开支、一般及行政开支、研发开支、金融资产减值损失、财务成本，以及资产与负债）不按个别分部计量。因此，本集团既不呈列分部资产与负债信息，亦不呈列有关资本性支出、利息收入与利息开支的信息。

报告期内向本集团最高级执行管理层提供用于资源分配与分部业绩评估的可呈报分部信息载列如下。

2026 H1 分部业绩（人民币千元）：

分部	本地化部署	云端部署	合计
分部收入（来自外部客户）	128,716	825,176	953,892
分部毛利	48,740	202,871	251,611

2025 H1 分部业绩（人民币千元）：

分部	本地化部署	云端部署	合计
分部收入（来自外部客户）	161,777	29,100	190,877
分部毛利	95,540	(116)	95,424

(ii) 可呈报分部损益的调节（人民币千元）：

项目	2026 H1	2025 H1
可呈报分部毛利合计	251,611	95,424
其他收入	41,436	4,614
销售及营销开支	(177,666)	(208,570)
一般及行政开支	(103,346)	(185,165)
研发开支	(2,131,171)	(1,594,661)
金融资产减值损失	(27,416)	(10,867)
财务成本	(95,925)	(53,270)
应占联营公司损益	144,152	14,147
按 FVPL 计量的金融工具公允价值变动	50,385	9,791
向投资者发行的金融工具账面金额变动	(22,090)	(429,295)
合并除税前亏损	(2,070,030)	(2,357,852)

附注 4：合并损益及其他全面收益表中的所得税

所得税列示（人民币千元）：

项目	2026 H1	2025 H1
当期所得税	1,962	-
递延所得税	-	-
合计	1,962	-

报告期内，在中国大陆成立的集团实体适用 **25%** 中国企业所得税税率（2025 年上半年：25%）。

在其他司法管辖区注册的子公司之所得税按相关司法管辖区之适用所得税率计算。报告期内本集团部分子公司已获得税务机关颁发的”高新技术企业”证书，享受 15% 的优惠税率。根据中国所得税法及相关规定，符合小型微利企业条件的实体，对应纳税所得额低于人民币 300 万元的部分享受 5% 的优惠所得税率；本集团部分子公司符合小型微利企业条件并于 2026 H1 和 2025 H1 享受该优惠税率。

附注 5：每股亏损

基本每股亏损按本公司普通权益股东应占亏损除以报告期内已发行或视同已发行普通股的加权平均数计算。

公司于 **2025 年 3 月 26 日** 变更为股份有限公司。公司实缴资本人民币 36,224,375 元相应转换为每股面值人民币 1.00 元的 36,224,375 股。为计算基本与摊薄每股亏损，公司变更为股份有限公司之前的视同已发行普通股加权平均数，假设该变更自 2025 年 1 月 1 日起即按 2025 年 3 月变更时确定的转换比例发生。

此外，报告期内普通股的加权平均数还就公司股份于 **2026 年 1 月 8 日** 上市完成前已生效的股份拆细的影响进行追溯调整。

每股亏损计算（人民币千元）：

项目	2026 H1	2025 H1
普通权益股东应占期内亏损（附注 5(a)）	(2,020,788)	(993,276)
普通股加权平均数（千股，附注 5(b)）	432,359	159,522
基本每股亏损（人民币元）	(4.67)	(6.23)

(a) 普通权益股东应占期内亏损（人民币千元）：

项目	2026 H1	2025 H1
本公司全体权益股东应占期内亏损	(2,071,326)	(2,351,173)
减：向投资者发行的金融工具应占期内亏损之分配	50,538	1,357,897
普通权益股东应占期内亏损	(2,020,788)	(993,276)

(b) 普通股加权平均数（千股）：

项目	2026 H1 股份数	2025 H1 股份数
1 月 1 日已发行普通股	40,281	36,224
首次公开发行的股份影响	40,361	-
视同已发行普通股的影响	-	1,536

项目	2026 H1 股份数	2025 H1 股份数
向投资者发行的金融工具的影响	(1,081)	(21,808)
股份拆细的影响	352,798	143,570
6 月 30 日普通股加权平均数	432,359	159,522

(c) 摊薄每股亏损

由于纳入计算会产生反摊薄效应，向投资者发行的金融工具未纳入摊薄每股亏损的计算。因此，截至 2026 年 6 月 30 日与 2025 年 6 月 30 日止六个月的摊薄每股亏损与基本每股亏损相同。

附注 6：物业、厂房及设备

自有资产购置 2026 年 5 月，公司完成对北京红钻全部股权的收购，总对价为人民币 360,226,000 元，包括：(1) 所收购股权的现金对价人民币 79,158,000 元（附注 9）；（2）用于协助偿还北京红钻相关债务的款项人民币 281,068,000 元。截至 2026 年 8 月 21 日，对价的全部款项已支付完毕。

公司选择采用《IFRS 3 —企业合并》允许的集中度测试，并认定收购北京红钻属于资产收购而非企业合并。于收购完成日，自北京红钻收购的可识别资产与负债如下（人民币千元）：

项目	公允价值
物业、厂房及设备	360,000
预付款项及其他应收款	869
已收购现金及现金等价物	49
应付账款及其他应付款	(692)
已付／应付对价合计	360,226

附注 7：应收账款及应收票据

项目	2026/06/30	2025/12/31
应收账款	838,139	339,198
应收票据	6,445	—
减：损失准备	(61,640)	(35,990)
账面价值	782,944	303,208

按发票日期的账龄分析（净额，人民币千元）：

账龄段	2026/06/30	2025/12/31
3 个月内	626,201	243,484
3 个月至 6 个月	23,697	19,831

账龄段	2026/06/30	2025/12/31
6 个月至 1 年	109,486	31,406
1 年至 2 年	23,538	8,415
2 年至 3 年	22	72
合计	782,944	303,208

本集团一般要求本地化部署及相关服务的客户按进度结算账单；云端部署客户在按使用量计费及相关服务合同下须于发票日后 0-60 天内付款，或在订阅合同及相关服务下按周期结算。

附注 8：预付款项及其他应收款

项目	2026/06/30	2025/12/31
押金	78,404	70,538
处置按 FVPL 计量的权益证券投资应收款	–	7,098
其他应收款	110,379	52,029
减：损失准备	(20,923)	(18,994)
小计	167,860	110,671
预付算力服务费及其他	1,243,736	116,295
待抵扣进项税	209,715	168,768
小计	1,453,451	285,063
合计	1,621,311	395,734

附注 9：应付账款及其他应付款

项目	2026/06/30	2025/12/31
应付账款	241,589	297,139
应付算力服务费	1,427,415	727,348
应付营销及推广服务费	51,004	42,812
应付职工薪酬	122,923	146,507
应付企业合并及资产收购对价（附注 6、11）	229,809	–
其他应付款及应计费用	26,869	54,437
按摊余成本计量的金融负债	2,099,609	1,268,243
其他应交税费	17,515	39,774
质量保证拨备	18,635	20,614
合计	2,135,759	1,328,631

按发票日期的应付账款账龄分析（人民币千元）：

账龄段	2026/06/30	2025/12/31
3 个月内	171,575	229,890
3 个月至 6 个月	15,254	46,448
6 个月至 1 年	50,208	19,506
1 年以上	4,552	1,295
合计	241,589	297,139

附注 10：股息

董事会不建议就截至 2026 年 6 月 30 日止六个月派付任何中期股息（2025 年上半年：无）。

附注 11：企业合并（XCore Sigma 收购）

2026 年 6 月 5 日，本集团（通过本公司）与北京中科加禾智能科技有限公司（“XCore Sigma”）的股东签订购买协议，以总现金对价人民币 291,916,000 元收购 XCore Sigma 60% 股权。根据购买协议，对价将分四期支付，2026 年 6 月已支付人民币 141,264,000 元。XCore Sigma 主要从事 AI 基础设施研发。该收购已于 2026 年 6 月完成。

收购日 XCore Sigma 可识别资产与负债的公允价值（人民币千元）：

项目	金额
物业、厂房及设备	1,309
无形资产	22,510
递延所得税资产	3,377
存货	3,934
应收账款及应收票据	1,373
预付款项及其他应收款	209
银行存款及现金	66,567
应付账款及其他应付款	(6,558)
合同负债	(9,906)
租赁负债	(304)
递延所得税负债	(3,377)
可识别净资产公允价值	79,134
减：非控制权益	(31,654)
加：商誉	244,436
收购对价合计	291,916

收购 XCore Sigma 的现金及现金等价物净流出分析（截至收购日，人民币千元）：

项目	金额
总现金对价	291,916
减：收购日取得的现金	(66,567)
减：应付对价（附注 9）	(150,651)
收购子公司产生的现金净流出（于简明合并现金流量表中列示）	74,698

附注 12：报告期后非调整事项

2026 年 7 月 13 日，公司完成配售并按每股 1,588 港元的价格发行 19,780,000 股普通股。本次配售所得款项总额约为 **314.1064 亿港元**（约合人民币 272.30826 亿元）。

六、其他信息 (Other Information)

6.1 中期股息

董事会不建议就截至 2026 年 6 月 30 日止六个月派付任何中期股息（2025 年上半年：无）。

6.2 上市证券的购买、出售或赎回

回购授权 (Repurchase Mandate)：经 2026 年 6 月 22 日股东决议通过，公司董事会获得一项一般授权，可不时在公开市场回购公司股份。该一般授权下，公司获授权回购不超过于 2025 年年度股东大会通过相关决议当日公司已发行 H 股股份总数的 10%。

公司 H 股于 **2026 年 1 月 8 日** 在香港联交所主板首次上市。自上市日起至 2026 年 6 月 30 日止，公司及其任何子公司概无购买、出售或赎回公司任何上市证券（含出售任何库存股份）。截至 2026 年 6 月 30 日，公司未持有《上市规则》含义内的任何库存股份。

6.3 报告期后重大事项

一般授权下配售新 H 股 2026 年 7 月 9 日（开市前），公司与中国国际金融香港证券有限公司（“独家配售代理”）订立配售协议（“2026 年 7 月配售”）。据此，公司同意委任独家配售代理、独家配售代理同意作为公司的代理，尽最大努力促使不少于 6 名承配人（专业、机构或其他投资者，其本人及最终实益拥有人均为独立第三方）按 **1,588.00 港元／股** 的配售价认购最多 **19,780,000 股新 H 股**。本次配售依据于 2026 年 6 月 22 日举行的公司年度股东大会授予董事的一般授权进行。

2026 年 7 月配售已于 **2026 年 7 月 13 日** 完成。2026 年 7 月配售所得款项净额（扣除佣金及估计开支后）约为 **313.7495 亿港元**。

除上述披露外，自报告期末至本公告日期，本公司并无其他重大事件。

6.4 企业管治守则合规

自上市日起至本公告日期，公司严格遵守《企业管治守则》所有适用条文。

6.5 标准守则合规

上市以来，公司已采纳标准守则作为规管董事及可能掌握内幕信息的员工进行公司证券交易的行为守则。经向全体董事作出特定查询后，全体董事确认自上市日起至 2026 年 6 月 30 日止期间均遵守标准守则。本集团可能掌握内幕信息的员工亦须遵守标准守则。自上市日起至 2026 年 6 月 30 日止期间，公司未发现任何相关员工违反标准守则的情况。

6.6 《上市规则》下的持续披露义务

除本公告披露外，公司无《上市规则》第 13.20、13.21、13.22 条所规定的其他披露义务。

6.7 公持量充足性

根据公开披露信息及董事所知，自上市日起至本公告日期，公司持续维持《上市规则》要求的最低公持量。

6.8 审核委员会

审核委员会由谢德仁博士、杨强博士及李涓子博士组成，谢德仁博士担任主席。审核委员会已审阅本集团截至 2026 年 6 月 30 日止六个月的中期业绩，以及按 IFRS 会计准则编制的本集团截至 2026 年 6 月 30 日止六个月的未经审核简明合并财务报表。董事会与审核委员会就报告期内所采纳的会计处理并无意见分歧。

6.9 中期财务资料审阅

本中期业绩公告所载财务资料摘录自已于 2026 年 6 月 30 日止六个月的中期财务报告（未经审核）。该中期财务报告已由毕马威会计师事务所根据香港会计师公会颁布的《香港审阅准则第 2410 号——由实体独立核数师审阅中期财务资料》进行审阅，其无保留意见的审阅报告将载于稍后寄发予股东的中期报告内。

6.10 中期业绩与中期报告的刊发

本业绩公告已登载于联交所网站 (www.hkexnews.hk) 及公司网站 (www.zhipuai.cn)。公司截至 2026 年 6 月 30 日止六个月的中期报告将于适当时候登载于上述网站，并按 H 股股东所选择的收取公司通讯方式寄发予 H 股股东。

七、术语表 (Definitions & Glossary)

7.1 定义 (Definitions)

术语	释义
审核委员会	董事会辖下的审核委员会
董事会	本公司董事会
企业管治守则	《上市规则》附录 C1 所载《企业管治守则》
中国／ PRC	中华人民共和国，仅就本公告而言，不包括香港、澳门特别行政区与台湾
公司／本公司	Z.AI Co., Ltd. (北京智谱华章科技股份有限公司)，于 2019 年 6 月 11 日依中国法律成立的有限责任
董事	本公司董事
全球发售	招股书所界定的香港公开发售与国际配售
本集团／我们	本公司及其子公司
H 股	本公司境外上市的股份，每股面值人民币 0.10 元，以港元交易，在联交所上市及买卖
港元	港元，香港法定货币
香港	中华人民共和国香港特别行政区
IFRS 会计准则	国际会计准则理事会颁布的 IFRS 会计准则
独立第三方	独立于本公司及其关联人士、与彼等概无关联的第三方
上市	H 股于 2026 年 1 月 8 日在联交所主板上市
上市日期	2026 年 1 月 8 日，H 股在联交所主板上市日期
上市规则	《香港联合交易所有限公司证券上市规则》
标准守则	《上市规则》附录 C3 所载《董事进行证券交易的标准守则》
招股书	公司于 2025 年 12 月 30 日就全球发售及上市刊发的招股书
人民币／ RMB	人民币，中国法定货币
报告期	截至 2026 年 6 月 30 日止六个月
股份	本公司股本中每股面值人民币 0.10 元的普通股，包括非上市股份与 H 股
股东	股份持有人
联交所／香港联交所	香港联合交易所有限公司
子公司	具有《上市规则》赋予的涵义
库存股份	具有《上市规则》赋予的涵义
非上市股份	本公司发行的每股面值人民币 0.10 元的、不在任何证券交易所上市的普通股
美元／ USD	美元，美国法定货币
%	百分比

7.2 技术术语表 (Glossary of Technical Terms)

术语	释义
智能体 (Agent)	能自主执行多步任务链的 AI 模型，可自行调用工具
智能体工程 (Agentic Engineering)	模型从简单代码生成演进到完整工程项目交付的工程方法
AGI 商业价值公式	公司用于衡量 AGI 商业价值的量化框架，公式为：商业价值 = 模型能力 - 成本
全栈基础设施战略 (All-in-Infra strategy)	公司自 2026 年初起全面实施的基础设施战略，覆盖从芯片到应用的全栈
算力乘数 (Computing Power Multiplier)	每 1 元人民币算力投入 (含训练与推理) 所产生模型能力的倍数
协作 (Co-work)	模型能力阶梯的第四级，模型交付” 专业人员使用模型完成复杂任务
对话 (Chat)	模型能力阶梯的第一级，模型单轮响应，任务简单且结构化
编程 (Coding)	模型能力阶梯的第二级，模型交付一段可运行代码
CyberGym	评估 AI 模型在网络安全领域能力的基准测试平台
ExploitGym	评估 AI 模型漏洞利用能力的基准测试平台
向纵深 Scaling (Scaling deep)	走深至后期训练、强化学习、长程轨迹与任务规划

术语	释义
向上 Scaling (Scaling up)	通过扩大模型规模、上下文长度、算力、系统
GLM-5.3-Flash (Ox-Alpha)	公司下一代轻量旗舰模型，测试代号 Ox-Alpha
GLM-5.3	公司旗舰大模型；与 GLM-5.2 采用相同架构、
中训练 (Mid-training)	公司在预训练与后期训练之间单独划拨算力的
基于评论器的 SAO (Critic-based SAO)	公司自研的基于评论器的单轨迹优势优化，由
OPD	公司在算法维度上由监督微调 (SFT) 扩展而
Index Share	公司自研技术，用于降低长上下文下稀疏注意
MTP	多 Token 预测 (Multi-Token Prediction) 技
LayerSplit	公司自研技术，用于优化 KV Cache 管理与集
ZCube	公司自研技术，与 LayerSplit 协同优化 KV C
mHC (流形约束超连接, manifold-constrained Hyperconnection)	GLM-5.3-Flash 采用的架构创新，全称为”流
Encode-Prefill-Decode 分离架构	公司在生产级集群层面采用的推理架构，将多
Loop Transformer	通过循环复用层以在不大幅增加参数的情况
国产芯片集群	以国产 AI 加速芯片为主构建的推理算力集群，

本翻译版说明：本翻译由 Claude 基于港交所 hkexnews 智谱 Z.AI (02513.HK) 2026 中期業績公告英文版 (编号 2026083101539.pdf, 发布日期 2026-08-31, 60 页) 整理。译稿保留所有原始财务数据、专业术语及表述。如英文版本与中文表述之间有任何歧义，应以英文版本为准。

智谱官方未在港交所发布该公告的中文版 PDF。本翻译版不应被视作官方版本。