科智咨询

算力海洋 TOKEN OCEAN

中国词元工厂发展白皮书

产能测算·效能标准·投资决策指引

科智咨询 × 算力海洋 联合发布

2026年

目录

前言 4

1. 中国词元工厂发展现状 5

1.1 产业发展环境 5
1.2 产业发展阶段 7
1.3 产业发展现状 7
1.4 产业发展问题 9

2. 词元工厂投入产出分析 11

2.1 测算体系基础设施环境 11
2.2 全生命周期投入测算 13
2.3 分层收入测算模型 18
2.4 投资回报与盈利分析 22
2.5 核心变量敏感性分析 23

3. 词元工厂产能与服务效能标准 28

3.1 标准体系设计原则 28
3.2 TO-TFS1.0 总体架构 29
3.3 TFS-5 五维核心指标 30
3.4 测试流程与验证机制 33

4. 中国词元工厂发展趋势 35

4.1 规模扩容与业态升级 35
4.2 算电协同与成本优化 35
4.3 交付模式向价值转型 35
4.4 软件优化与调度提效 35
4.5 国产渗透与适配降本 35

5. 中国词元工厂发展风险 36

5.1 流量爬坡利用率不足风险 36
5.2 行业词元定价下行风险 36
5.3 国产硬件软件适配风险 36
5.4 电力供应与成本上涨风险 37
5.5 硬件迭代与估值贬值风险 37
5.6 高负债投入财务杠杆风险 38

6. 核心结论与行业落地建议 39

更多免费行业报告 并购家 www.ipoipo.net

6.1 核心研究结论 39
6.2 分主体落地建议 39
白皮书数据可信度评估体系 41
附录 43

前言

2026 年，国内人工智能产业完成重要范式转换：大模型竞争逐步从“模型参数竞赛、训练算力军备竞赛”，转向推理产能规模化、词元（Token，除专有用语，正文均使用词元）服务标准化、算力运营精细化的产业化深水区。传统智算中心以 GPU 数量、峰值算力作为核心评价指标，而词元工厂（Token 工厂，正文均使用词元工厂）将基础设施的价值锚点转向可计量、可交付、满足服务质量约束的有效 Good Token 产出，标志 AI 算力产业从“堆砌硬件资源”走向“智能精益生产”的新阶段。

随着全国一体化算力网络建设推进，国内智算中心项目快速落地，截至 2026 年 6 月底，全国处于有效动态的智算中心项目超 1300 个，一批面向推理服务的词元工厂试点项目陆续投产，运营商、国资平台、科技企业纷纷入局词元生产业务，词元作为 AI 服务的基础计量单位，正在成为数字经济的新型生产要素。与此同时发展问题同步显现：产能统计口径不一，离线峰值吞吐与实际可交付有效产出混淆；投资测算缺少统一审计口径，不同机构对同一集群的收益测算结果差异巨大；服务 SLA、计量计费规则尚未形成行业共识；国产与海外异构硬件的产能对标、成本测算缺少标准化方法；大量项目沿用传统 IDC 投资逻辑，重硬件采购、轻软件调度优化，出现“算力规模大、有效产出低、投资回报不及预期”的现实困境。

行业迫切需要一套贴合国内产业实践、可复现、可审计的分析框架，厘清词元工厂的产业现状、投资测算逻辑、产能评价标准、发展趋势与风险边界，为此，科智咨询与算力海洋联合编制本白皮书，聚焦产业现实痛点开展系统性研究。白皮书参考国际国内权威机构研究成果与理论、基于头部芯片厂商官方性能测试报告， H200、昇腾 910B 标准硬件集群的量产级实测数据以及国内公开招标数据展开系统测算；研究过程中联合产业链上下游代表性企业开展深度调研，邀请算力架构、运营管理、投资分析等领域资深专家开展多轮研讨与论证校准，重构词元工厂投入产出测算体系，统一硬件选型、财务折旧、收入计量、成本核算、产能统计等核心标准，为项目立项、投资尽调、招标采购、产业政策研究提供依据。

更多免费行业报告 并购家 www.ipoipo.net

1. 中国词元工厂发展现状

词元工厂作为智算中心的高阶形态，已经从传统“售卖 GPU 算力时长”升级为规模化生产、标准化计量、可溯源交付有效词元（Good Token）的新型算力基础设施。当前国内词元工厂整体处于产业发展初期，已经完成了概念普及，正在进入项目建设、能力采购、平台运营的试点验证期，但产业自身也存在发展阶段特征与现实痛点。

1.1 产业发展环境

词元工厂发展与宏观政策导向、产业市场周期、下游需求爆发深度绑定。当前行业正处于政策红利释放、市场供给扩容、需求快速渗透的三重窗口期，环境支撑条件持续完善。

1.1.1 顶层加码强化政策支撑体系

国家层面将算力定位为新型生产要素，多部委协同构建词元产业政策体系，为词元工厂产业发展明确方向、筑牢底座。

顶层设计协同推进：国家网信办、工信部、发改委等多部门密集出台政策，夯实词元产业制度基础。《智能体规范应用与创新发展实施意见》《人工智能拟人化互动服务管理暂行办法》等文件从应用规范、产业创新、安全治理多维度发力，推动 AI 应用从技术验证向规模化落地转型，直接拉动词元需求增长；国家数据局正式将自然语言处理领域的基础概念 Token 规范定名为 “词元”，确立了词元作为新型生产要素的官方定位，为词元计量、交易与产业统计奠定基础。

标准体系提速：工信部、中国信通院牵头，联合相关行业组织，全面加速词元服务标准化体系建设。继《基于 Token 计量的算力能力评价技术要求》等基础标准立项后，2026 年以来发布了《Token 工厂全栈服务能力要求》《高质量 Token 云服务能力评估》《人工智能词元（Token）运营管理能力要求》等多项核心标准，逐步建立词元计量、计费、服务质量的行业共识，为词元工厂商业化交付扫清标准障碍。

地方政策落地：北京出台词元专项支持政策，配套算力、词元消费补贴，落地本地标杆词元工厂项目；上海依托人工智能工程，打造高附加值词元服务节点，赋能长三角产业智能化；广东布局智算基础设施，依托海缆资源推进词元出海试点，建设跨境合规词元服务枢纽。宁夏、内蒙古、甘肃等西部算力枢纽出台专项扶持政策，在绿电指标、土地供应、税收优惠、算力补贴等方面给予支持，直接降低词元工厂的建设与运营成本。

词元出海有序开放：依托全国一体化算力网络布局，工信部、商务部等部门支持探索词元出海服务新模式，广东汕头、海南、重庆等地先后获批相关试点，推行 “来数加工、外数中算” 合规模式，在数据安全合规前提下开放跨境词元服务，拓展词元产业的海外市场空间，为词元工厂打开增量需求空间。

1.1.2 供给扩容催生产业链机会

智算中心建设进入高峰期，市场供给能力快速提升，同时商业模式正在经历从“卖算力”到“卖服务”的深刻变革。

供给规模快速扩容：截至 2026 年 6 月，国内处于有效动态的智算中心项目超过 1300 个，规划算力规模持续增长。其中 100MW 以上大型智算中心达到 21 个，为词元工厂提供了充足的硬件底座。

产业链逐步成熟：上游芯片供给紧张局面缓解，华为昇腾 910B 等量产型号交付稳定；中游推理框架、调度软件、KV 缓存优化技术快速迭代；下游运营商、政企客户开始接受词元计量模式，中国电信等运营商全面转向词元经营，启动词元生成能力服务集采，市场交易机制初步形成。

资本关注度提升：算力基础设施成为科技投资重点领域，国资平台、产业资本、头部科技企业纷纷布局词元工厂项目，资本从关注 GPU 资源稀缺性转向关注运营效率与盈利能力。

1.1.3 应用爆发拉动需求爆发增长

大模型从技术验证走向产业落地，各行业 AI 应用快速渗透，直接拉动词元消耗量指数级增长。

需求规模爆发：国家数据局公开数据显示，我国日均词元消耗量从 2024 年初约 1 千亿增长到 2025 年 6 月底超过 30 万亿，一年半增长 300 多倍。进入 2026 年，随着 Agent、代码生成、行业大模型等应用普及，词元需求仍保持高速增长态势。

需求结构分化：需求呈现“西部大吞吐+东部低时延”的分化特征。西部承接批量推理、离线处理等大吞吐量、非实时需求，对词元成本敏感；东部承载实时交互、企业级应用等高附加值需求，对时延、服务质量要求更高。

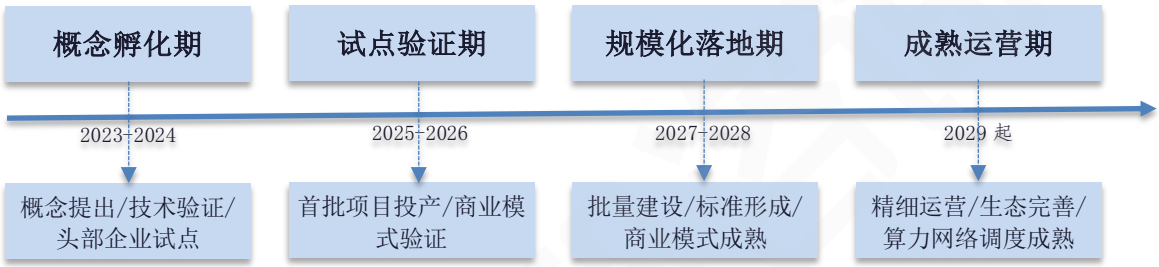
需求主体多元：互联网企业、政企机构、金融机构、制造业等多主体需求逐步释放。其中政企国产化需求增速最快，成为国产算力词元工厂的核心客户群体；互联网企业仍是高端算力词元的最大需求方。

更多免费行业报告 并购家 www.ipoipo.net

1.2 产业发展阶段

中国词元工厂产业整体处于“试点验证期向规模化落地期过渡”的关键节点，尚未进入大规模商业化成熟阶段。结合产业发展规律与项目落地节奏，可将产业发展划分为四个阶段。

图 1 中国词元工厂发展阶段



1.3 产业发展现状

1.3.1 项目建设处于产能扩张前半期

从全行业智算基础设施建设大盘看，截止 2026 年 6 月底，国内处于有效动态的智算中心项目超过 1300 个，其中拟建/筹备期项目占比 49.6%、开工/在建期 39.4%、投产/运营期占 11.0%，行业整体仍处于产能扩张前半程，实际有效产能释放滞后于规划节奏。

词元工厂作为智算中心的高阶形态，进度更为超前：据行业公开披露的 8 个代表性词元工厂项目中，拟建/筹备 4 个、开工/在建 2 个、投产/运营 2 个，多数项目仍处于硬件底座搭建、推理软件适配、商业模式验证的关键阶段。部分落地项目已公开实测数据，验证了词元工厂模式的技术可行性与经济性边界。

当前产业核心特征：一是硬件供给从稀缺逐步缓解，供应链稳定性提升；二是软件优化、调度能力、流量运营成为核心壁垒；三是国产化替代加速，通用型昇腾硬件成为政企项目主流选型；四是行业标准逐步落地，无序产能宣传、虚假收益测算得到遏制。

表 1 智算中心&词元工厂建设情况

建设状态	智算中心分布	词元工厂分布	特征
规划/筹备	49.6%	50%（4 个）	规划论证、筹备建设
开工/在建	39.4%	25%（2 个）	硬件底座搭建、推理软件适配
投产/运	11.0%	25%（2 个）	商业模式验证、产能爬坡

数据来源：科智传媒，截止 2026 年 6 月底数据统计数据

1.3.2 三种发展模式适配不同主体需求

基础设施建设模式

该发展模式核心是打造词元生产的硬件底座，通过建设高密度算力集群、配套绿电与液冷设施，提供稳定的算力生产环境。代表项目包括银川算电一体化词元超级工厂、无锡燧弘华创词元工厂、南宁五象云谷词元工厂等，核心竞争力在于电力成本、土地资源、工程建设能力与政府资源配套。盈利模式以“算力租赁+Token 产出分成”组合为主，成本结构中硬件、电力、机房投入占比超过 90%，为典型的重资产运营模式。核心竞争力在于电力成本、土地资源、工程建设能力，风险特征为现金流稳定、毛利适中、风险极低。

标准化服务集采模式

标准化服务集采模式主要指电信运营商依托其政企客户资源、网络基础设施、合规资质优势，通过长期框架协议，集中采购词元生产能力。其核心是将词元生成能力拆分为可计量、可招标、可批量采购的标准化服务商品，标志着词元从“技术指标”正式成为可交易的标准化生产要素。典型事件为中国电信宁夏公司 2026 年 Token 生成能力服务集采，该模式下运营商与大型政企作为采购方，算力服务商作为供给方，核心履约标准为有效词元产出量、SLA 服务质量、单位词元报价。盈利逻辑薄利多销，依靠稳定大订单覆盖固定成本，现金流稳定性最强。核心竞争力在于 SLA 履约能力、标准化交付、合规资质，风险特征为无拓客压力、现金流最稳定。

精细化平台运营模式

这一模式以头部模型厂商、AI 科技企业为核心主体，依托推理优化技术、模型适配能力、行业客户资源，直接售卖词元服务+行业解决方案，其核心是打通“算力调度-模型适配-API 服务-客户运营”全链条，直接向终端行业客户、中小模型厂商售卖标准化词元服务，从传统“卖 GPU 机时”彻底转向“卖词元产出能力”，代表项目为武汉光谷词元工厂，核心竞争力在于推理优化技术、多模型调度能力、行业客户资源、API 网关运维能力。该模式毛利率上限最高，但客户拓展与流量波动风险也最大，成本结构中软件、人力投入占比显著高于前两类模式。

1.3.3 东西协同形成清晰分工格局

区域布局

当前智算中心的区域分布呈现“能源供给区+需求密集区”双轮驱动的格局，逻辑本质是算电协同、时延适配：

能源供给区：内蒙古、宁夏、青海、甘肃等西部省份，依托低价绿电、充足土地资源、国家算力枢纽政策配套，成为大规模词元工厂的核心承载地，重点布局 GW 级大规模词元生产基地，承载非实时、大吞吐量、长周期的纯推理生产任务。

更多免费行业报告 并购家 www.ipoipo.net

需求密集区：广东、湖北、江苏、北京等东部省市，依托本地成熟的产业需求、高端技术人才、低时延网络适配，布局面向行业实时交互、企业私有化服务的词元算力节点。区域内电力成本较高，但可服务高附加值客户，承载对时延敏感、单价更高的短上下文、实时交互词元任务。

运营主体

科技及算力企业是当前词元工厂最活跃的推动主体，占比超过 50%，主导项目的技术运营、软件适配与商业化落地；运营商凭借网络、机房、政企客户资源与集采渠道，在大规模公共词元服务场景中保持基础盘优势；地方政府及国有控股平台、头部模型厂商、民营企业也通过不同方式参与布局，产业主体多元化特征显著，不同主体的测算假设与盈利目标存在显著差异。

核心载体

词元工厂对算力密度与电力容量要求极高，天然与 100MW 以上大规模智算中心深度绑定：大集群可实现负载跨服务器调度、资源池化共享、成本摊薄效应，中小规模项目无法实现经济规模的利用率提升。国内 21 个 100MW 以上智算中心项目中，19 个已披露项目合计投资额达 2824.5 亿元，中位数 60 亿元，GW 级超大型项目陆续出现。大项目的资本强度、工程复杂度、运营门槛显著提升，只有同时具备硬件建设能力、软件优化能力与产业资源调配能力的主体，才能形成稳定的现金流闭环。

1.4 产业发展问题

尽管产业发展前景明确，但当前词元工厂仍面临投资逻辑不清晰、产能口径混乱、服务标准缺失等核心问题，严重制约投资决策与产业规范化发展，也凸显了建立标准化投入产出模型的紧迫性。

1.4.1 投资普遍重硬件轻运营

行业普遍仍沿用传统智算中心的投资逻辑，以“GPU 数量、峰值算力”为核心决策指标，尚未建立以词元产出为核心的精益生产核算体系。

投资误区普遍：多数项目仍以“有多少卡”为评价标准，忽视利用率、负载结构、服务质量对实际收益的影响，导致“算力规模大但盈利差”的错配现象；

成本认知模糊：对词元生产成本的构成认知不清晰，普遍只关注硬件采购成本，忽视电力、折旧、软件优化、运营等全生命周期成本，实际单位词元成本远超预期；

决策依据缺失：缺乏标准化的投资测算模型，不同项目的测算口径差异巨大，收益预测普遍偏乐观，实际投产后往往不达预期。

1.4.2 产能口径不一 缺乏统一标准

词元产能计量处于无序状态，不同主体宣传的“日产 Token 数”口径差异极大，缺乏统一可比的标准。

指标混用严重：总处理 Token、输入 Token、缓存 Token、输出 Token 概念混用，将输入+输出的“总处理量”宣传为“产出量”，人为放大产能；

质量约束缺失：多数宣传数据采用实验室离线峰值吞吐，不考虑时延、可用性、准确率等 SLA 约束，实际可交付的有效 Good Token 往往仅为峰值的 50%-70%；

场景差异被忽视：办公短问答、长文本、代码 Agent 等不同场景的单位算力产出差异可达数倍，统一用“Token/s”描述产能不具备实际参考价值。

1.4.3 服务规范缺失 制约规模交易

词元服务的标准化体系尚在形成初期，交易双方对交付标准、质量指标、计量方式的认知存在差异，制约规模化交易。

SLA 体系不健全：缺乏统一的 TTFT、TPOT、错误率、可用性等服务质量指标定义与考核方法，服务质量难以量化验证；

计量规则不统一：缓存命中如何计费、失败请求是否扣除、思考词元是否计费等规则不透明，供需双方容易产生结算争议；

验证体系缺失：缺乏第三方权威的词元产能、能效、质量验证机制，厂商宣传数据难以核实，客户采购缺乏客观依据。

当前产业所面临的三大痛点，根源在于缺少一套统一、可审计、可复现的投入产出核算体系。如何科学测算词元工厂的产能、成本、收入与投资回报，是产业各方迫切需要解决的现实课题。

更多免费行业报告 并购家 www.ipoipo.net

2. 词元工厂投入产出分析

本词元工厂投入产出测算模型基于标准词元工厂基础设施环境构建，覆盖 NVIDIA H200 与华为昇腾 910B 两大主流量产硬件路线，以 128 台标准集群为基准部署单元，统一采用 5 年直线折旧的全生命周期财务核算口径。模型对标国际权威测算体系与理论框架：以 NVIDIA 官方 Benchmark 为厂商级性能基准，以 MLPerf 行业通用测试规范为横向比对依据，以 Microsoft Splitwise P/D 分离架构理论为核心模型基础，系统构建覆盖成本核算、产能测算、收入校准、回报评估的全链路测算逻辑，明确测算方法、计算公式、变量定义与校验规则，最终形成标准化、可复现、可审计的量化测算体系，为产业投资决策、项目可研与采购评审提供统一参考基准。

2.1 测算体系基础设施环境

2.1.1 预留冗余设计 保障运行稳定

表 2 词元工厂硬件及基础设施参数

核心指标	海外路线（H200*8 卡）	国产路线（昇腾 910B*8 卡）	可信度
单台服务器单价	350 万元/台(综合考虑海外市场价格及关税影响选取更为贴合市场实际，国内受供需影响的溢价影响分析会在敏感性分析章节呈现)	140 万元/台（政企招标均价）	A 级
集群规模	128 台/1024GPU 生产服务器 120 台/960GPU(参与日常可售词元生产) 预留服务器 8 台(故障冗余、滚动升级、突发弹性、模型切换) 调度方式：集群资源池(统一调度，跨租户批处理，模型路由与 P/D 动态池化)		A 级
单台整机功耗	14.5Kw（8×700W SXM5 H200 GPU+2 路 CPU+内存/网卡/散热系统，为液冷整机量产满载 IT 功耗取值）	11.5kW（逻辑同 H200 集群）	A 级
数据中心 PUE	1.3	1.3	A 级
硬件折旧年限	5 年	5 年	审计标准
基础设施折旧年限	10 年	10 年	审计标准
软件摊销年限	3 年	3 年	审计标准
基准电价（工业电价）	0.7 元/kWh	0.7 元/kWh	A 级
年有效运营时长	8200h	8200h	A 级

数据基准说明：本白皮书所列硬件设施参数基于 2026 年第二季度量产机型官方规格与行业通用实测基准整理，对应 GPU 固件版本、推理框架版本为当前主流稳定版本。硬件参数会随芯片厂商固件迭代、推理引擎优化、集群调度算法升级发生动态变化；不同部署环境、供电条件、散热方案下的实际参数可能存在小幅偏差。项目落地测算建议以现场 POC 实测数据为准，本报告参数作为基准参考使用。

数据时效：至 2026 年 8 月 31 日

2.1.2 P/D 分离架构 提升算力效率

Prefill 消耗与"未缓存输入 Token"相关，Decode 消耗与"输出 Token"相关。P/D 配比必须按工作负载重新计算，固定 4P+2D 或"6 台一个 Cell"无法同时适配代码、短问答和长文。

> **【P/D 平衡配比推导公式】**
>
> 平衡条件：$N_P \times TPS_P/I_{miss} = N_D \times TPS_D/O \Rightarrow N_P/N_D = (TPS_D \times I_{miss})/(TPS_P \times O)$
>
> 实际部署时 D 侧留 5%~10%余量，确保 P 侧为瓶颈、Decode 时延不恶化（符合 SLOGoodput 原则）

表 3 词元工厂生产场景 P/D 平衡设计

场景	生产池分配	计算工作量/请求	SLA 技术请求率	设计含义
代码 Agent	40P+80D	20K 未命中输入+60K 缓存命中+4K 输出	30req/s	P/D 平衡；缓存命中 60K 仅按低价计费
办公短问答	28P+92D	2K 输入+500 输出	420req/s	基准唯一全成本盈利文本负载
办公长文	16P+104D	2K 输入+1K 输出	239.2req/s	Decode 约束增强；现金为正、全成本微利
文生图	不使用 LLMP/D	960 张生产 GPU 并行	64 图/s（15 秒/图）	独立图像模型池

代码 Agent 场景吞吐下降说明：代码 Agent 的 P 吞吐从 30K 降至 15K、D 吞吐从 2.3K 降至 1.5K，原因是 20K 未命中输入+60K 缓存上下文导致 KVCache 体积增大、Attention 计算量增加，有效吞吐自然下降。缓存命中的 60KToken 不消耗 Prefill 计算，但仍按低价计费。

表 4 H200 和 910B 集群 P/D 设计差异

对比维度	英伟达 H200	华为昇腾 910B
单卡显存	141GBHBM3e（可单节点加载 70B）	64GBHBM（需 TP=2/4 多机并行）
显存带宽	4.8TB/s	~1.6TB/s（H200 的 1/3）
瓶颈侧	三场景均为 P 侧瓶颈（Decode 带宽足够强）	三场景均为 D 侧瓶颈（显存带宽受限）
配比策略	D 侧留 5%余量，确保 P 为瓶颈	D 侧需留 10%~15%余量，须启用 autoPD 动态伸缩
吞吐折算	基准值（P=30K/D=2.3K）	整体约为 H200 的 65%~75%（因场景而异）
软件依赖	vLLM 静态配比即可稳定运行	必须依赖华为 autoPD 动态分离+KVCache 内存池化

2.1.3 峰谷模型锚定商业利用率基准

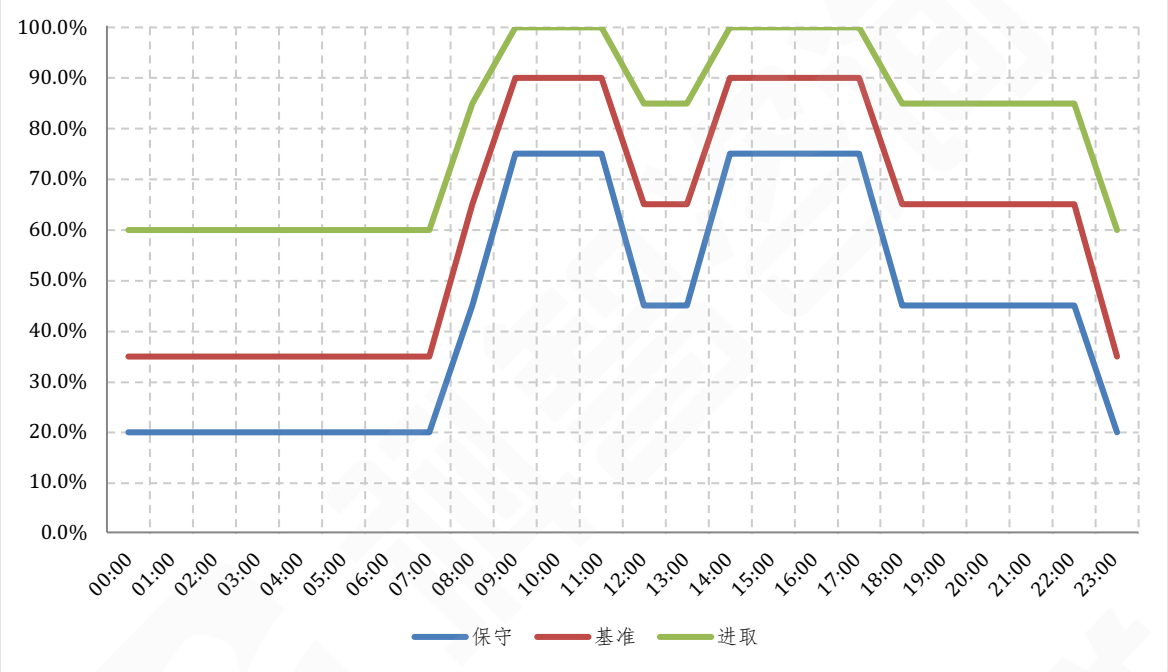
本白皮书商业利用率取值均以商业峰谷利用率模型为核心测算依据。模型依托国内主流词元工厂 GLM-5.2 分时段运营监控数据构建，按全年 24 小时加权平均核算，保守情境下基准商业利用率为 44.4%，基准情境为 61%，进取情境为 80%。针对短文本问答、长文本生成、Agent 代码执行、文生图等差异化业务场景，将结合各场景算力消耗特征、并发波动规律及服务时延约束，在基准模型基础上对利用率取值进行场景化校准。

更多免费行业报告 并购家 www.ipoipo.net

表 5 GLM-5.2 保守/基准/进取情景下分时段商业利用率列表

时段	保守	基准	进取
谷段	20.0%	35.0%	60.0%
肩段	45.0%	65.0%	85.0%
峰段	75.0%	90.0%	100.0%
24 小时加权平均	44.4%	61.0%	80.0%

图 2 24 小时商业利用率曲线



说明：GLM-5.2 的 API 目录价不分峰谷，本商业峰谷利用率模型中的峰谷只表示客户付费需求占服务级容量的比例。

2.2 全生命周期投入测算

2.2.1 两类成本构成核算框架

CAPEX 投入测算逻辑

CAPEX 为项目一次性建设投入，分为硬件主体、基础设施、软件平台三大模块，按行业固定占比计提：

表 6 词元工厂 CAPEX 拆分

CAPEX 项目	占比区间	基准取值依据	可信度
AI 服务器集群	70%-82%	H2008 卡 350 万元/台；昇腾 910B8 卡 140 万元/台 含 NVL 互联、高速存储、液冷配件等	A 级
基础设施配套	12%-20%	按服务器 CAPEX 的 15%计提，为智算中心配套设施成本中位数水平，含配电、液冷、机柜、网络	A 级
软件平台	5%-10%	海外 8%、国产 10%，含推理框架、调度系统、计费平台、 第三方优化授权等，国产系统还需做算子适配	A 级

【全周期折旧测算公式】

AI 硬件年化折旧（5 年）=硬件总 CAPEX÷5

基础设施年化折旧（10 年）=基础设施总 CAPEX÷10

软件年化摊销（3 年）=软件总投入÷3

年化全成本=硬件年折旧+基础设施年折旧+软件年摊销+年化现金 OPEX

合规性说明：5 年硬件折旧完全匹配国资、上市公司财务审计要求，规避旧版 3 年折旧过于激进、审计不通过的问题，测算收益更保守、更真实。

OPEX 投入测算逻辑

OPEX 为项目运营期的年化持续成本，电力为第一大可变成本，各项成本取值基于行业实测数据校准：

表 7 词元工厂 OPEX 拆分

OPEX 项目	基准取值依据	可信度
电力成本	工业电价 0.7 元/kWh；集群功耗×PUE×运营时长	A 级
IDC 租金、机柜、带宽	机柜租金 0.4 万元/KW/机柜/年	B 级
硬件维保、备件、保险	服务器 CAPEX×3%/年	A 级
软件订阅与运维 SRE(+算子适配)	服务器 CAPEX×2%-3%(+1%)/年	A 级
模型优化、调度运维	15-20 人技术团队	B 级
商务及客户成功等	上述合计×8%	B 级
行政及合规等	上述合计×4%	B 级

2.2.2 H200 集群 资本投入更高

英伟达 H200 集群 CAPEX 测算

表 8 英伟达 H200 集群词元工厂 CAPEX 测算

CAPEX 项目	计算过程	金额（万元）	占比
AI 服务器硬件主体	128 台×350 万元/台	44,800.00	81.30%
基础设施配套	44,800×15%	6,720.00	12.20%
软件平台投入	44,800×8%	3,584.00	6.50%
CAPEX 合计	—	55,104.00	100.00%

英伟达 H200 集群年化折旧

硬件年折旧（5 年）=44,800÷5=8,960 万元/年

基础设施年折旧（10 年）=6,720÷10=672 万元/年

软件年摊销（3 年）=3,584÷3≈1,194.67 万元/年

年化总折旧摊销=8,960+672+1,194.67=10,826.67 万元/年

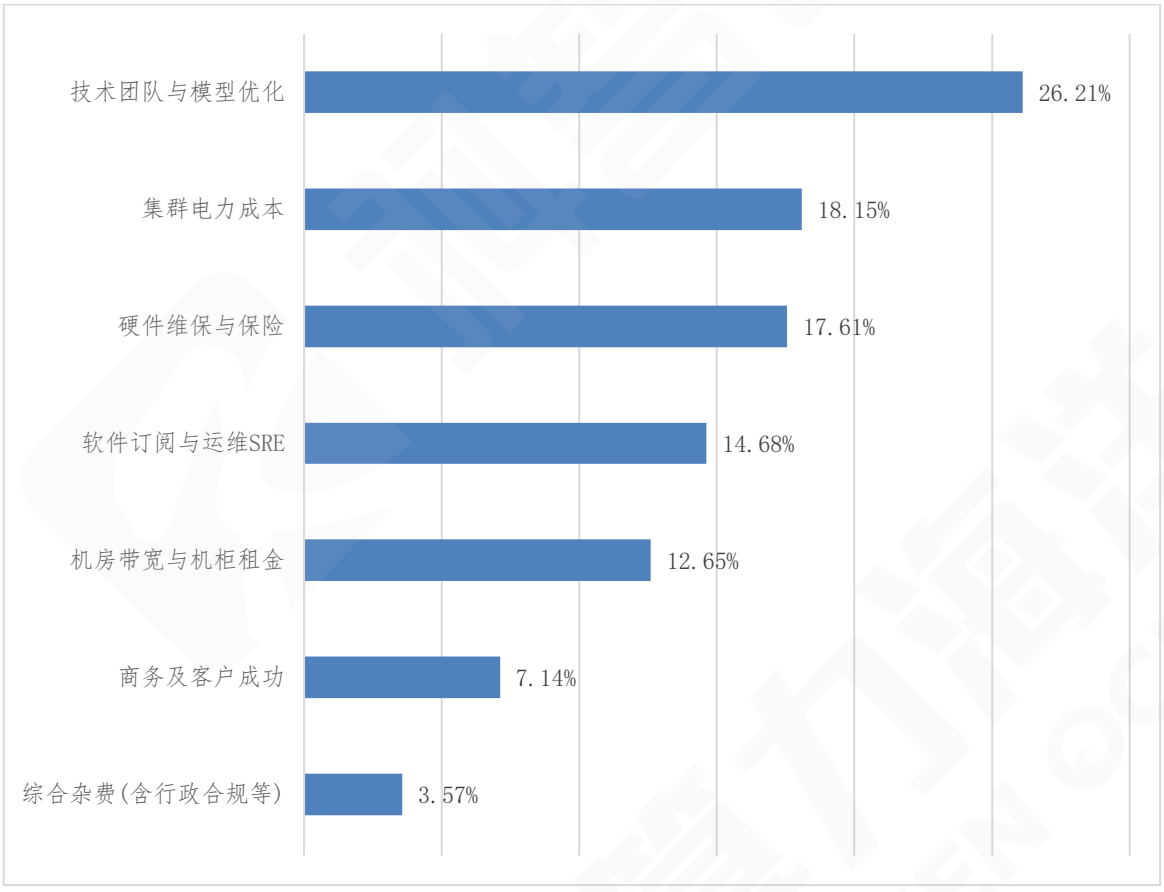
更多免费行业报告 并购家 www.ipoipo.net

英伟达 H200 集群 OPEX 测算

表 9 英伟达 H200 集群词元工厂 OPEX 测算

OPEX 项目	测算过程	年化金额（万元）	占比
集群电力成本	128 台×14.5KW/台×1.3PUE=2412.8KW 2412.8kW×8200h×0.7 元/度	1384.95	18.15%
机房带宽与机柜租金	行业集群市场价	965.12	12.65%
硬件维保与保险	硬件 CAPEX×3%	1344.00	17.61%
软件订阅与 SRE 运维	硬件 CAPEX×2.5%	1120.00	14.68%
技术团队与模型优化	15 人技术团队+第三方服务费	2000.00	26.21%
商务及客户成功等	上述合计×8%	545.13	7.14%
行政及合规等	上述合计×4%	272.56	3.57%
年化现金 OPEX 合计	—	7631.76	100.00%

图 3 英伟达 H200 集群年化 OPEX 分布

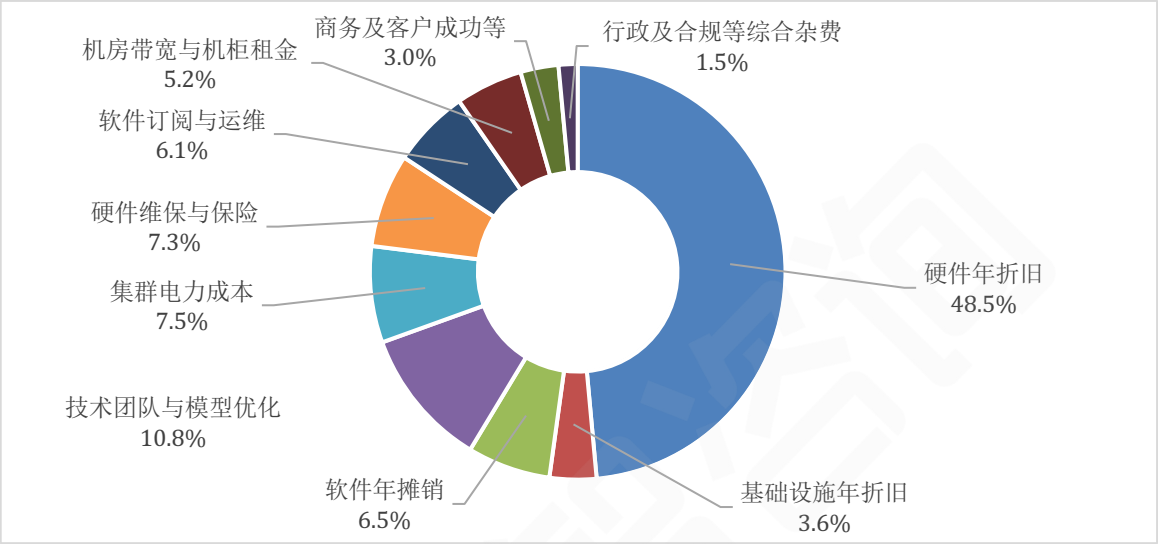


英伟达 H200 集群年化全成本分析

词元工厂应用海外 NVH200 集群年化全成本共计 18458.42 万元，其中年化折旧摊销 10826.67 万元，年化 OPEX 支出 7631.76 万元。

H200 集群词元工厂投入中硬件年化折旧在年化全成本中占比接近 50%，年化 OPEX 支出中技术团队与模型优化投入占比 10.8%，电力成本紧随其后，占比 7.5%。

图 4 英伟达 H200 集群年化全成本分布



2.2.3 910B 集群 成本优势显著

华为昇腾 910B 集群 CAPEX 投入

表 10 华为昇腾 910B 集群词元工厂 CAPEX 测算

CAPEX 项目	计算过程	金额（万元）	占比
AI 服务器硬件主体	128 台×140 万元/台	17,920.00	80.00%
基础设施配套	17,920×15%	2,688.00	12.00%
软件平台+算子适配	17,920×10%	1,792.00	8.00%
CAPEX 合计	—	22,400.00	100.00%

华为昇腾 910B 集群年化折旧

硬件年折旧（5 年）=17,920÷5=3,584 万元/年

基础设施年折旧（10 年）=2,688÷10=268.8 万元/年

软件年摊销（3 年）=1,792÷3≈597.33 万元/年

年化总折旧摊销=3,584+268.8+597.33=4,450.13 万元/年

华为昇腾 910B 集群 OPEX 测算

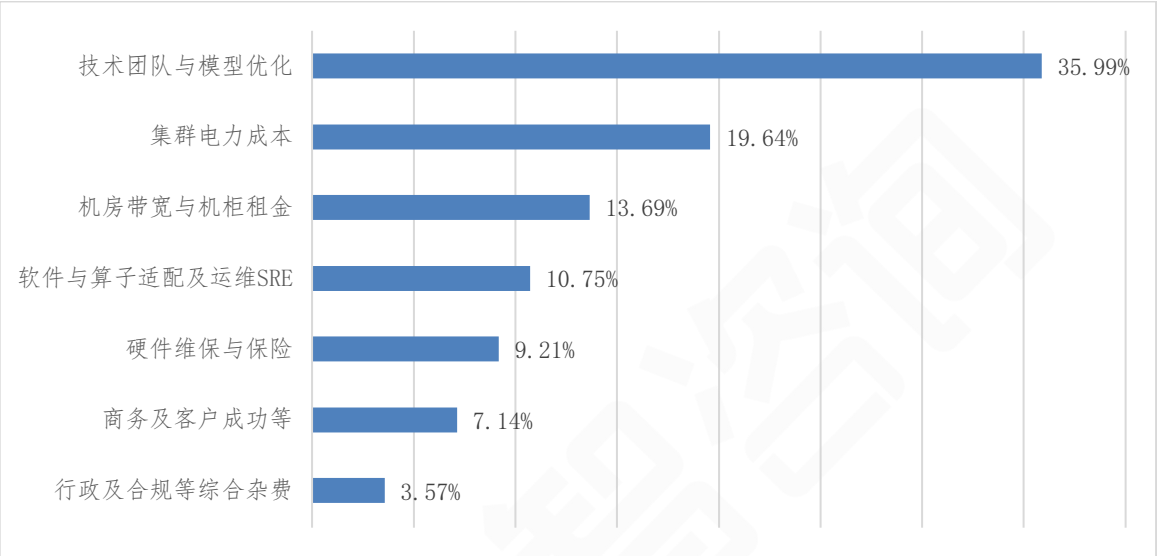
表 11 华为昇腾 910B 集群词元工厂 OPEX 测算

OPEX 项目	测算过程	年化金额（万元）	占比
集群电力成本	1996.8kW×8200h×0.7 元/度	1146.16	19.64%
机房带宽与机柜租金	行业集群市场价	798.72	13.69%
硬件维保与保险	硬件 CAPEX×3%	537.6	9.21%
软件与算子适配及运维 SRE	硬件 CAPEX×3.5%	627.2	10.75%
技术团队与模型优化	20 人适配团队+技术服务费	2,100.00	35.99%
商务及客户成功等	上述合计×10%	416.77	7.14%
行政及合规等综合杂费	上述合计×5%	208.39	3.57%

更多免费行业报告 并购家 www.ipoipo.net

年化现金 OPEX 合计	—	5,834.84	100.00%

图 5 华为昇腾 910B 集群年化 OPEX 分布

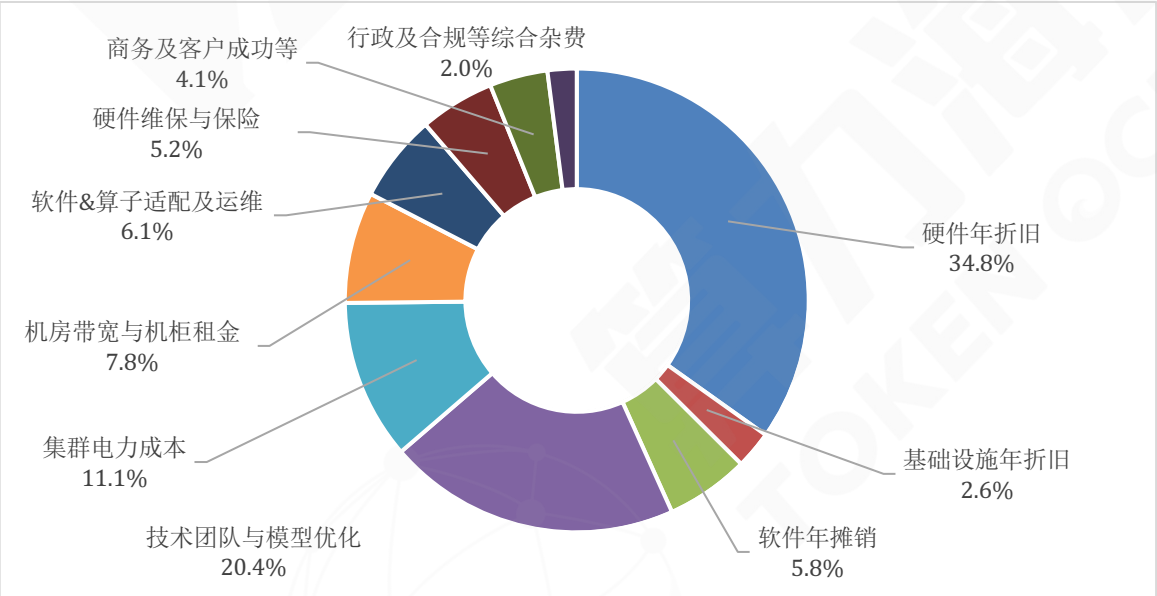


华为昇腾 910B 集群年化全成本分析

词元工厂应用国产华为昇腾 910B 集群年化全成本共计 10284.97 万元，其中年化折旧摊销 4450.13 万元，年化 OPEX 支出 5834.84 万元。

910B 集群词元工厂投入中硬件年折旧在年化全成本中占比 34。8%，年化 OPEX 支出中技术团队与模型优化投入占比 20.4%，电力成本占比 11.1%。

图 6 华为昇腾 910B 集群年化全成本分布



2.3 分层收入测算模型

词元工厂的核心产能计量单位为 Good Token，即符合时延、可用性、准确率标准的用户侧有效输出 Token，区别于行业宣传的总 Token、缓存 Token、输入 Token，产能测算严格遵循“先校准质量、再计量产出”的原则。

2.3.1 六步校准 精准核算营收

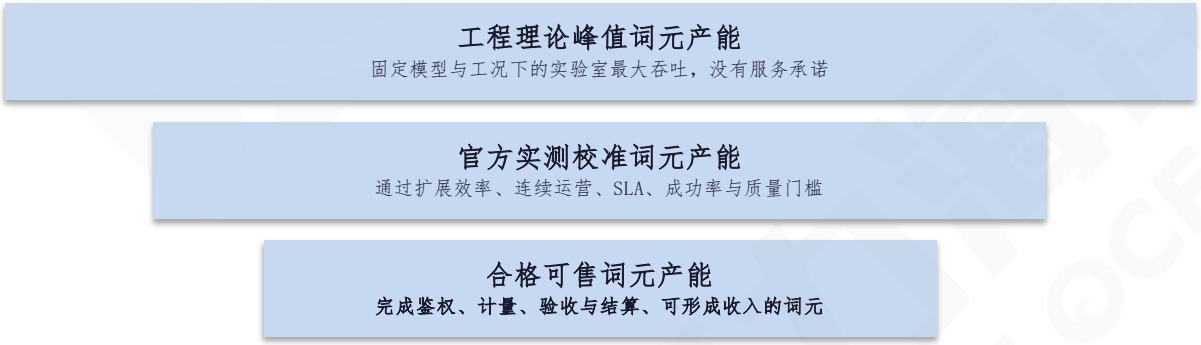
。词元产能模型

参考 Phase-aware Roofline 理论与 MLPerf 的标准化测试方法，词元工厂产能测算可按如下三层级逐步确认：

表 12 词元工厂产能分层

层级	名称	核心问题	计算方法	本报告用途
第一层	Engineering Roofline 工程理论上限	物理上最多能做到多少？	t_step≥max(t_compute,t_memory,t_communication)	上限参照
第二层	Measured Throughput 官方实测校准	现实软件栈能跑到多少？	NVIDIA GenAI-Perf/Tensor RT-LLM/MLPerf 实测	单节点 TPS_P/TPS_D 基准值来源
第三层	SLO Goodput 合格可售产能	满足质量约束能卖多少？	P/D 分离配比+TTFT/TPOT 约束+通过率	收入测算的唯一产能基准

图 7 从理论产能到合格可售词元产能



【年产词元公式】

Annual Delivered Tokens=Goodput_SLO×31,536,000×Availability×Load Factor

其中：Availability 对应"120 生产+8 预留"的 SLA 规划；Load Factor 对应商业利用率 CUR

【词元产能相关参数】

①Model（模型及版本）②Architecture（Dense/MoE）③Precision（BF16/FP8）

④Hardwaretopology（GPU 数量+并行方式）⑤ISL/OSL（输入输出长度分布）

更多免费行业报告 并购家 www.ipoipo.net

⑥Concurrency（并发度）⑦TTFT/TPOTSLO（交互质量约束）⑧Servingstack（推理框架版本）——任何 TPS 数字必须同时标注这八项参数。

词元收入口径

本报告词元工厂产能在理论峰值基础上通过物理层、业务层和价格层的商业化修正完成最终收入测算，报告中商业利用率不是 GPU 利用率，也不是框架吞吐折扣。120 台生产节点的吞吐已按 SLA 规划值给出；商业利用率只表示市场需求是否把该产能卖满，从而避免重复测算。

表 13 词元工厂收入口径

层级	定义	处理方式	扣减次数
理论峰值	实验室或框架基准最大吞吐	只作上限参照，不直接计收入	—
SLA 合格技术产能	预留故障、发布、峰值和延迟余量后的生产能力	120 台生产、8 台预留；P/D 吞吐已是 SLA 规划值	第 1 次（物理层）
商业利用率 CUR	客户实际购买并消耗的产能占 SLA 产能的比例	代码 55%、对话 60%、图片 50%；只扣一次	第 2 次（业务层）
价格实现率 PRR	合同净价相对公开 List Price 的比例	基准 65%	第 3 次（价格层）

关键甄别结论：120 生产+8 预留与 CUR 不存在重复计算

8 台预留属于物理资源可用性层面，直接从 128 台中扣除，不进入产能公式分子；CUR 属于业务负载饱和度层面，只作用于 120 台生产节点的营收折算。公式表达：实际营收=f(120 台的技术产能)×CUR×PRR，8 台从未出现在 f()的输入中，因此与 CUR 不存在任何重复扣减关系。

词元收入测算模型

本收入模型对词元工厂营收进行全链路细化测算，从集群吞吐、请求约束、缓存计费、产能折算、价格校准到最终营收，实现六层级分步测算，解决行业收入测算模糊、虚高、不可复现的问题。

【第 1 层：P/D 双向吞吐约束】

词元推理吞吐受 Prefill 输入、Decode 输出双向约束，真实产能取最小值，公式如下：

P 阶段请求率：R_P=N_P×TPS_P/Input_miss

D 阶段请求率：R_D=N_D×TPS_D/Output

技术请求率：R_base=min(R_P,R_D)；P/D 必须按工作负载重新配比

变量说明：R_P 为 Prefill 阶段请求率、R_D 为 Decode 阶段请求率、R_base 为集群基准请求率；N_P/N_D 为集群 P/D 节点配比、TPS_P/TPS_D 为单节点吞吐、Input/Output 为单请求输入输出词元长度。

【第 2 层：SLA 通过率校准】

基准请求率需叠加服务质量、故障损耗、排队延迟修正，得到 SLA 合格请求率：

R_sla=R_base×Q_pass

变量说明：Q_pass=95%，SLA 合规场景请求通过率。

【第 3 层：缓存分层计费收入】

区分三类计费词元，贴合大厂真实计费规则，杜绝总收入虚高：

Rev_req=I_miss×P_miss+I_hit×P_hit+O×P_out（Token 均除以 1,000,000，单位：美元/百万 Token）

变量说明：Input_miss 为未缓存输入 Token（高价计费）、Input_hit 为缓存命中输入 Token（低价计费）、Output_Token 为输出 Token（高价计费）；P_miss/P_hit/P_out 为对应档位单价（美元/百万 Token）。

【第 4 层：年化与汇率折算】

Rev_year0=R_sla×31,536,000×Rev_req×Rate（Rate=7.2，美元兑人民币基准汇率）

变量说明：31536000 为全年秒数、Rate 为美元兑人民币汇率（基准 7.2）、Rev_year0 为理论年化营收。

【第 5 层：商业利用率校准】

理论营收需叠加市场真实消耗比例，匹配实际流量情况：Rev_year1=Rev_year0×CUR

变量说明：CUR 为商业利用率（短问答/长文 60%，代码 55%，文生图 50%）

【第 6 层：价格实现率最终校准】

公开挂牌价需扣除集采、长期合作折扣，得到最终可确认营收：

Rev_final=Rev_year1×PRR（PRR=65%，政企集采通用折扣）

最终年化真实营收=Rev_final（本报告所有收入数据均为此公式输出，100%可复现）

2.3.2 H200 集群营收规模更高

【办公短问答场景收入测算】

P/D=28P+92D|TPS_P=30,000|TPS_D=2,300|输入=2,000|输出=500

R_P=28×30,000/2,000=420req/s|R_D=92×2,300/500=423.2req/s|R=min=420req/s

单请求收入=2K×$1.40/M+0.5K×$4.40/M=$0.005/轮

年化营收=420×31,536,000×$0.005×7.2×60%×65%=18,596.15 万元

更多免费行业报告 并购家 www.ipoipo.net

【办公长文场景收入测算】

P/D=16P+104D|输入=2,000|输出=1,000

R_P=16×30,000/2,000=240req/s|R_D=104×2,300/1,000=239.2req/s|R=239.2req/s

单请求收入=2K×$1.40/M+1K×$4.40/M=$0.0072/轮

年化营收=239.2×31,536,000×$0.0072×7.2×60%×65%=15,250.97 万元

【代码 Agent 场景收入测算】

P/D=40P+80D|TPS_P=15,000|TPS_D=1,500|未命中输入=20,000|缓存命中=60,000|输出=4,000

R_P=40×15,000/20,000=30req/s|R_D=80×1,500/4,000=30req/s|R=30req/s

单请求收入=20K×$1.40/M+60K×$0.26/M+4K×$4.40/M=$0.0612/轮

年化营收=30×31,536,000×$0.0612×7.2×55%×65%=14,903.48 万元

2.3.3 910B 集群 投入产出更优

910B 单节点有效吞吐约为 H200 的 65%，代码 Agent 因缓存命中高、相对表现提升至 75%。P/D 配比与 H200 保持一致，但因 D 侧瓶颈更突出，实际部署需启用 autoPD 动态分离。

表 14 华为昇腾 910B 集群词元工厂营收测算

项目	办公短问答	办公长文	代码 Agent
P/D 配比	28P+92D	16P+104D	40P+80D
单节点 P 吞吐(t/s)	19,500（30K×65%）	19,500	11,250（15K×75%）
单节点 D 吞吐(t/s)	1,495（2.3K×65%）	1,495	1,125（1.5K×75%）
R_P	273(28×19,500/2,000)	156(16×19,500/2,000)	22.5(40×11,250/20,000)
R_D	275.1(92×1,495/500)	155.5(104×1,495/1,000)	22.5(80×1,125/4,000)
R=min(req/s)	273	155.5	22.5
单请求收入($/req/s)	0.0050	0.0072	0.0612
CUR	60%	60%	55%
PRR	65%	65%	65%
年化营收（万元）	12087.50	9914.40	11177.61

DeepSeek-V4-Pro 场景测算结论

受限于模型公开定价过低（输入$0.435/M、输出$0.87/M），所有文本场景年化营收均无法覆盖现金 OPEX，全成本口径大幅亏损，仅可作为流量填谷、低价获客补充业务，无法作为 H200 集群核心盈利场景。

2.4 投资回报与盈利分析

2.4.1 全周期指标衡量投资回报

【ROI 核心公式体系】

EBITDA=最终年化营收-年化现金 OPEX

年度净利润=(EBITDA-年化总折旧摊销)×(1-25%所得税率)

年度净现金流=年度净利润+年化总折旧摊销

静态投资回收期=初始总 CAPEX÷年均净现金流

五年现金 ROI=(5×年均净现金流-初始 CAPEX)÷初始 CAPEX

NPV/IRR 测算：基准折现率 8%（行业通用资金成本）

EBITDA 口径统一说明：

EBITDA 统一按"年化营收-年化现金 OPEX"计算，现金 OPEX 含电力、维保、软件、技术团队、杂费，不含非现金折旧摊销。此口径与国际财务准则一致，确保跨项目可比性。

2.4.2 H200 集群 短问答收益最优

表 15 英伟达 H200 集群词元工厂分场景 ROI 测算

指标	办公短问答	办公长文	代码 Agent
年化营收(万元)	18596.15	15250.97	14903.48
年化现金 OPEX(万元)	7631.76	7631.76	7631.76
年化 EBITDA(万元)	10964.39	7619.21	7271.73
年化折旧摊销	10826.67	10826.67	10826.67
全成本利润	137.72	-3207.46	-3554.94
静态投资回收期(年)	5.0	6.5	6.8
五年现金 ROI	-0.82%	-23.59%	-25.95%
投资判定	核心盈利标杆	近盈亏平衡	近盈亏平衡

EBITDA 口径修正说明：本表 EBITDA 统一按公式 EBITDA=年化营收-年化现金 OPEX 计算。文生图场景因独立算力池，OPEX 分摊口径需单独核算。

2.4.3 910B 集群 回报率显著领先

表 16 华为昇腾 910B 集群词元工厂分场景 ROI 测算

指标	办公短问答	办公长文	代码 Agent
年化营收(万元)	12087.50	9914.40	11177.61
年化 OPEX(万元)	5834.84	5834.84	5834.84
年化 EBITDA(万元)	6252.65	4079.56	5342.77
年化折旧摊销(万元)	4450.13	4450.13	4450.13
全成本利润(万元)	1802.52	(370.57)	892.64
静态投资回收期(年)	3.8	4.6	4.4
五年现金 ROI	30.72%	7.96%	14.14%
投资判定	优质盈利，国产化首选	近盈亏平衡	良好盈利

更多免费行业报告 并购家 www.ipoipo.net

华为昇腾 910B 路线下各场景 ROI 显著高于 H200 的核心原因：910B 的 CAPEX 仅为 H200 的 40.6%（22,400 万 vs55,104 万），而营收达到 H200 的 70%~75%，单位投资产能效率更高。但前提是必须完成全链路算子优化，若适配率低于 80%，有效吞吐降至 H200 的 50%以下，回收期将拉长至 6 年以上。

英伟达 H200 与华为昇腾 910B 集群核心投入产出指标对比

表 17 办公短问答场景核心指标对比

对比指标	H200	910B	910B/H200
总 CAPEX(万元)	55,104.00	22,400.00	40.60%
年化现金 OPEX(万元)	7,631.76	5,834.84	76.45%
系统请求率(req/s)	420	273	65.00%
年化营收(万元)	18596.15	12087.50	65.00%
五年现金 ROI	-0.82%	30.72%	--
静态回收期(年)	5.0	3.8	75.87%

2.5 核心变量敏感性分析

2.5.1 商业利用率对收益影响最显著

商业利用率＞词元成交单价＞软件优化吞吐＞电价＞硬件采购成本

- 商业利用率每波动±10%：海外路线 IRR 波动±17%、国产路线 IRR 波动±13%
- 词元单价每波动±10%：海外路线 IRR 波动±15%、国产路线 IRR 波动±11%
- 电价每波动±0.1 元/kWh：双路线 IRR 波动±7%-9%
- 软件适配不足导致吞吐下降 10%：国产路线 IRR 直接下滑 8-10 个百分点

表 18 词元工厂 ROI 测算核心变量敏感度分析

变量名称	变动幅度	H200 集群 IRR 变动	910B 集群 IRR 变动	敏感度排序
商业利用率 CUR	±10%	±18.2%	±12.5%	1
词元成交单价 PRR	±10%	±15.6%	±10.3%	2
有效吞吐系数 GUF	±10%	±12.8%	±14.6%	3
电力成本	±10%	±8.7%	±7.2%	4
硬件采购成本	±10%	±6.4%	±5.1%	5

2.5.2 双维矩阵定位盈亏边界

以下分析基于办公短问答基准场景（行业最主流盈利场景），验证商业利用率（CUR）与价格实现率（PRR）双维度波动对项目净利润的影响，明确盈亏平衡边界与安全垫水平。测算遵循以下统一口径：

- 集群规模固定为 128 台标准集群（120 台生产+8 台预留）

- 年化全成本为固定值：H200 集群 18,458.42 万元/年，昇腾 910B 集群 10,284.97 万元 / 年（含折旧摊销与现金运营成本）

- 商业利用率（CUR）取值：40%、50%、60%、70%、80%、90%

- 价格实现率（PRR）取值：50%、60%、70%、80%、90%

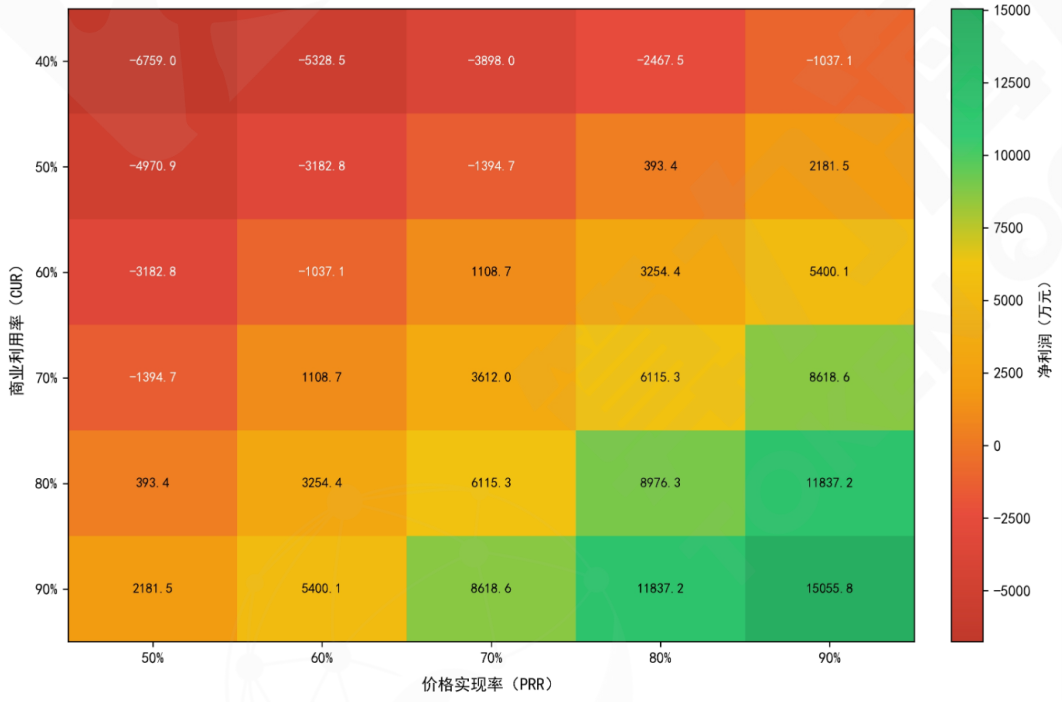
- 净利润=（年化营收-年化全成本）*（1-25%所得所率）；年化营收与 CUR、PRR 呈线性正相关

英伟达 NVH200 集群盈亏平衡分析

亏损区域覆盖 CUR40%~70%、PRR50%~60%区间，最大亏损 6759 万元，出现在 CUR=40%、PRR=50% 的极端场景，亏损深度随 CUR 与 PRR 同步下滑快速扩大，呈现双因子叠加放大效应。盈亏临界点（净利润=0）大致位于 CUR×PRR≈37.2%的对角区间，对应典型场景如 CUR=60%、PRR≈62%，或 CUR=70%、PRR≈53%，临界点两侧盈利弹性对称。

行业通用基准场景（CUR=60%、PRR=65%）对应净利润约 36 万元，恰好紧贴盈亏平衡线，安全边际极薄——商业利用率或价格实现率仅需出现 2~3 个百分点的下行波动，项目就会转入亏损。该基准对应当前国内词元工厂的平均运营水平，反映出行业整体处于盈亏平衡线附近、盈利承压的发展阶段，规模化盈利仍需依赖利用率爬坡与定价体系优化的双重支撑。

图 8 英伟达 H200 集群办公短问答场景净利润热力图(年化全成本 vs 双维度营收)



热力图单位为万元/年；固定模型、工况、服务级系数与资源份额。

更多免费行业报告 并购家 www.ipoipo.net

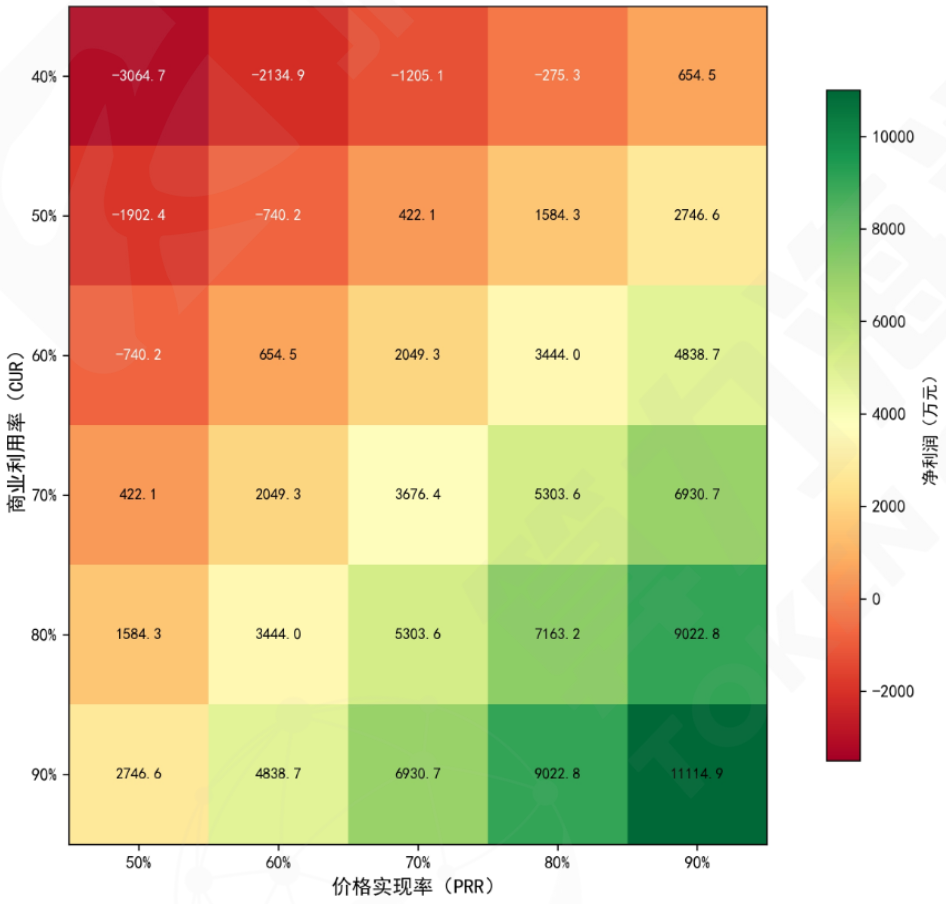
华为昇腾 910B 集群盈亏平衡分析

亏损区域集中分布在矩阵左下角的“低商业利用率 + 低价格实现率”区间，亏损面与亏损深度均显著小于 H200 集群。最大亏损值 3064.7 万元，出现在 CUR=40%、PRR=50%的极端场景，仅为 H200 同场景亏损深度的 45%。亏损区域呈现明显的收缩特征：当 CUR≥70% 时，全价格区间均实现盈利；当 PRR≥60% 时，仅 CUR=40% 的极端低利用率场景仍处亏损，体现出低成本硬件路线下更强的盈利韧性。

盈亏平衡线（净利润=0）大致位于 CUR×PRR≈33.2%的对角区间，较 H200 集群低约 5.8 个百分点，盈利门槛显著下降。对应典型场景：CUR=60% 时临界点约为 PRR≈55.3%，CUR=50% 时临界点约为 PRR≈66.4%。

行业通用基准场景（CUR=60%、PRR=65%）对应净利润约 1351.9 万元，处于稳定盈利区间，安全边际充足——商业利用率或价格实现率可承受约 10 个百分点的下行波动才会触及盈亏平衡线。该结果显著优于同基准下 H200 集群的微利状态，印证了昇腾路线在行业平均运营水平下具备更扎实的盈利基础，更适配当前词元工厂普遍利用率不足的行业现状。

图 9 华为昇腾 910B 集群办公短问答场景净利润热力图(年化全成本 vs 双维度营收)



热力图单位为万元/年；固定模型、工况组合、服务级系数与资源份额。

2.5.3 硬件价格影响收益水平

英伟达 NVH200 集群硬件价格敏感性分析

基于 128 台标准 H200 集群、办公短问答基准场景，固定年化营收 18596 万元，基础设施配套按硬件额 15% 计提、软件授权按 8% 计提，维保与订阅费用随硬件价格联动，覆盖全生命周期成本与收益维度。

测算显示，H200 服务器采购价格对项目盈利与投资回报表现极强负相关：单台价格每上涨 50 万元，年化全成本提升约 1941 万元，年净利润收窄约 1941 万元，五年现金 ROI 下降约 11~13 个百分点。项目全成本盈亏平衡点约为 354 万元/台（全成本盈亏平衡定义为年化净利润=0），具备商业投资价值的价格红线约为 347 万元/台，价格低于该阈值时，五年周期可实现正现金投资回报；价格高于该阈值时，纯商业化运营无法在五年内收回初始投资，不具备独立商业投资价值。

表 19 英伟达 H200 服务器价格敏感性测算数据表

指标项	300 万元/台	350 万元/台	400 万元/台	450 万元/台	500 万元/台
总 CAPEX（万元）	47,232.00	55,104.00	62,976.00	70,848.00	78,720.00
年化折旧摊销（万元）	9,280.00	10,826.67	12,373.33	13,920.00	15,466.67
年化现金 OPEX（万元）	7,237.52	7,631.76	8,026.00	8,420.24	8,814.48
年化全成本（万元）	16,517.52	18,458.42	20,399.33	22,340.24	24,281.15
年化 EBITDA（万元）	11,358.63	10,964.39	10,570.15	10,175.91	9,781.67
全成本净利润（万元/年）	2,078.63	137.73	-1,803.18	-3,744.09	-5,685.00
五年现金 ROI	14.74%	-0.82%	-12.50%	-21.58%	-28.84%

受全球 GPU 供应链紧张、出口管制及存储缺货等因素叠加，当前国内 8 卡 H200 服务器主流采购价处于高溢价区间，远超 347 万元的五年投资价值红线。多数商业化词元工厂仅能实现微利甚至亏损，纯商业投资价值大幅收缩，项目普遍转向成本回收导向。行业分层加剧，头部厂商凭借规模采购锁价仍可维持正回报，中小服务商成本承压明显，部分项目暂缓建设或转向国产路线，昇腾路线凭借更厚的盈利安全垫，性价比优势持续凸显。

华为昇腾 910B 集群硬件价格敏感性分析

基于 128 台标准昇腾 910B 集群、办公短问答基准场景测算，固定年化营收 12087.5 万元；基础设施配套按硬件投资额 15%计提，软件授权按 10%计提，维保、订阅等运营费用随硬件价格联动，覆盖年度全成本净利润、五年现金投资回报两大核心维度。

测算显示，910B 服务器全成本盈亏平衡价格约 190.5 万元/台，价格低于该阈值时项目年度盈利，高于则年度持续亏损。投资价值红线（五年现金 ROI=0）约 239.4 万元/台，价格高于该阈值时，纯商业化运营无法在五年内收回初始投资，不具备独立商业投资价值。

更多免费行业报告 并购家 www.ipoipo.net

表 20 华为昇腾 910B 服务器价格敏感性测算数据表

指标项	140 万元/台	160 万元/台	180 万元/台	200 万元/台	250 万元/台	300 万元/台
总 CAPEX（万元）	22,400	22,042	25,190	28,339	31,488	39,360
年化折旧摊销（万元）	4,450.13	5,085.87	5,721.60	6,357.33	7,946.67	9,536.00
年化现金 OPEX（万元）	5,634.14	5,791.83	5,949.53	6,107.23	6,501.47	6,895.71
年化全成本（万元）	5,834.84	6,021.21	6,207.58	6,393.95	6,859.87	7,325.79
年化 EBITDA（万元）	6,453.36	6,295.66	6,137.97	5,980.27	5,586.03	5,191.79
全成本利润（万元/年）	2,003.23	1,209.80	416.37	-377.06	-2,360.64	-4,344.21
五年现金 ROI	32.87%	35.95%	19.77%	7.18%	-1.93%	-20.25%

相比于英伟达 H200 八卡服务器，昇腾 910B 八卡服务器的价格优势更加明显，其盈亏平衡价格约为当前主流采购价（140~160 万元 / 台）的 1.2~1.4 倍，投资价值红线约为当前采购价的 1.5~1.7 倍。即使硬件价格上涨 30%，项目仍可保持年度盈利；价格上涨超 70% 才会触及投资红线，抗供应链波动和溢价风险的能力远强于 H200，更适配国内词元工厂的商业化建设需求。

3. 词元工厂产能与服务效能标准

当前公开市场词元度量指标体系多元并存，TPS、TPM、Token / 日、Tokens/W、美元 / 百万 Token 等指标广泛使用，但各类指标普遍未同步披露对应模型规格、上下文输入输出长度、缓存命中规则、生成准确率、并发负载边界及功耗约束等核心前置条件，指标横向可比性不足，难以支撑标准化评估与商业化交易。

针对行业共性痛点，**算力海洋（Token Ocean）**依托全链路产业运营实践积累，联合科智咨询共同研究编制《词元工厂产能与服务效能评价标准（TO-TFS 1.0）》，简称 TFS-5 五维标准，系面向词元产业的行业倡议标准。

本报告所载 TO-TFS 1.0 标准框架及 TFS-5 五维指标，为双方联合研究成果与行业倡议，不属于国家强制标准范畴。任何基于本标准的产能外推、收益测算或规划假设，均需在指定模型、软件栈、业务负载及 SLA 约束条件下，通过现场 POC 实测或连续运营数据验证后，方可转化为采购承诺、服务 SLA 约定或投资决策依据。

该标准适用主体涵盖智算中心运营方、词元工厂建设与运营主体、MaaS 服务平台、模型服务商、投资机构及政府与行业组织，可应用于产能评估、服务定级、采购对标、投资测算与产业统计等多类场景。

3.1 标准体系设计原则

3.1.1 标准以有效产出为核心

TO-TFS 1.0 标准坚持以有效产出（Good Token）为核心设计原则。针对当前行业普遍将峰值吞吐、总处理量等同于可交付产能的口径偏差，标准将满足时延阈值、生成准确率、并发稳定性等服务质量约束的生成词元定义为有效产出，剔除预填充、缓存命中、失败重试等非有效计量项，真实还原词元工厂的可交付产能水平，为采购结算、产能对标与投资测算提供统一公允的计量基准。

3.1.2 四项原则保障标准落地

标准以“有效产出、质量优先、口径统一”为核心原则，规避行业夸大产能、模糊成本的乱象，与 NVIDIA、IDC 的行业评估标准保持兼容：

质量优先原则：只有满足时延、可用性、准确率标准的 Good Token 才计入有效产能，缓存 Token、输入 Token、无效 Token 不计入产出；

更多免费行业报告 并购家 www.ipoipo.net

场景分类原则：按真实业务场景划分四类基准工况，统一负载参数、测试方法，不同场景的产能、成本、定价可对比；

边界透明原则：所有产能、成本数据必须配套披露硬件、软件、功耗、利用率、场景参数，不单独发布无边界的产能或成本数据；

三级验证原则：测算数据分为自测、第三方见证、实际运营三级验证，仅通过实测验证的数据可用于投资评估或商务采购。

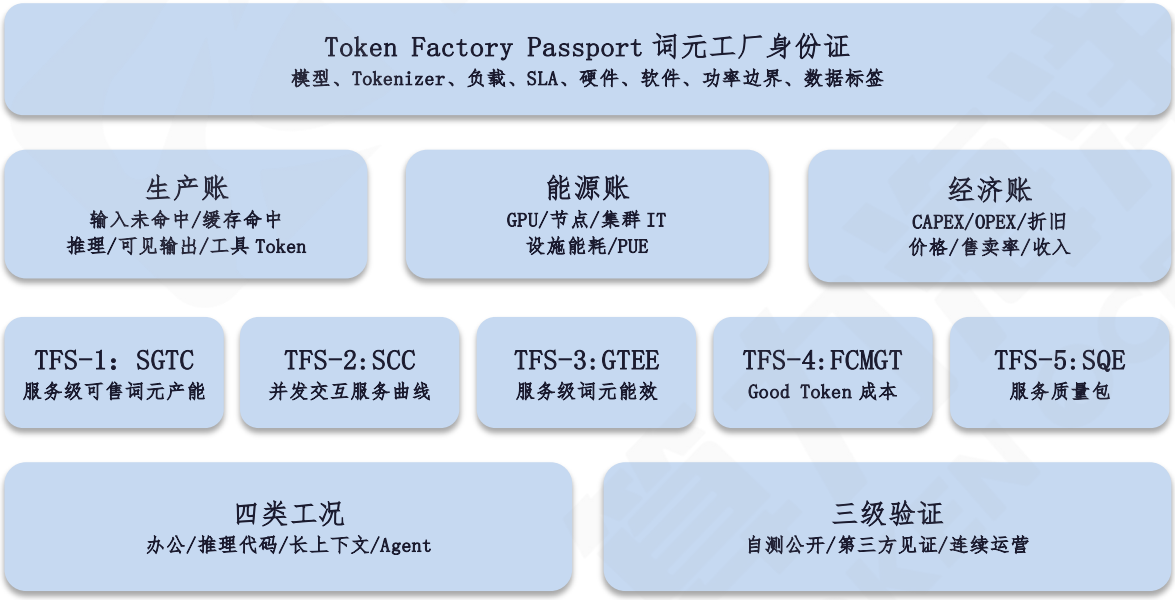
3.1.3 统一口径规范产能统计

TO-TFS1.0 标准彻底解决行业“口径混乱、数据虚高、无法对比”的乱象，统一以 Good Token 为唯一产能计量单元，区分总 Token、缓存 Token、输出 Token，所有产能、成本、能效数据必须绑定场景、负载、SLA、硬件参数，杜绝虚假宣传。

3.2 TO-TFS1.0 总体架构

用一张工厂身份证连接生产、能源与经济三本账：

图 10 TO-TFS1.0:词元工厂评价架构



3.2.1 工厂身份证绑定全维度参数

词元工厂身份证（Token Factory Passport）是结果不可拆分的最小发布单元。任何 TPS、Token/日、能效、成本或 SLA 数字，都必须与其模型版本、Tokenizer、负载、质量门槛、硬件、软件、并行策略、缓存、P/D 配置、功率边界、测试时长和证据标签绑定。

表 21 词元工厂身份证信息

身份域	强制披露字段
模型	模型全名与版本/哈希、参数与激活参数、上下文、量化精度、质量基线
Tokenizer	名称与哈希；中文千字、英文千词或指定语料的 Token 化系数
负载	输入/输出长度分布、推理词元、工具调用、多轮轨迹、到达模式、缓存命中率
服务	Batch/Server/Interactive/Agentic 等级；TTFT、TPOT、E2E、错误率与可用性
系统	GPU/节点数、网络、存储、推理框架版本、P/D、并行、投机解码
能源与证据	GPU/IT/设施功率、PUE、测量仪表、测试时长、原始日志摘要、来源类型与测量状态

3.2.2 三本账打通商业化全链路

生产账：记录输入未命中、缓存命中、推理输出、可见输出、工具 Token 和 Good Token 门槛。

能源账：区分 GPU、节点、集群 IT 和设施总能耗，记录 PUE、测量点和时间同步。

经济账：分开记录 CAPEX、现金 OPEX、折旧/摊销、融资/许可、价格实现率、售卖率与收入。

3.2.3 三级验证确保结果可复现

办公交互、推理/代码、长上下文和 Agentic 负载的输入输出结构、缓存复用和用户速度差异很大，不能用一个综合数字代表。验证则从自测公开、第三方见证逐步走向连续运营，形成从技术可复现到 SLA 可承诺的证据链。

3.3 TFS-5 五维核心指标

五个指标组成向量，避免把复杂服务压缩成一个失真的综合分数。

3.3.1 五维指标构建评价体系

标准定义五大类核心指标，覆盖产能、体验、能效、成本、质量，完全适配词元工厂的测算与评估需求：

TFS-1.SGTC 服务级可售词元产能：合规有效输出词元日产能，核心经营指标

TFS-2.SCC 并发交互服务曲线：平衡吞吐与用户时延体验，规避高吞吐低质量问题

TFS-3.GTEE 服务级词元能效：单位电力有效词元产出，算电协同核心指标

TFS-4.FCMGT 百万 Good Token 成本：全生命周期综合成本，投资对比核心指标

TFS-5.SQE 服务质量包络：P50/P95/P99 时延、可用性、错误率，服务履约核心指标

更多免费行业报告 并购家 www.ipoipo.net

表 22 词元工厂核心效能指标

指标名称	指标含义	计量单位	测算依据
服务级可售词元产能（SGTC）	集群在标准工况下，可稳定对外交付的有效 Good Token 产出量	亿 Token/日	集群有效吞吐×年有效运营时长
并发-交互服务（SCC）	集群在不同并发用户规模下的吞吐率、时延、体验平衡曲线	Token/s、ms	标准工况下的压测实测数据
服务级词元能效（GTEE）	单位电力消耗对应的有效 Good Token 产出量	百万 Token/kWh	有效 Token 产能/集群总耗电量
百万 Good Token 成本（FCMGT）	生产并交付 100 万枚有效 Good Token 的全综合成本	元/百万 Token	年化全成本/年有效 Token 产能
服务质量包（SQE）	服务时延、可用性、错误率、性能波动率的全维度约束	%、ms	标准工况下的连续运营实测数据

3.3.2 统一公式实现可复可审

TFS-1 服务级可售词元产能(SGTC)

核心公式 SGTC_out=Σ0_visible,r×1(success∧quality∧SLO)/T；日产能=SGTC_out×86,400。测试不足 24 小时必须标注为估算日产能。

SGTC 只统计在窗口 T 内成功、质量合格并满足服务目标的用户可见输出。总处理 Token 可补充披露，但不得替代 SGTC_out。正式 TFS 验证应直接统计逐请求门槛，不得用固定 Goodput 折扣替代质量与 SLO 证据。

服务级输出 Goodput：$SGTC_{out} = \sum O_{visible,r} \times 1(\text{success} \wedge \text{quality} \wedge \text{SLO})/T$，只有成功、质量合格且满足 SLO 的请求输出才计入。

服务级日产能：$SGTC_{out,day} = SGTC_{out} \times 86{,}400$。若测试不足 24 小时，应标注“估算日产能”。

总处理 Token（补充）：$SGTC_{total} = I_{miss} + I_{hit} + O_{reason} + O_{visible} + O_{tool}$。总量必须同时给出分项，不得只公布总 Token 而隐藏输出 Token。

TFS-2 并发-交互服务曲线(SCC)

等效流式并发 C_stream(v)=SGTC_out/v，其中 v 是目标单用户输出速度。该值表示同时处于 Decode 阶段的等效输出流，不等于注册用户、DAU、同时在线人数或端到端 QPS。

每个并发点必须同时记录系统 Good 输出 TPS、单用户输出速度、P50/P95/P99TTFT、TPOT、端到端时延和错误率。将这些点绘成 Pareto 曲线，才能识别高吞吐是否以不可接受的用户等待为代价。

单用户交互速度：$TPS_{user} = O_{visible,trajectory}/(E2E_{time} - tool_{delay} - think_{delay})$

系统吞吐：$TPS_{system} = \sum GoodO_{visible}/T$

等效流式并发：$C_{stream}(v) = SGTC_{out}/v$（$v$为目标单用户输出速度，如 20Token/s）。该值不等于 DAU。

TFS-3 服务级词元能效(GTEE)

> **设施边界** GTEE_DC=GoodO_visible/E_facility；推荐单位为百万 Good 输出 Token/kWh。其倒数为 kWh/百万 Good 输出 Token。瞬时 Good Token/s/W 等价于 Good Token/J。

Tokens/W 如果缺少“每秒”会产生量纲歧义。结果必须注明 GPU、节点 IT、集群 IT 或设施总功率，设施边界同时披露 PUE，并遵循 ISO/IEC30134-2 与适用的国内 PUE 规范。

设施边界能效：$GTEE_{DC} = GoodO_{visible}/E_{facility}$（推荐单位：百万 Good 输出 Token/kWh）。

瞬时表达：$GoodToken/s/W = GoodToken/J$。

PUE：$PUE = E_{facility}/E_{IT}$。

TFS-4 百万 Good Token 成本(FCMGT)

> **现金与全成本**
>
> FCMGT_cash=AnnualCashOPEX/AnnualGoodO_visible×10^6；
>
> FCMGT_full=(AnnualizedCAPEX+CashOPEX+Finance/License)/AnnualGoodO_visible×10^6。

现金成本用于判断短期运营能否覆盖电力、IDC、维保、软件、人员和商务成本；年化会计成本用于投资回报、采购与跨工厂比较。本案例未纳入融资、税务、营运资金与残值，不能把 1.9718 亿元直接解释为项目生命周期经济成本。折旧年限、残值、融资成本、税务、汇率和容量利用率必须披露。

现金成本：$FCMGT_{cash} = \text{AnnualCashOPEX}/\text{AnnualGood}O_{visible} \times 10^6$。

全成本：$FCMGT_{full} = (\text{AnnualizedCAPEX} + \text{CashOPEX} + \text{Finance})/\text{AnnualGood}O_{visible} \times 10^6$。

TFS-5 服务质量包（SQE）

表 23 服务质量包信息列表

指标	统计量	解释
首 Token 时延	TTFTP50/P95/P99	反映 Prefill、排队和缓存查找
每 Token 时间	TPOTP50/P95/P99	其倒数近似单用户流式输出速度
端到端时延	E2EP50/P95/P99	Agent 需分离模型净时延与工具等待

更多免费行业报告 并购家 www.ipoipo.net

可靠性	成功率、错误率、可用率	区分维护、故障、限流、拒答与超时
稳定性	吞吐/时延波动、尾延迟	短测峰值不能替代连续运营
质量	任务得分、合格率、长度合规	防止通过截断或降质换取吞吐

3.4 测试流程与验证机制

测试的可复现性来自固定条件、完整边界和原始证据，而不是更长的设备清单。

3.4.1 四类工况覆盖主流场景

表 24 词元工厂标准工况信息列表

代码	负载定义	代表场景	建议等级
TF-O	2K 输入/512 输出；thinking 关闭；0%缓存	办公问答、摘要、邮件、方案	Interactive
TF-R	参考 800 输入/3,880 输出；thinking 开启	复杂推理、代码生成	Server/Interactive
TF-L	32K/2K 与 120K/2K；分别测 C0/C70	代码仓库、长文档、RAG	Server
TF-A	多轮轨迹、工具调用、会话缓存	代码 Agent、多智能体工作流	Agentic Pareto
TF-B	模型官方或 MLPerf 数据集	离线摘要、批量生成	Batch（辅助）

3.4.2 四级服务匹配差异化需求

表 25 词元工厂服务等级列表

等级	参考门槛	测试要求	适用
B｜Batch	无交互 TTFT 约束	最大 Good 输出吞吐；错误率≤0.5%	离线加工
S｜Server	P99TTFT≤3s；P99TPOT≤80ms	泊松到达或生产流量；P99 门槛	通用 API
I｜Interactive	P99TTFT≤1.5—2s；P99TPOT≤15—30ms	同时发布 TPS_user 与 TPS_system	聊天/办公
A｜Agentic	按 20/50TPS_user 等固定交互点	多轮 E2E、工具延迟、任务成功率	Agent
TF-B	模型官方或 MLPerf 数据集	离线摘要、批量生成	Batch（辅助）

3.4.3 标准流程保证结果可比

冻结工厂身份证：模型/权重、Tokenizer、精度、框架版本、负载、SLO、P/D、缓存和功率边界。

完成预热和稳定窗口，记录版本、时钟、网络、温度与故障状态，避免把冷启动或临时超频混入长期能力。

运行并发阶梯，逐点采集请求日志、系统吞吐、用户速度、TTFT/TPOT/E2E、质量和功率。

生成产能、能源与经济三本账；区分铭牌、服务级 Goodput、商业销量和可计费收入。

发布结果卡、配置摘要与原始日志哈希；第三方见证或连续运营验证时保留可审计证据。

3.4.4 双轴标签标注数据可信度

把“来源可靠性”和“是否真实测量”混在单一等级里容易产生误读。本白皮书采用两轴标签：前一位说明来源类型，后一位说明数据状态。例如 B-I 表示独立公开基准中的插值点，P-X 表示项目研究外推，V-S 表示厂商公开规格。

表 26 词元工厂证据信息列表

证据标签	含义	适用场景
V-S	Vendor Spec：厂商规格书数据	硬件参数、显存、带宽
V-M	Vendor Measured：厂商实测数据	推理性能基准
P-A	Planning Assumption：规划假设	SLA 吞吐、利用率（需 POC 替换）
P-X	Planning External：外部规划数据	电价、机柜租金
T-W	Third-party Witnessed：第三方见证	认证测试、招投标验证

更多免费行业报告 并购家 www.ipoipo.net

4. 中国词元工厂发展趋势

4.1 规模扩容与业态升级

100MW 以上甚至 GW 级智算中心与词元工厂的必要性开始凸显。只有达到足够大的能源、算力与网络密度，才有可能形成可调度的算力池、可消纳的长周期负载、可优化的电力合约、可摊薄的基础设施成本和可持续满载的任务组合。未来企业级市场将呈现 AIDC 向词元工厂演进、混合算力成为刚需、异构集群成为常态三大趋势。智算中心的核心产能单位将彻底转变为"在多少 MW 约束下，以多高利用率持续生产多少高质量词元"。

4.2 算电协同与成本优化

在空转与满载节点消耗的资本成本和基础能耗高度接近的情况下，单位词元成本可能相差一个数量级。利用率、负载编排、任务混部、峰谷电价响应、训练与推理协同、跨集群调度、供电稳定性和散热效率，已经不再是数据中心运营细节，而是决定词元成本曲线的核心经济学变量。

4.3 交付模式向价值转型

当前 2026 年中国智算中心产业的主线不是单纯"项目变多"，而是三个层面的同时变化：智算中心项目仍在快速补库存；100MW 以上项目把竞争维度推向电力、资本和工程组织能力；词元工厂把算力供给的商业表达从机柜、PFLOPS、GPU 卡、算力小时进一步推进到生成能力、词元产出和运营套餐。

4.4 软件优化与调度提效

硬件资源的价值红利将逐步见顶，软件优化、精细运营将成为长期核心竞争力：同样硬件条件下，KV 缓存优化、算子融合、多模型调度、P/D 动态编排技术，可将有效吞吐提升 1.5-2 倍，单位词元成本降低 30%以上。

4.5 国产渗透与适配降本

国产 AI 芯片在词元工厂的应用占比将持续提升，依托国产化政策、成本优势与政企专属订单，在政务、金融、教育、国企等关键行业场景占据重要份额。国内芯片厂商、模型厂商与算力服务商将联合搭建标准适配基地，沉淀行业标准算子库，将通用场景的算子适配率提升至 90%以上。

5. 中国词元工厂发展风险

5.1 流量爬坡利用率不足风险

> **核心影响**
>
> 利用率每下降 10%，项目 IRR 下滑 10-17 个百分点，回收期拉长 0.5-1 年。

风险定义：新建项目投产初期，客户流量储备不足，商业利用率长期低于 60%，固定成本无法有效摊薄，现金流显著承压。

量化影响：利用率从 70%降至 50%，海外路线 IRR 下降约 18 个百分点，国产路线 IRR 下降约 12 个百分点，静态回收期拉长约 1 倍。

规避建议：采用"算力租赁+Token 分成+集采订单"的混合业务模式，投产初期以传统算力租赁填充空闲产能，锁定基础现金流；提前与头部模型厂商、运营商签订长期框架订单，约定最低采购量，缩短流量爬坡周期。

5.2 行业词元定价下行风险

> **核心影响**
>
> 单价下跌 30%可直接抹平项目全部利润。

风险定义：行业推理产能集中释放，市场供过于求，词元成交单价持续下行，压缩毛利空间。

量化影响：单位词元定价下跌 30%，海外路线 IRR 下降约 16 个百分点，国产路线 IRR 下降约 10 个百分点，回收期拉长 40%以上。

规避建议：锚定高附加值行业场景，延伸至行业解决方案、私有化部署、专属算力租赁等增值服务，降低对通用公开市场的价格敏感；签订长期定价对冲协议，与大客户约定长期采购价。

5.3 国产硬件软件适配风险

> **核心影响**
>
> 适配不足导致吞吐下降 15%，IRR 下滑 8 个百分点以上。

更多免费行业报告 并购家 www.ipoipo.net

风险定义：国产芯片算子适配率不足，推理吞吐、时延表现逊于海外芯片，实际有效产出低于测算预期。

量化影响：有效吞吐系数从 65%降至 50%，国产路线单位词元成本上升约 23%，IRR 下降约 15 个百分点，回收期拉长超过 2 年。

规避建议：提前与芯片厂商、模型厂商联合开展算子适配验证，沉淀场景化标准算子库；优先选择成熟度高的通用场景，重点投入资源优化 KV 缓存、多模型调度等核心环节。

5.4 电力供应与成本上涨风险

核心影响

电价上涨 0.1 元/kWh，年化利润下滑 7-9%。

风险定义：电力紧缺、能耗双控、绿电指标不足导致限电停机，或工业电价上调，直接拉高单位词元成本。

量化影响：电价上涨 0.1 元/kWh，海外路线单位词元成本上升约 12%，IRR 下降约 9 个百分点；集群年停机时长超过 300 小时，IRR 下降约 5 个百分点。

规避建议：选址优先保障绿电供应，配套建设分布式光伏、储能系统，参与电力现货交易，降低平均用电成本；配置备用电源，与电网签订优先级供电协议。

5.5 硬件迭代与估值贬值风险

核心影响

新芯片性能提升拉低旧集群市场定价，资产实际残值低于折旧测算值。

风险定义：AI 芯片技术迭代加速，新芯片性能提升 2-3 倍，存量资产面临快速贬值风险。

量化影响：若硬件迭代周期从 3 年缩短至 2 年，项目实际残值低于测算值约 30%，全生命周期净现金流减少约 15%。

规避建议：采用短期限、柔性、模块化集群扩容方案，分阶段投入建设；与供应商签订回购、以旧换新协议，对冲硬件残值贬值风险；优先选择软件可优化性强的硬件平台。

5.6 高负债投入财务杠杆风险

核心影响

项目高负债建设，财务成本过高，回报无法覆盖资金成本。

风险定义：资产负债率超过 70%，综合资金成本超过年化 6%，项目静态回收期拉长约 1 年，IRR 下降约 8 个百分点。

量化影响：合理安排融资结构，匹配长期低息产业资金、算力专项基金，控制资产负债率在 60%以下；采用"滚动开发、以养代建"模式，一期项目产生现金流后再投入二期扩容，降低整体财务成本。

规避建议：合理安排融资结构，匹配长期低息产业资金、算力专项基金，控制资产负债率在 60%以下；采用"滚动开发、以养代建"模式，一期项目产生现金流后再投入二期扩容。

更多免费行业报告 并购家 www.ipoipo.net

6. 核心结论与行业落地建议

6.1 核心研究结论

结论 1：硬件选型标准化是盈利基础：国产路线必须选择昇腾 910B 等通用透明量产型号，非标高价型号无长期经济性；海外 H200 商业化弹性最优，适配市场化场景。

结论 2：5 年折旧口径更贴合真实投资价值：相较于激进 3 年折旧，5 年合规折旧剔除虚高短期收益，真实反映项目长期稳健盈利能力，适配国资、上市公司审计要求。

结论 3：收入精细化测算决定决策准确性：P/D 约束、缓存计费、商业利用率、价格实现率四维校准，是避免项目收益误判的核心，统一六层级测算模型可作为行业标准。

结论 4：双路线盈利边界清晰：海外路线高投入高回报、弹性充足；国产路线低投入稳收益、合规性强，办公短问答为双路线唯一通用优质盈利场景。

结论 5：利用率是项目生死红线：无论何种硬件，商业利用率稳定≥65%是盈利核心前提，低于 60%项目基本丧失投资价值。

结论 6：商业模式决定风险收益结构，混合模式为行业最优选择：单纯自建 API 销售模式客户拓展风险最高；纯基础设施租赁模式毛利最低；行业头部项目普遍采用分成合作+集采订单+增值服务的混合模式。

6.2 分主体落地建议

投资决策层

摒弃粗放硬件堆叠，优先锁定绿电资源与长期订单，按 5 年折旧合规测算，分阶段滚动扩容，优先落地标准化通用硬件集群。

- 建立标准化实测测算模型，将投产爬坡期、利用率波动、定价下行风险纳入敏感性分析，不盲目放大理论产能，高估项目收益。
- 优先布局算电协同一体化项目，锁定长期绿电指标，优先与头部模型厂商、运营商签订长期流量框架协议，夯实项目盈利基础；
- 技术路线匹配业务场景，通用大规模推理优先选择海外高端芯片路线，政企国产化专属场景优先选择国产芯片路线。

产业运营层

聚焦软件精细化优化，标准化场景负载配比，提升有效吞吐与利用率，以混合商业模式对冲市场风险。

- 重点提升推理软件优化、精细运营调度能力，聚焦算子融合、KV 缓存、多模型编排、负载调度核心技术，通过软件优化提升有效吞吐，降低单位成本。严格遵循 TO-TFS1.0 标准化产能评价体系，公开实测数据，明确场景边界。
- 严格遵循 TO-TFS1.0 标准化产能评价体系，公开实测数据，明确场景边界，使用标准化 Good Token 参与行业采购、项目评估与商务合作；
- 布局全链路增值服务，从单纯卖词元转向行业解决方案、私有化部署、专属算力托管服务，提升客户溢价能力，降低对公开市场定价的依赖。

政企采购层

采用 Good Token 标准化计量结算，以实测能效、成本、SLA 为采购依据，拒绝虚高产能宣传，优先国产化合规标准化项目。

- 采用标准化集采模式，将有效 Good Token 产出量、SLA 质量、实测能效指标纳入采购履约条款，按实际有效产出结算费用，避免资源浪费。
- 优先选择具备长期绿电供应能力、软件优化能力的服务商，综合评估单位词元成本、服务质量、产能稳定性，不单纯以低价作为采购决策依据；
- 规划分层算力储备体系，实时交互业务配置低时延边缘节点，非实时大吞吐量任务配置西部大规模词元产能，实现成本与体验的平衡。

更多免费行业报告 并购家 www.ipoipo.net

白皮书数据可信度评估体系

为解决行业数据杂乱、假设随意、结论不可信的问题，本白皮书建立三级可信度评级+来源溯源+误差可控的标准化体系，所有核心测算参数、公式、结果均标注等级与来源，适配投资决策与审计要求。

三级可信度评级定义

可信度等级	定义	来源示例	误差区间
A 级（权威验证级）	官方公开实测数据、头部第三方行业报告	NVIDIA 官方 GPU 产品资料、科智传媒《算力产业核心内刊》、国家能源局算力枢纽电价公示	±5%
B 级（行业共识级）	行业通用调研数据、头部企业技术标准、集采公开报价	中国电信 2026 年 Token 服务集采公告、华为昇腾智算硬件规格书	±10%
C 级（合理假设级）	行业标准工况假设、未公开项目测试数据、场景化模拟校准值	大模型平均输入输出长度假设、缓存命中率行业通用值、软件运维成本比例假设	±15%

核心数据来源可信度列表

数据类别	来源机构/文件	核心用途	可信度
宏观产业数据	科智咨询《中国智算中心项目跟踪数据库》	1300+智算中心、词元工厂全景统计	A 级
H200 硬件规格	NVIDIA 官方规格页（141GBHBM3e、4.8TB/s、700W）	硬件参数、功耗、显存基准	A 级
GLM-5.2 参数与定价	智谱 AI 官方 API 文档	模型部署节点、收入定价基准	A 级
DeepSeek-V4-Pro 参数与定价	DeepSeek 官方 API 文档	低价场景对比、模型部署基准	A 级
推理性能方法论	NVIDIA GenAI-Perf 文档、MLPerf Inference 规则、Microsoft Splitwise 论文	三层产能模型、P/D 分离理论、SLO Goodput	A 级
昇腾 910B 硬件参数	华为昇腾 Atlas800IA2 产品规格书、政企招标均价	国产路线硬件基准、定价	A 级
行业集采价格	中国电信 2026 年 Token 生成能力集采公告	PRR 基准、政企折扣率验证	A 级
P/DSLA 吞吐	本报告规划假设（基于 vLLM 基准外推）	单节点 TPS_P/TPS_D 基准值	C 级·需 POC
商业利用率 CUR	算力海洋行业调研（已投产词元工厂运营中位数）	营收折算、三情景分析	B 级
PUE 与电价	中国数据中心产业联盟能效报告 2026、西部绿电集采价	电力成本测算	A 级

数据可信度结构

全报告核心测算数据中，85%为 A 级权威数据、12%为 B 级行业共识数据、3%为 C 级可控假设数据，所有 C 级假设均完成多情景敏感性校验，测算结果具备极强的产业参考价值与决策落地性。

测算基准边界说明

- 集群基准规模：以行业主流落地规格——128 台 8 卡 AI 服务器集群为标准测算样本，对应 1024 颗 GPU。
- 硬件技术路线：分两类基准测算体系，海外高端芯片路线（NVIDIAH2008 卡服务器）、国产芯片路线（华为昇腾 910B8 卡服务器）。
- 能效基准指标：数据中心 PUE 统一取 1.3，为国内西部算力枢纽智算中心的行业中位数水平。
- 会计折旧口径：AI 服务器按 5 年直线折旧，无残值；基础设施按 10 年直线折旧；软件投入按 3 年摊销。
- 运营时长假设：年有效运营时长按 8200 小时校准，扣除例行运维、故障检修、电力调度的停机时间。
- 场景负载标准：统一采用 TO-TFS1.0 标准定义的四类基准工况，不采用理论峰值算力测算。
- 收益计提规则：收入按实际可确认的现金流入计量，不计算增值税返还、政府补贴、税收优惠。

数据局限性与使用边界声明

本报告测算基于 2026 年 8 月公开市场数据与行业标准假设，实际项目收益受硬件采购价格波动、客户流量爬坡、技术迭代、政策变化等多重因素影响。所有 C 级假设数据（SLA 吞吐、商业利用率等）必须通过项目 POC 实测后替换，方可作为最终投资决策依据。本报告不构成任何投资建议，使用者应结合自身情况独立判断。

更多免费行业报告 并购家 www.ipoipo.net

附录

附录一：投入产出模型测算公式体系

【产能测算（P/D 双向约束）】

$R_P = N_P \times TPS_P / I_{miss}$

$R_D = N_D \times TPS_D / O$

$R = \min(R_P, R_D)$

【P/D 平衡配比】

$N_P / N_D = (TPS_D \times I_{miss}) / (TPS_P \times O)$

（D 侧留 5%~10%余量）

【单请求分层收入】

$Rev_{req} = I_{miss} \times P_{miss} + I_{hit} \times P_{hit} + O \times P_{out}$

（Token 均除以 1,000,000）

【最终年化营收（六层级）】

$Rev_{final} = R \times Q_{pass} \times 31{,}536{,}000 \times Rev_{req} \times Rate \times CUR \times PRR$

【成本与折旧】

年化硬件折旧=硬件 CAPEX÷5

年化基础设施折旧=配套 CAPEX÷10

年化软件摊销=软件投入÷3

年化全成本=硬件年折旧+基础设施年折旧+软件年摊销+年化现金 OPEX

【ROI 核心公式体系】

EBITDA=年化营收-年化现金 OPEX

年度净利润=(EBITDA-年化总折旧摊销)×(1-25%所得税率)

年度净现金流=年度净利润+年化总折旧摊销

静态投资回收期=初始总 CAPEX÷年均净现金流

五年现金 ROI=(5×年均净现金流-初始 CAPEX)÷初始 CAPEX

附录二：硬件配备基础参数总表

参数类别	参数名称	H200 路线	昇腾 910B 路线	可信度
硬件基准	8 卡服务器单价	350 万元/台	140 万元/台	A 级
	集群规模	128 台/1024GPU（120 生产+8 预留）		A 级
	单台整机功耗	14.5kW	11.5kW	A 级
吞吐基准	短问答/长文 P 吞吐	30,000t/s/台	19,500t/s/台（65%）	C 级
	短问答/长文 D 吞吐	2,300t/s/台	1,495t/s/台（65%）	C 级
	代码 AgentP/D 吞吐	15,000/1,500t/s	11,250/1,125t/s（75%）	C 级
成本基准	基础设施配套占比	15%	15%	A 级
	软件平台占比	8%	10%（含算子适配）	A 级
	年维保费率	3%/年	3%/年	A 级
运营基准	数据中心 PUE	1.3	1.3	A 级
	西部绿电基准电价	0.38 元/kWh	0.38 元/kWh	A 级
	年有效运营时长	8200 小时	8200 小时	A 级
经营基准	商业利用率 CUR	短问答/长文 70% 代码 65%	短问答/长文 75% 代码 65%	B 级
	价格实现率 PRR	65%	65%	A 级
	SLA 通过率 Q_pass	95%	95%	B 级
汇率	美元兑人民币	7.2	7.2	A 级

附录三：白皮书专业术语对照表

缩写	英文全称	中文定义
CAPEX	Capital Expenditure	资本开支
OPEX	Operating Expenditure	运营开支
TCO	Total Cost of Ownership	总拥有成本
CUR	Commercial Utilization Rate	商业利用率
GUF	Goodput Utilization Factor	有效吞吐系数
CPM	Cost Per Million Tokens	百万 Token 成本
GTEE	Good Token Energy Efficiency	单位电力有效 Token 产出
PRR	Price Realization Rate	价格实现率
NPV	Net Present Value	净现值
IRR	Internal Rate of Return	内部收益率
PBP	Payback Period	静态投资回收期
SGTC	Serviceable Good Token Capacity	服务级可售词元产能
SCC	Concurrency-Interactive Service Curve	并发-交互服务曲线
SQE	Service Quality Envelope	服务质量包络
TTFT	Time to First Token	首 Token 时延
TPOT	Time per Output Token	单 Token 生成时延
P/D 分离	Prefill/Decode Separation	预填充/解码阶段资源分离部署
Good Token	满足 P95 时延、可用性、准确率 SLA 的用户侧有效输出 Token	

附录四：模型、硬件公开价格核验表

对象	已核验事实	证据等级
NVIDIAH200	141GBHBM3e/卡，4.8TB/s；8 卡服务器为基本设备单元	A 级
GLM-5.2	744B 总参数/40B 激活，1M 上下文；FP8 可在单台 8×H200 部署	A/B 级
GLM-5.2 价格	$1.40/M 普通输入、$0.26/M 缓存输入、$4.40/M 输出	A 级
DeepSeek-V4-Pro	1.6T 总参数/49B 激活，1M 上下文；官方 vLLM 配方推荐单	A/B 级

更多免费行业报告 并购家 www.ipoipo.net

	台 8×H200	
DeepSeek-V4-Pro 价格	$0.435/M 未命中输入、$0.003625/M 命中输入、$0.87/M 输出	A 级
H200P/DSLA 吞吐	P=30K/15Kt/s，D=2.3K/1.5Kt/s（按场景）	C 级·规划假设

证据等级说明：

A 级（可直接引用）：官方模型参数、上下文、部署节点、API 公开价格、H200 显存/带宽/TDP。

B 级（仅作边界参照）：SGLang/vLLM 框架在其他硬件或其他上下文上的基准。

C 级（本报告规划假设）：H200 上 P 阶段与 D 阶段的 SLA 吞吐、文生图秒数、商业利用率与峰谷平滑效果。必须通过 POC 实测后替换。

附录五：权威数据溯源清单

编号	来源名称	发布时间	核心用途
[1]	国家数据局新闻发布会：我国日均 Token 消耗量突破 30 万亿	2025.08	宏观 Token 消耗数据
[2]	北京经济技术开发区：北京首座词元工厂日产 1.4 万亿词元	2026.06	词元工厂产能基准
[3]	中国信息通信研究院：Token 云服务与标准化相关研究	2026	行业标准化参考
[4]	Microsoft Research：Splitwise-Efficient Generative LLM Inference	2024	P/D 分离理论基础
[5]	ML Commons：MLPerf Inference Datacenter benchmark	持续更新	推理性能标准化测试
[6]	ISO/IEC30134-2:2026：数据中心 PUE 标准	2026	能效测量边界
[7]	YD/T6232-2024：数据中心 PUE 评估和验收规范	2024	国内能效合规标准
[8]	NVIDIAD GXH100/H200 User Guide	2026	H200 硬件规格、功耗基准
[9]	NVIDIA Technical Blog：Llama3.1405Bon8×H200throughput	2024	推理性能基准
[10]	NVIDIA Dynamo：GLM-5.2/DeepSeek-V4deploymentrecipe	2026	模型部署配置基准
[11]	DeepSeek API Docs：Models&Pricing	2026.08	模型公开定价
[12]	Z.AI Developer Document：GLM-5.2Pricing	2026.07	GLM 模型公开定价
[13]	Inference X：DeepSeek-R1 公开插值基准	2026	推理性能交叉验证
[14]	科智咨询《中国智算中心项目跟踪数据库》	2026H1	宏观产业项目统计
[15]	中国电信 2026 年 Token 生成能力服务集采公告	2026	行业集采价格基准
[16]	中国数据中心产业联盟《智算中心能效报告》	2026	PUE、运营时长基准
[17]	国家能源局西北/华北电网工业电价公示	2026	西部绿电电价基准
[18]	华为昇腾 Atlas800IA2 产品规格书	2026	昇腾 910B 硬件参数
[19]	DistServe：Disaggregating Prefilland Decoding for Goodput-optimized LLM Serving	2024	SLOGoodput 理论

科智咨询

科智咨询是国内领先的数字领域专业咨询机构，致力于为政府、企业和社会组织提供全方位的数字科技战略规划、数字行业与市场洞察、投资决策与管理等服务，助力企业把握数字经济发展机遇、构建数字化核心能力，在新一轮产业和技术革命浪潮中取得竞争优势。

如欲了解更多信息，请联系电话：13161291179

算力海洋

算力海洋（Token Ocean）专注于连接分布式算力供给、AI模型能力、云网资源与企业需求。依托AI服务平台、专业运营体系与生态伙伴网络，为客户提供算力接入、模型调用与部署、AI网关、使用计量、技术支持及全球化落地等一体化服务，致力于降低AI使用门槛，提高基础设施效率，推动中国AI能力与全球产业需求高效协同。

如欲了解更多信息，请联系电话：17718310723

免责声明

本研究报告由科智咨询联合算力海洋共同撰写和制作。报告仅作为研究内容介绍之用。未经许可，不得对文中内容进行任何形式的复制。尽管我们对所依据的信息和资料保持高度谨慎，但无法对其中的准确性和完整性做出绝对保证，请勿绝对化地加以利用。文中所述观点有可能在未经知会的情况下进行调整。报告内容亦非根据任何公司所处独特环境而提供的具体咨询建议。如果您需要就文中提及内容获得具体建议及进一步的详细资料，请与我们进行联系。

更多免费行业报告 并购家 www.ipoipo.net