

Hong Kong Exchanges and Clearing Limited and The Stock Exchange of Hong Kong Limited take no responsibility for the contents of this announcement, make no representation as to its accuracy or completeness, and expressly disclaim any liability whatsoever for any loss howsoever arising from or in reliance upon the whole or any part of the contents of this announcement.



Z.AI Co., Ltd.

北京智譜華章科技股份有限公司

(A joint stock company established in the People's Republic of China with limited liability)

(Stock Code: 2513)

INTERIM RESULTS ANNOUNCEMENT FOR THE SIX MONTHS ENDED JUNE 30, 2026

The Board is pleased to present the unaudited condensed consolidated interim results of the Group for the six months ended June 30, 2026, together with the comparative figures for the corresponding period in 2025.

Certain amounts and percentage figures included in this announcement have been subject to rounding adjustments, or have been rounded to one or two decimal places. Any discrepancies between the total amounts listed in any tables, charts, or other data and the sum of the individual amounts listed therein are due to rounding.

FINANCIAL SUMMARY			
	For the six months ended June 30,		Year-on-Year
	2026	2025	Change
	<i>RMB in thousands</i>		<i>%</i>
	(Unaudited)	(Unaudited)	
Revenue	953,892	190,877	399.7
Gross profit	251,611	95,424	163.7
R&D expenses	(2,131,171)	(1,594,661)	33.6
Loss for the period	(2,071,992)	(2,357,852)	-12.1
Adjusted net loss for the period	(1,964,138)	(1,751,974)	12.1

	For the six months ended June 30,				Year-on-Year Change in Revenue Growth
	2026		2025		
	<i>RMB in thousands</i>	<i>% of total</i>	<i>RMB in thousands</i>	<i>% of total</i>	<i>%</i>
	(Unaudited)		(Unaudited)		
Revenue					
Open platform and API	825,176	86.5	29,100	15.2	2,735.7
Enterprise-level agents	55,561	5.8	13,739	7.2	304.4
Enterprise-level general-purpose large models	67,037	7.0	147,555	77.3	-54.6
Technical services and others	6,118	0.7	483	0.3	1,166.7
Total	953,892	100.0	190,877	100.0	399.7

MANAGEMENT DISCUSSION AND ANALYSIS

Business Review

In the first half of 2026, the Company validated the deliverability of large models in the domain of coding – progressing from early-stage simple code recommendations, to full-file code generation, and then to the implementation of complete engineering projects (Agentic Engineering). It has fully traversed this trajectory: the model is no longer merely a tool, but possesses the capability to complete an entire task independently. Accordingly, its revenue structure shifted: the share of open-platform and API services – billed on a per-invocation basis – rose from 26.3% of the previous full year to 86.5%, while on-premises model deployment revenue fell from 73.7% to 13.5%. The restructuring of our core revenue mix in the first half of 2026 is essentially an inevitable consequence of a generational leap in model capability. Cybersecurity Co-work is an extension of the same capability curve. Software vulnerability discovery does not rely on memorization, but on the model’s reasoning depth and long-range autonomy – precisely the direction that naturally follows after coding capability crosses the threshold of engineering project delivery. During the Reporting Period, the Company’s model achieved a score of 84.5% on CyberGym (surpassing Fable 5 and GPT-5.6 Sol) and completed 130 tasks on ExploitGym (ranking second only to Fable 5 and GPT-5.6 Sol).

Underpinning these two trajectories is infrastructure. Since the beginning of 2026, the Company has fully implemented an All-in-Infra strategy, investing in both training and inference, and integrating domestic chips as a primary inference compute resource. It has achieved large-scale, low-cost inference across a cluster of over 100,000 domestic chips, with unit Token inference costs declining by 80% from the beginning of the year. Given the persistent global supply constraints on compute, this path ensures that the Company's cost curve remains within its own control. As at the end of the Reporting Period, the number of enterprise and developer users on the GLM MaaS platform exceeded 5.8 million. The Company's total revenue amounted to approximately RMB954 million (approximately USD142 million), representing **a nearly 400% increase year-on-year**. Among this, open-platform and API revenue amounted to approximately RMB825 million (approximately USD123 million), **up 2,735.7% year-on-year**.

In August 2026, the Company released GLM-5.3 and GLM-5.3-Flash (Ox-Alpha), and simultaneously made the GLM Coding Plan subscription available to all users, covering Lite, Pro, Max and Team editions – making it one of the earliest large-model services in China to deliver coding-agent capabilities under a subscription model. As of the date of this announcement, the GLM MaaS platform had over 7.4 million enterprise and developer users.

I. Strategic Focus

Intelligence upper bound defines the task boundary; task value defines the commercial space

At the beginning of 2026, we set out a formula: AGI Commercial Value = Upper Bound of Intelligence × Token Consumption Scale. This time, we would like to elaborate a bit on this.

The upper bound of intelligence determines which tasks the model can handle and how much user value can be created per Token. A higher intelligence upper bound means that the model can handle tasks with higher value density, and each unit of intelligence corresponds to greater economic output. Token consumption scale reflects the breadth and depth of intelligence utilization; the upper bound of intelligence first unlocks the task boundary, and the task boundary then determines where Tokens are assigned to in terms of value density. The same volume of Token used for casual chat versus fixing a production-environment vulnerability yields an output difference of several orders of magnitude.

Capability Ladder. The evolution of model capabilities is not a set of parallel product lines, but a ladder that can only be climbed sequentially:

Chat → Coding → Agent → Co-work → Autonomous AI

Chat delivers a single response; the task ends within one round, with the lowest value density but broadest coverage. Coding delivers runnable code, introducing the first objective acceptance criterion – it either runs or it does not. Agent delivers a completed multi-step task chain, where the model begins to invoke tools and handle intermediate failures by itself. Co-work delivers a work product that can be reviewed by a professional, entering real workflows and priced by completed tasks. Autonomous AI delivers continuous operation capability, where the model undertakes long-term objectives without human supervision. Each step upward is gated by a concrete technical threshold. From Chat to Coding, the threshold is verifiability of results; from Coding to Agent, it is long-horizon planning and error recovery; from Agent to Co-work, it is reliability to the level that professionals are willing to review rather than redo from scratch; from Co-work to Autonomous AI, it is whether the model can judge the correctness of its own outputs. If the threshold cannot be crossed, the business model for that level does not hold.

Market Space. As the task boundary moves upward, the market space expands correspondingly. By rough order-of-magnitude estimates: Chat corresponds to the search market of approximately USD100 billion – the answer given by the Internet era; Coding corresponds to roughly USD500 billion in software R&D expenditure; Co-work corresponds to approximately USD5 trillion in knowledge-work spending; and in the longer term, AI-driven re-architecting of operating systems and human-computer interaction, together with industry-wide upgrades, corresponds to an addressable space of about USD30 trillion. (The above market space figures are the Company’s rough estimates based on publicly available industry data, used solely to illustrate the correspondence between task tiers and market magnitudes, have not been independently verified, and do not constitute a forecast or guarantee by the Company of its business, financial condition or results of operations. The shareholders and investors of the Company should not place undue reliance on such figures.) Correspondingly, industry evaluation criteria are also shifting: from single-turn response quality and Token consumption volume to long-horizon task success rate, unit intelligence cost, result verifiability, and the actual business outcomes delivered to customers.

During the Reporting Period, the Company’s focus lay between the second rung (Coding) and the fourth rung (Co-work) of the ladder. The coding direction has been fully traversed; cybersecurity has entered the co-work validation phase; and other areas such as legal remain at early stages. The Company’s R&D expenditure for the first half of 2026 amounted to approximately RMB2.131 billion (approximately USD317 million), demonstrating a steadfast commitment to scaling new heights.

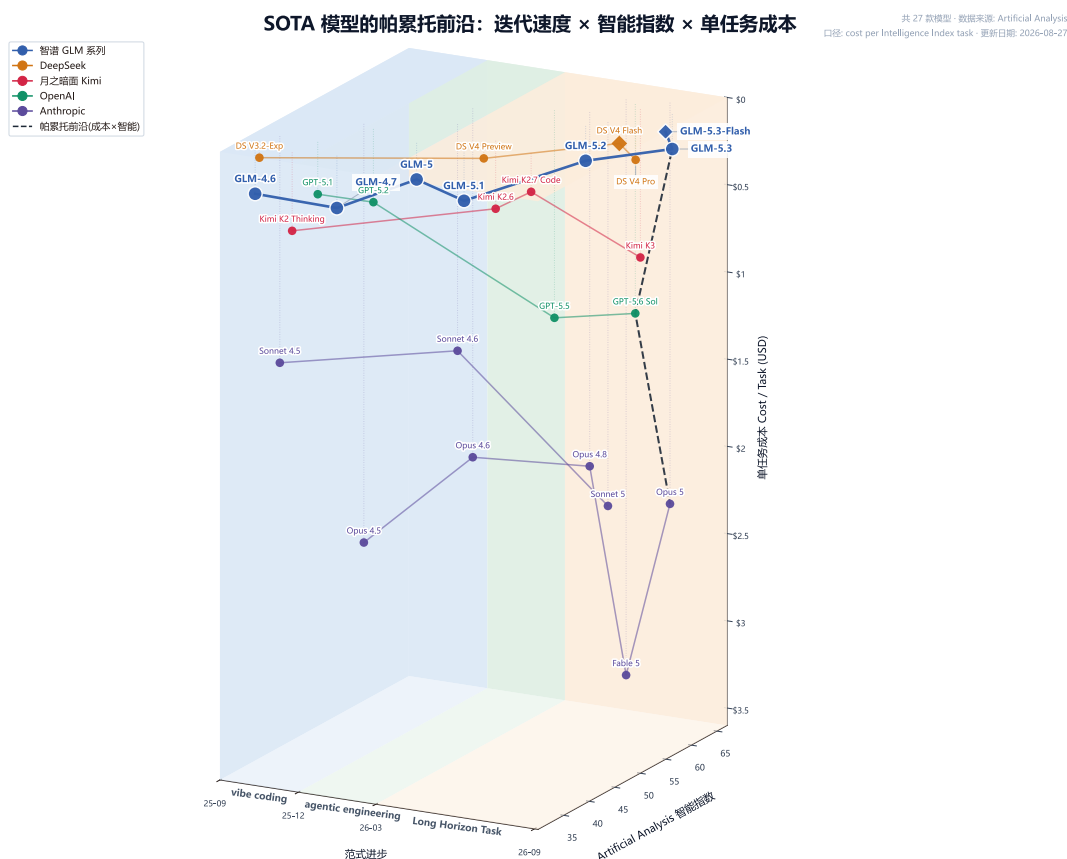
Why this order. We hold a straightforward understanding of AGI: it is not the intelligence of a single person, but the sum total of human intelligence, and should ultimately possess the capability to generate original knowledge on par with the theory of relativity. What determines progress along this path is the generational level of model capabilities. This is the starting point for all of the Company’s products and business logic.

The Company was among the earliest teams in China to invest in large-model R&D. From the GLM pre-training architecture and the Wudao series to today’s GLM-5 series, we have focused essentially on one thing: training more powerful models. When the industry was hot, we did not divert resources to roll out multiple product lines; when the industry cooled, we did not pause model iteration. In retrospect, every sharp increase in revenue occurred after – not before – model capabilities improved and new task scenarios were unlocked.

II. Technology Roadmap and Model Progress

The core capability of a model-building company lies in the efficiency of converting unit resources into intelligence. Compute, data and talent all carry costs; how efficiently the Company can convert these inputs into model intelligence determines how far it can go. Our optimal organizational approach is that research problems originate from real tasks, and research outcomes go directly into the next-generation model, delivering end-to-end service capabilities to users.

- *How We Understand SOTA*



Over the past year, the industry’s understanding of “SOTA” has been rewritten. Leading a point-in-time leaderboard is no longer the most important thing. What truly matters is who can sustainably bring more capable models to market at a faster pace and with better cost-efficiency. We place 27 flagship models from five vendors from September 2025 to the present into a single coordinate system: the horizontal axis is paradigm progress, the vertical axis is intelligence index, and the depth axis is cost per task. In this framework, a single-point champion is merely a fleeting dot; sustained optimality is a trajectory that keeps advancing upward. All players are moving forward, but the pace and cost differ. The GLM series completed six iterations – 4.6, 4.7, 5, 5.1, 5.2 and 5.3 – over approximately 11 months, with the intelligence index rising from 32 to 60 and per-task costs stable at around USD0.20, spanning all three paradigm stages shown in the figure. GLM-5.3-Flash, with a per-task cost of USD0.045, has entered the intelligence-cost frontier. (Intelligence Index and per-task cost source: Artificial Analysis Intelligence Index, as of August 27, 2026; task metric: cost per Intelligence Index task.)

During the Reporting Period, the Company’s technology roadmap comprised three generations of models.

1. GLM-5: Using Coding as the Entry Point to Unlock Long-Horizon Intelligence

We chose coding as the entry point for capability breakthroughs because software engineering is the closest real-world environment to “verifiable intelligence”: task goals are clear, the toolchains are complete, and whether code runs, tests pass and issues are fixed all constitute clear feedback signals. Without verifiable feedback, there are no targets to optimize against.

The capability focus of GLM-5 extends from code generation to system understanding, task planning, tool invocation, continuous debugging and engineering delivery, driving the model from Vibe Coding towards Agentic Engineering.

The value of coding to the Company goes beyond helping developers write code. Over longer task chains, the model needs to continuously understand goals, invoke tools, handle exceptions and adjust strategies based on environmental feedback. Through coding – a high-feedback-density real-world environment – the long-horizon task capabilities accumulated can be transferred to cybersecurity, data analytics, complex automation and other professional domains. During the Reporting Period, the coding direction had already become the Company’s primary revenue source, and the capability transfer mentioned above was validated in Cybersecurity Co-work.

2. GLM-5.2 and GLM-5.3: Scaling Deep While Scaling Up

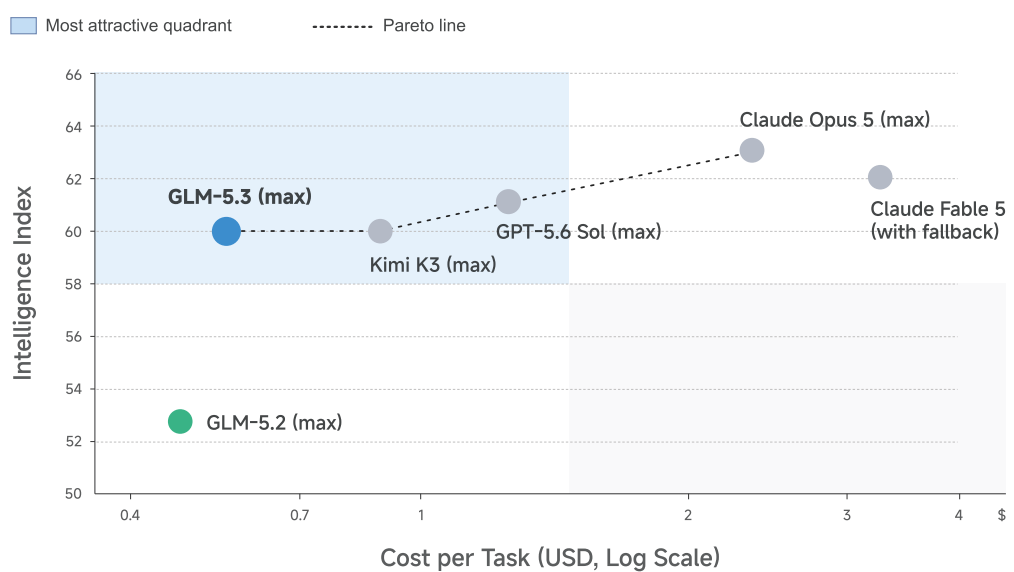
Scaling up means expanding model size, context length, compute power and system capacity; scaling deep means investing in post-training, reinforcement learning, long-horizon trajectories and task feedback. At the current stage, we prioritize resource allocation for the latter. There have been several historical precedents. Kaplan’s Law drove parameter growth far ahead of data growth; Chinchilla revised this path using the “20 Tokens per parameter” ratio: the 70B model trained on 1.4T Tokens comprehensively outperformed the 280B Gopher. In the inference era, models are invoked billions of times per day, inference costs exceed training costs, and the optimum shifts toward “smaller yet trained longer.” MoE further splits parameters into two dimensions: total parameters determine how much is memorized, while active parameters and depth determine how deep the reasoning goes. Recent research also indicates that the optimal Token-to-parameter ratio is task-dependent: memory-oriented tasks favor larger parameters, while reasoning-oriented tasks favor more data. Tasks like cybersecurity do not require memorizing more CVEs; they require pushing a twenty-step causal chain to the end without breaking – a capability that does not reside in total parameters.

CVE (Common Vulnerabilities and Exposures)

MoE (Mixture-of-Experts)

GLM-5.3 is a test of this judgment. It shares the same architecture, total parameter count and active parameter count with GLM-5.2; the only variable is the scale of post-training – using one month to scale up long-horizon task environments and reinforcement learning. In the Company’s self-built real-world coding evaluation, the end-to-end completion rate improved by over 50% compared with GLM-5.2 (evaluation methodology and sample size are detailed in the GLM-5.3 technical blog at <https://z.ai/blog/glm-5.3>). This approximates a controlled experiment, demonstrating that scaling is not just about parameters, and that post-training currently offers the largest headroom for performance gains.

Frontier Models: Intelligence vs. Cost per Task



Source: Artificial Analysis Intelligence Index

Our thesis is that a capable model can only be obtained by applying appropriate data at the right time and training at a suitable scale. The objective is not simply to pursue maximum size, but to deliver the model that, under current data and compute constraints and with sufficient training depth, delivers the highest intelligence upper bound.

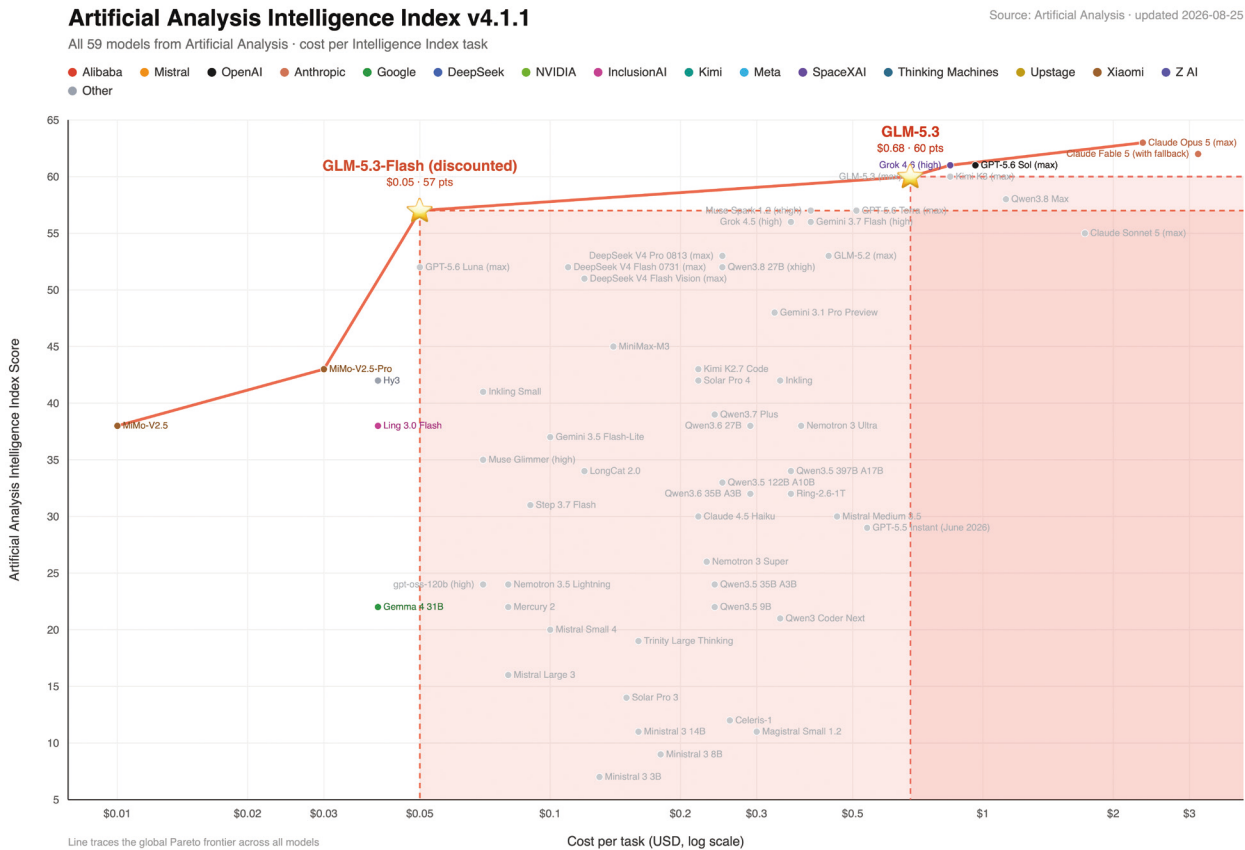
This approach unfolds across four dimensions. (i) Data: continuously increase knowledge density and task complexity, and accumulate high-quality long-horizon Agent trajectories through the developer ecosystem, industry partnerships and Model-with-Model Self-Play, which are then filtered, validated, synthesized and fed back into training; (ii) Mid-training: between pre-training and post-training, the Company has separately allocated compute for the mid-training stage to increase knowledge density, extend effective context length, and establish foundational patterns for tool invocation and long-horizon tasks. The marginal returns of the three stages differ, and compute is budgeted separately rather than as a single one-off investment; (iii) Algorithms: in response to new challenges in reward allocation, feedback utilization and continual learning over long trajectories, we are moving from GRPO to explore Critic-based SAO, and extending from SFT towards OPD. Starting with GLM-5.2, we have introduced Critic-based single-trajectory optimization for ultra-long trajectories; (iv) Engineering and Environment: Slime improves the efficiency of large-scale reinforcement learning training; Index Share reduces the overhead of long-context sparse attention; MTP optimizes inference decoding; and LayerSplit and ZCube improve KV cache management and cluster network efficiency. At the same time, we continue to expand Coding Agent and more real-world interactive environments across additional domains, enabling the model to learn through continuous execution, failure and replanning. As the number, type and complexity of environments grow, the model learns not only how to answer, but how to continuously complete complex work in dynamic environments. These four dimensions interlock: environments generate trajectories; algorithms improve trajectory utilisation; engineering supports larger-scale training and inference; and data sets the starting point.

3. GLM-5.3-Flash (Test Codename Ox-Alpha): Next-Generation Architecture and System Capability

If GLM-5.3 represents the current capability upper bound, GLM-5.3-Flash represents the cost frontier. The two serve different task regimes: the former targets long-horizon tasks requiring the highest intelligence; the latter targets high-frequency, high-volume, cost-sensitive invocations.

GLM-5.3-Flash adopts a brand-new architecture, purpose-built for ultra-low cost. It has 320B total parameters, 18B active parameters and 45 layers. It employs a hybrid architecture of sparse attention and linear attention, which substantially reduces long-context serving costs while maintaining long-context precision, and uses manifold-constrained hyperconnection (mHC) to enhance scaling capability. Combined with a 30T-scale multimodal pre-training corpus, GLM-5.3-Flash outperforms GLM-5.2 on benchmarks and in real-world applications, at one-tenth of the latter’s price.

Before the official release, we deployed the anonymous model Ox-Alpha on OpenCode and OpenRouter for real-world invocation testing. On its first day online, it reached the top of the OpenRouter leaderboard and set a new single-day Token usage record on that platform; within one week, it became the most-invoked model on both platforms, with a 6-day cumulative Token volume of more than 62 trillion, driving an over 20% increase in overall platform traffic.



4. Domestic Chips Power Domestic Models

The inference of GLM-5.3-Flash marks the first time the Company has fully deployed a domestic chip cluster to serve ultra-large-scale real-world traffic, with chips interconnected via a self-developed high-bandwidth network. The way this work was accomplished is itself worth recording: the entire inference engine was substantially accelerated by an infra agent powered by GLM-5.3, which assisted engineers in developing and optimizing operators, diagnosing performance bottlenecks, and improving the deployment service stack. The model optimizes the system, and the system hosts the model – this loop ran internally for the first time. Operator development cycles were shortened by half. This also serves as an internal validation of the Company’s own co-work approach.

The main technical constraint is the memory capacity and bandwidth of domestic chips, making 1M context support particularly challenging. To address this, we built a dedicated inference engine on top of SGLang, with custom optimizations that trade compute for bandwidth and communication for memory, and adopted at the cluster level a production-grade Encode – Prefill – Decode disaggregated architecture, splitting multimodal encoding, prompt pre-filling and per-Token decoding into independently schedulable and scalable worker pools. Compared with the initial baseline on the same hardware, end-to-end service performance tripled; in coding scenarios, the per-Token inference cost on domestic chips has reached a level comparable in cost-performance to mainstream imported GPUs. Domestic chips can economically support inference for cutting-edge models at large scale.

III. Commercialization Progress: The Right Model Will Natively Go to Market

(I) Revenue Structure Flip

The evolution of the Company’s revenue structure is essentially a progression of transaction modes driven by model capability improvement:

Sell Model → Sell Invocations → Sell Subscriptions → Sell End-to-End Task Outcomes

These four transaction modes are not parallel product lines, but forms unlocked sequentially as model capability improves. The more complete the tasks the model can independently accomplish, the closer the customer’s purchase is to the outcome itself, and the unit of revenue measurement changes accordingly.

1. On-premise Deployment Phase: Completing Deep Industry Accumulation and Validation

In the early days, when the model was not yet capable of independently completing a full task, customers purchased tools that needed to be integrated, customized and deployed in their own environments. Combined with requirements for data security, compliance and autonomous control, on-premise deployment became the natural starting point for enterprise adoption of large models. The transaction mode at this phase was selling models – one-time licensing, project-based delivery. During this phase, the Company completed penetration into key industries, developed understanding of complex scenarios, and built enterprise-grade delivery capabilities. These accumulations remain valuable today – what has changed is the delivery form, not the industry understanding.

2. Coding Phase: From One-Time Delivery to Continuous Invocation

The improvement in GLM's coding capability drove rapid growth in API and Coding Plan adoption. During this phase, the Company proactively scaled back its on-premise model software licensing business, shifting toward continuously providing invocable, scalable and quantifiable intelligence services. By end-2025, API invocation volume had grown approximately seven times from end-June 2025, with the revenue from the open platform and API accounting for approximately 26.3% of total revenue, and gross margin improving to 18.9%. Coding became the first scaled entry point for Token consumption and cloud revenue growth. The significance of this step extends beyond growth rate: revenue recognition shifted from one-off to recurring, giving the Company predictable recurring revenue for the first time.

3. Agent and Co-work Phase: From Model Invocation to Task Delivery

In the first half of 2026, GLM's capabilities in Agent planning, tool invocation, continuous execution and long-horizon tasks continued to improve, extending the business from the single-point application in coding to high-value scenarios such as software engineering, cybersecurity, data analytics and complex automation. The transaction mode evolved further from selling invocations to selling subscriptions and end-to-end task outcomes.

During the Reporting Period, revenue from the open platform and API reached RMB825 million, representing a year-on-year increase of over 27 times and accounting for approximately 86.5% of total revenue, compared with 15.2% in the same period last year – a rise of 71.3 percentage points.

Co-work represents the part currently being unlocked: the model connects to enterprise data, tools and business systems, embedding into procurement, finance, customer service, R&D, IT operation and maintenance and other functions, moving from understanding requirements to executing tasks. In turn, enterprises' real workflows become learnable environments and data for the model. During the Reporting Period, the Company initiated pilot efforts in cybersecurity, legal, finance, education and other sectors, with cybersecurity progressing most rapidly; the others remain at early stages and have not yet generated material revenue.

Looking back at these three phases, they form a cohesive trajectory: with each leap in model capabilities, what customers purchase moves one step closer to tangible outcomes, and the Company's revenue structure is reshaped accordingly.

(II) Volumes Surge While Prices Rise, with Improving Gross Profit and Narrowing Loss

The above growth did not come at the expense of price. As of the date of this announcement, the Token invocation volume of the MaaS platform had increased over 40 times from the beginning of 2026, with Coding Plan invocation volume growing over 23 times. During the same period, the average API selling price increased by approximately 101%, and Coding Plan subscription prices also went up. Volumes surged while prices rose.

During the Reporting Period, the gross profit margin for the open platform and API business increased to 24.6%, an improvement of approximately 25 percentage points year-on-year and up approximately 6 percentage points from 2025. The Company's loss for the period decreased by 12.1% year-on-year, the loss ratio (loss/revenue) **narrowed by 4.7 times compared with the same period last year**, and the adjusted net loss ratio for the period (adjusted loss/revenue for the period) **shrunk by 3.5 times compared with the same period last year**. While the scale expanded, unit economics improved, which is the set of metrics we value more.

(III) From Coding to Co-work

Coding is currently the largest-consumption and most fully validated scenario. Leveraging code generation, engineering understanding, tool invocation and long-horizon task capabilities, the Company is extending from code assistance to the full software development lifecycle. Through Harness, ZCode undertakes the tasks of context management, tool invocation, task scheduling, caching and result verification, improving the model's performance in real-world engineering. As long-horizon capabilities advance, the complexity ceiling of coding tasks continues to rise.

The Company uses three dimensions to decide which co-work directions to enter: intellectual threshold – the requirements for reasoning, planning, tool invocation and long-horizon execution; unit intelligence consumption – the effective Token, inference duration, computing power and human intervention required to complete a task; and unit output value – the revenue generated by a single delivery, labor time saved, and risk reduced. Together, these three factors determine whether it is necessary to handle over a task to the model today.

For large enterprises and public institutions, the Company continues to provide on-premise deployment and organization-level intelligent infrastructure to satisfy the demands for data security, permission management, system integration and domestic computing power adaptation. Such demands will not disappear, but their roles in revenue have shifted from primary contributor to supplementary contributor.

(IV) User Scale and Engagement Depth

As of the date of this announcement, the Company's MaaS platform had over 7.4 million users, an increase of 144% from the beginning of 2026, with the number of paid daily active users rising by 603%. For the top ten customers by revenue, the average daily invocation volume soared 98 times from the beginning of 2026.

These three figures point to the same direction, with increasing magnitudes: more users, faster growth in paying users, and deeper usage among heavy paying users. The third figure is especially important as it indicates that the model is moving from trial to production within customer organizations.

(V) Compute Multiplier

During the Reporting Period, the Company's compute multiplier, defined as the open platform and API revenue generated from every RMB1 of computing power investment (including both training and inference), increased by approximately 14 times year-on-year.

This improvement in the curve comes from both ends: the intelligence per Token is increasing, while the Token and cost consumed per task are decreasing. The former is determined by model capability; the latter by infrastructure. Only when both of them improve simultaneously can there be a Pareto advance in the "intelligence-cost" frontier. The revenue recovered per RMB1 of computing power investment is increasing, meaning the model's self-sustaining capability is strengthening. We intend to continue pushing this curve upward.

IV. Ecosystem and Safety: Earning Real-World Trust for Capability

Between model capability and a dependable service lies one thing: whether the real world trusts it. The Company works on two fronts simultaneously: open-source and openness, so that safety capabilities grow in tandem with model capabilities. On one hand, the Company makes model weights, the MaaS platform, Coding Plan and the global developer network available for users, allowing the model to enter as many real-world scenarios as possible. On the other hand, the Company establishes verifiability of capabilities through permission controls, process supervision, risk assessment, vulnerability disclosure and external validation. They are mutually reinforcing: without openness and open-source, safety cannot be examined; without safety, openness and open-source cannot be sustained.

Safety is not merely a compliance cost; it can also be a verifiable public good, helping the market judge whether a model can be safely, stably and durably deployed in real-world environments.

From Capability to Public Good

Since the launch of GLM-5.2, the Company, in collaboration with multiple security teams, has conducted ongoing red-team testing and safety evaluations. As of the end of the Reporting Period, a total of 2,436 vulnerabilities had been identified (after initial screening and deduplication), of which 1,097 were medium- or high-severity, covering 269 projects including system kernels, operating systems, browser engines, open-source foundational components, Internet applications and network protocols. A considerable portion of the vulnerabilities are long-standing, previously undiscovered historical vulnerabilities, with the oldest having existed for more than 20 years.

These discoveries have evolved into an ongoing vulnerability discovery and disclosure effort, with related vulnerabilities being submitted to national vulnerability databases such as CNNVD/CNVD. To enhance transparency, the Company has established a Security Disclosure Ledger (<https://cvd.z.ai/>) that records the entire process from discovery, confirmation to remediation. The ledger is continuously updated, which only discloses the vulnerabilities that have been disclosed and remediated by vendors; for those still in the coordinated disclosure phase, only hash values are made public for subsequent verification, without disclosing reproducible exploit code.

The significance of this mechanism is that it places the model's capabilities in a position subject to third-party verification. Benchmark scores are reported by us; vulnerability disclosures are confirmed by national vulnerability databases and vendors, with the latter being the source of external trust.

Open Source and Openness

With the launch of GLM-5.3, the Company simultaneously launched the “Open Source Shield” initiative, comprising three ongoing commitments. First, we conduct continuous security audits of key open-source projects, helping maintainers identify and remediate risks free of charge, with audit results made public through OpenVuln (<https://huggingface.co/spaces/zai-org/OpenVuln>). Second, we provide free model quotas and defence access for the open-source community, allowing open-source project maintainers to apply for these resources for security audit and defence tasks; Third, we offer code audit functionality in ZCode, integrating security checks into daily R&D workflows.

The common thread across these three initiatives is putting the Company’s most advanced capabilities into the hands of those who need them most but can least afford them. Open-source foundational components underpin almost all modern software, while maintainers are often a few unpaid individuals. While attackers’ cost of using cutting-edge models is falling rapidly, defenders may not be able to afford them. If the strongest spear is locked in the hands of a few, the best shield must belong to everyone. Before open-sourcing, the Company, together with top domestic security teams, has conducted sustained red-team testing and automated jailbreak evaluations, using newly generated samples to iteratively harden the review system, and invited professional teams to independently assess the model’s high-risk task capabilities. The balance between capability release and risk mitigation is judged by external parties, not by the Company itself.

From a commercial perspective, open-source and safety are not cost items. The distribution scale brought by open-source is a prerequisite for the model to receive real-world task feedback; the external confirmation brought by security disclosures is the basis upon which enterprise customers decide whether to integrate the model into their production systems.

Capability Boundaries and Controlled Access

The model’s capabilities in cybersecurity have dual-use implications. For the most sensitive capabilities, the Company adopts a controlled-access approach: which are made available to verified security researchers and institutions, while all other users follow standard security policies. Before open-sourcing the model, the Company completes safety assessments, model hardening and usage boundary design.

Future Outlook

In the first half of 2026, we have validated that the model can independently complete an entire task. The next step is to enable it to undertake an entire business process.

(I) Tasks: From Long-Horizon Tasks to Autonomous Systems

Longer time horizons. We will enable the model to stay consistent with its goals over longer periods and continuously perform task decomposition, tool invocation, environmental interaction, error remediation and result verification. Today's frontier is engineering tasks on the order of hours, while the next step will be complete delivery on the order of days.

Multi-agent collaboration. We will move from a single Agent to multi-Agent collaboration, with the model undertaking complete business processes, shifting from assisting work to executing work. The threshold at this step lies not entirely in intelligence itself, but is more about in reliability and auditability, indicating the model must be able to operate without step-by-step human inspection, while leaving verifiable traces.

Each outward expansion of the task boundary raises the requirements on model capability by an order of magnitude. The consequent question is: along which dimensions can capability continue to improve.

(II) Capability: Scaling Never Stops

Scaling remains the main path to higher intelligence. What has changed is what we are scaling.

Over the past few years, scaling was almost synonymous with parameter count. However, once parameters pass the threshold of “housing the world”, the marginal return from further width diminishes rapidly, while other dimensions still have substantial untapped headroom. What determines a model's upper bound today is a combination of four factors: base model scale, effective depth, training depth and task environments. Their marginal returns differ and rotate across stages; the Company's resource allocation is about judging which factor offers the highest immediate return at each stage.

Base Model Scale: The upper bound is set by pre-training. Post-training is the knob with the highest immediate return, but the ceiling on intelligence is determined by the base model. How deep a model can be trained depends on how much of the world it has seen during pre-training and how many structures remain available for activation. After post-training is pushed to its extreme, the next phase of growth must return to the base model. The Company has never ceased base-model iteration. GLM-5.3-Flash has already used a new-generation architecture and a 30T-scale multimodal pre-training corpus, validating the new architecture’s efficiency at medium parameter scale. Building on this, the Company is advancing the training of the next-generation base model, targeting larger effective scale, longer native context and unified native multimodal modelling.

Effective Depth: Let the model think several steps ahead. Parameters determine how much is remembered; depth determines how far it can think. Loop Transformer increases effective computational depth through layer recycling without significantly increasing parameter count, shifting compute from “remembering” to “thinking” (the Company has preliminarily validated this concept in early experiments). This path aligns with our judgment on tasks: cybersecurity-type work requires pushing a twenty-step causal chain to the end without breaking, a capability that comes from depth, not parameter count.

Training Depth: Currently the largest untapped margin. GLM-5.3 is direct evidence for this judgment. It shares the same architecture, total parameters and active parameters as GLM-5.2, with the only variable being post-training scale. It achieved an end-to-end completion rate improvement of over 50%. This curve has not yet plateaued. The Company will continue to scale up reinforcement learning over long-horizon tasks, extend effective training trajectories and improve trajectory utilization efficiency.

Task Environments: A newly emerging variable. When algorithms and computing power are no longer the sole bottleneck, the number, type and complexity of environments begin to directly determine what the model can learn. This is a dimension that has emerged over the past two years and is among the most concentrated areas of the Company’s next-stage investment.

(III) Systems: Fully Self-training

Among the four knobs, the speed of environment building is becoming the new bottleneck. And at a deeper level, early evidence is emerging that AI is taking over all three components of the training system: data, environments and infrastructure.

Data is starting to be self-generated. High-quality human-generated data is peaking, with publicly available data nearly exhausted, while next-generation base model requires trillions of Tokens of ever-improving quality. Our solution is a model-vs-model self-play pipeline: one model generates solutions and content, while another model reviews, tests and executes validation, with only data that passes adversarial checks moving to the next stage. This is followed by four layers of filtering – rule-based verification confirms compilability, reasoning consistency and format compliance; execution validation actually runs in a sandbox, where execution results themselves are the highest-quality training signal; model review adds another judgment layer; and finally manual spot checks are conducted for calibration. What is accumulated is verifiable, executable and debuggable data that feeds back into all three stages of pre-training, mid-training and post-training.

Thus, the upper bound on data quality is no longer determined by human annotation speed. Instead, it is determined by the model’s own verification capability. The stronger the model, the harder the problems it generates and the stricter its validation, leading to higher-quality data; better data then trains a stronger next model. Data transforms from a one-time consumable into a reusable capital asset that spans generations.

Environments are starting to be self-built. A useful training environment must be executable, verifiable and close to real-world professional work, and there must be thousands of environments, which cannot be built through manual work. In the GLM-5.3 technical blog, we shared some of our approaches: research agents capture task patterns from real-world work, converting work that would take senior engineers several days into runnable long-horizon environments containing multi-step dependencies and hidden states; judge agents test-run each task to confirm solvability; verifiers are synthesized without access to ground-truth answers, and solver trajectories are then used to detect and seal reward shortcuts; verifiers that pass oracle, no-op and unsolved checks ultimately produce binary rewards directly usable for reinforcement learning. The Slime architecture reduces mathematics, code, sandboxes and verifiers to plug-and-play data components, turning environments into standard products on an assembly line.

It should be noted that this pipeline still involves a considerable amount of human intervention. Our next step is to make environment building and validation more autonomous.

Infrastructure is starting to self-optimize. Inference system optimization, including kernels, scheduling, quantization and operators, is essentially a code-engineering task, which happens to be the model’s strongest domain. The inference engine for GLM-5.3-Flash was built through the following approach: an infrastructure agent powered by GLM-5.3 assisted engineers in developing and optimizing kernels, diagnosing performance bottlenecks and improving the service stack, ultimately achieving a 300% improvement in end-to-end service performance on domestic chips. The same logic is also operating on the training side: Slime enables training and large-scale inference rollout to share a common data flow, and configuration experience and scheduling strategies accumulated in reinforcement learning can be directly transferred to production serving.

These three threads point to the same end state: the model participates in its own improvement, forming recursive self-improvement. This is not a distant concept and its early form has already emerged in the first half of 2026: GLM-5.3 contributed to the optimization of the inference engine serving GLM itself, and environments generated by GLM are training the next-generation GLM. The only difference today is that a large number of engineers still stand in this loop. Each time the loop tightens, the pace of intelligence progress is increasingly determined by the speed at which environments grow.

This is our technical vision: Fully Self Training. The next-generation GLM will be trained in the environments built by the previous-generation GLM.

(IV) Boundaries and Governance

The technical aspiration of the Company is that the model can automatically participate in the building of the training system. However, we must ensure that the assessment of its capabilities and the judgment of the risks ultimately remain human-mediated, with external institutions invited to perform evaluation independently. In the first half of 2026, we have already done so in cybersecurity: the most sensitive capabilities are subject to controlled access, open-sourcing is preceded by independent evaluation by third-party professional teams, and vulnerabilities are disclosed through national vulnerability databases. Benchmark scores are reported by us; risk judgments are subject to external confirmation. This division of labor will not change as capabilities improve; it will only be reinforced as capabilities ascend.

The Company will continue to seek new balances among capability ceiling, inference efficiency, task value density and safety governance, so that higher-level intelligence can enter the real world stably and controllably, and everyone can benefit from universal access to intelligence.

Financial Review

The following discussion is based on the financial data and accompanying notes set out in other sections of this interim results announcement and should be read in conjunction therewith.

ANALYSIS OF KEY OPERATING PERFORMANCE ITEMS

Revenue

During the Reporting Period, we achieved substantial revenue growth. Our revenues were RMB190.9 million for the six months ended June 30, 2025 and RMB953.9 million for the six months ended June 30, 2026, respectively, representing an increase of 399.7%. The expansion in revenue scale was primarily driven by the explosive growth of the cloud-based deployment business, marking a shift in the Group's primary revenue growth driver from on-premise deployment business to cloud-based deployment business.

By deployment method:

Cloud-based Deployment

Revenue from cloud-based deployment increased from RMB29.1 million for the six months ended June 30, 2025 to RMB825.2 million for the six months ended June 30, 2026, representing an increase of 2,735.7%. This growth primarily reflected the Company's gradual enhancement of its model capabilities for high-value task scenarios, and the continuous progress in the model's long-horizon task capabilities, enabling users to consistently leverage the model in complex scenarios, thereby driving a significant increase in average daily Token usage per user. Meanwhile, the number of users on the MaaS platform continued to expand, and the momentum of simultaneous growth in both volume and price persisted. Revenue from cloud-based deployment as a percentage of total revenue increased from 15.2% for the corresponding period of the previous year to 86.5%, and has become the Group's primary source of revenue.

On-premise Deployment

Revenue from on-premise deployment decreased from RMB161.8 million for the six months ended June 30, 2025 to RMB128.7 million for the six months ended June 30, 2026, representing a decrease of 20.5%. This decrease was primarily due to the Group's greater emphasis on the quality of its on-premise business and its selection of industries amid the explosive growth of its cloud-based deployment business, thereby laying a solid foundation for the future development of its model capabilities from general-purpose scenarios to high-value, specialised task scenarios in vertical industries.

	For the six months ended June 30,			
	2026		2025	
	<i>RMB in</i>	<i>% of total</i>	<i>RMB in</i>	<i>% of total</i>
	<i>thousands</i>		<i>thousands</i>	
	(Unaudited)		(Audited)	
Cloud-based deployment	825,176	86.5%	29,100	15.2%
On-premise deployment	128,716	13.5%	161,777	84.8%
Total	<u>953,892</u>	<u>100.0%</u>	<u>190,877</u>	<u>100.0%</u>

By business form and core product lines:

Open Platform and API

Revenue from the Open Platform and API, or cloud services provided by the Group based on general-purpose large models, increased from RMB29.1 million for the six months ended June 30, 2025 to RMB825.2 million for the six months ended June 30, 2026, representing an increase of 2,735.7%. This growth primarily reflected the Company's gradual enhancement of its model capabilities for high-value task scenarios, as well as the continuous improvement and qualitative advancement in long-horizon task capabilities driven by multiple rounds of iteration of the GLM series of models. These improvements enable users to continuously invoke the models in complex scenarios, thereby leading to a significant increase in average daily Token usage per user. Meanwhile, the user base of the MaaS platform continued to expand, and the momentum of simultaneous growth in both volume and price persisted.

Enterprise-level Agents

Revenue from enterprise-level agents, or autonomous intelligent systems developed by the Group for complex corporate scenarios, increased from RMB13.7 million for the six months ended June 30, 2025 to RMB55.6 million for the six months ended June 30, 2026, marking an increase of 304.4%. This was attributable to the accelerated adoption of intelligent agents by enterprise customers. On the one hand, the execution capabilities of the GLM series models in long-horizon tasks and complex scenarios continued to improve, enabling enterprise workflow-oriented agent products to cover more high-value scenarios; on the other hand, during the Reporting Period, enterprise customers demonstrated rapidly growing demand for the deployment of agent-based use cases, thereby contributing to the revenue growth of this segment.

Enterprise-level General-purpose Large Models

Revenue from enterprise-level general-purpose large models, namely a matrix of pre-trained models independently developed by the Group primarily delivered through private on-premise deployment, decreased from RMB147.6 million for the six months ended June 30, 2025 to RMB67.0 million for the six months ended June 30, 2026, representing a decrease of 54.6%. The decrease was primarily attributable to the Group's greater emphasis on the quality of its enterprise-level general-purpose large language model business and its selection of industries, thereby laying a solid foundation for the future development of its model capabilities from general-purpose scenarios to high-value, specialised task scenarios in vertical industries. Meanwhile, some enterprise-level general-purpose large language model customers gradually migrated to Open Platform and API services, which created a positive flywheel effect.

Technical Services and Others

Revenue from technical services and others increased from RMB0.5 million for the six months ended June 30, 2025 to RMB6.1 million for the six months ended June 30, 2026, representing a period-on-period increase of 1,120.0%, primarily due to the Group's exploration of extending the capabilities of its self-developed large models to more application scenarios. The revenue base of this segment remained relatively small, and its changes had a limited impact on the Group's overall revenue structure.

	For the six months ended June 30,			
	2026		2025	
	<i>RMB'000</i>	<i>% of Total</i>	<i>RMB'000</i>	<i>% of Total</i>
	(Unaudited)		(Unaudited)	
Open platform and API	825,176	86.5%	29,100	15.2%
Enterprise-level agents	55,561	5.8%	13,739	7.2%
Enterprise-level general-purpose large models	67,037	7.0%	147,555	77.3%
Technical services and others	6,118	0.7%	483	0.3%
Total	953,892	100.0%	190,877	100.0%

Gross Profit and Gross Profit Margin

The Group's gross profit increased from RMB95.4 million for the six months ended June 30, 2025 to RMB251.6 million for the six months ended June 30, 2026, representing an increase of 163.7%. The Group's gross profit margin decreased from 50.0% for the six months ended June 30, 2025 to 26.4% for the six months ended June 30, 2026, primarily due to the change in revenue mix, as the proportion of cloud-based deployment business expanded rapidly. This

business segment, being in its early stage of rapid growth, was still experiencing a gross profit margin ramp-up, which structurally diluted the overall gross profit margin. Nevertheless, the gross profit margin of the cloud-based business itself turned from negative to positive and increased rapidly from the corresponding period of the previous year. During the Reporting Period, the Group's gross profit increased by more than 1.6 times, reflecting that the change in revenue mix has made a positive contribution to gross profit.

By deployment method:

Cloud-based Deployment

Gross profit from the cloud-based deployment business increased from RMB-0.1 million for the six months ended June 30, 2025 to RMB202.9 million for the six months ended June 30, 2026, representing a turnaround over the prior period. The gross profit margin increased from -0.4% for the six months ended June 30, 2025 to 24.6% for the six months ended June 30, 2026. This improvement was primarily attributable to: first, the economies of scale resulting from the explosive growth in revenue, with the exponential increase in Token usage significantly diluting fixed costs; second, the gradual upward adjustment of pricing strategies in line with the increase in the upper bound of model intelligence, while Token usage continued to grow, maintaining the momentum of simultaneous growth in both volume and price; and third, continuous full-stack self-developed engineering optimization that reduced model inference costs.

On-premise Deployment

Gross profit from the on-premise deployment business decreased from RMB95.5 million for the six months ended June 30, 2025 to RMB48.7 million for the six months ended June 30, 2026, representing a period-on-period decrease of 49.0%. The gross profit margin decreased from 59.1% for the six months ended June 30, 2025 to 37.9% for the six months ended June 30, 2026. Gross profit and gross profit margin both decreased, primarily due to the Group's efforts to advance the adjustment of its business model, resulting in changes in the product mix of its on-premise deployment business. The scale of private delivery of enterprise-level general-purpose large models with relatively high gross profit margin has contracted, while the proportion of enterprise-level agent products, which remained at the stage of large-scale expansion, has increased. Meanwhile, certain customers migrated from on-premises deployment to cloud services.

	For the six months ended June 30,			
	2026		2025	
	Gross profit (RMB'000) (Unaudited)	Gross profit margin (%)	Gross profit (RMB'000) (Audited)	Gross profit margin (%)
Cloud-based deployment	202,871	24.6	(116)	-0.4
On-premise deployment	48,740	37.9	95,540	59.1
Total	251,611	26.4	95,424	50.0

By business form and core product lines:

Open Platform and API

Gross profit from the Open Platform and API business increased from RMB-0.1 million for the six months ended June 30, 2025 to RMB202.9 million for the six months ended June 30, 2026, representing a turnaround over the prior period. The gross profit margin improved from -0.4% for the six months ended June 30, 2025 to 24.6% for the six months ended June 30, 2026. This improvement was primarily attributable to the realization of economies of scale, upward adjustments to pricing, and continued full-stack engineering optimization that reduced inference costs.

Enterprise-level Agents

Gross profit from the enterprise-level agents business increased from RMB8.9 million for the six months ended June 30, 2025 to RMB19.8 million for the six months ended June 30, 2026, representing a period-on-period increase of 122.5%. The gross profit margin decreased from 64.6% for the six months ended June 30, 2025 to 35.6% for the six months ended June 30, 2026. Gross profit increased while gross profit margin decreased, primarily due to increasingly growing demands from enterprise customers for the deployment of agent-based use cases during the Reporting Period. To ensure delivery quality and the depth of scenario coverage, relatively significant development and delivery costs were made to meet diversified customer needs; as economies of scale gradually emerge, the gross profit margin is expected to stabilize and recover.

Enterprise-level General-purpose Large Models

Gross profit from the enterprise-level general-purpose large models business decreased from RMB86.3 million for the six months ended June 30, 2025 to RMB28.4 million for the six months ended June 30, 2026, representing a period-on-period decrease of 67.1%. The gross profit margin decreased from 58.5% for the six months ended June 30, 2025 to 42.3% for the six months ended June 30, 2026. Gross profit and gross profit margin both decreased, primarily due to the Group's greater emphasis on the quality of this type of business and its selection of industries, thereby laying a solid foundation for the future development of its model capabilities from general-purpose scenarios to high-value, specialised task scenarios in vertical industries. Meanwhile, the delivery and maintenance costs of this type of business have a certain degree of rigidity, resulting in reduced economies of scale. In addition, as customer needs have been upgraded, certain projects have shifted towards enterprise-level agent solutions, which also had a temporary impact on gross profit margin.

Technical Services and Others

Gross profit from the technical services and others business increased from RMB0.3 million for the six months ended June 30, 2025 to RMB0.6 million for the six months ended June 30, 2026, representing a period-on-period increase of 100.0%. The gross profit margin decreased from 65.8% for the six months ended June 30, 2025 to 9.4% for the six months ended June 30, 2026. Gross profit increased while gross profit margin decreased, primarily due to increased investment in certain exploratory scenario projects during the Reporting Period, and the relatively small revenue base of this segment, which resulted in a more significant impact on gross profit margin.

	For the six months ended June 30,			
	2026		2025	
	Gross profit (RMB'000) (Unaudited)	Gross profit margin (%)	Gross profit (RMB'000) (Unaudited)	Gross profit margin (%)
Open platform and API	202,871	24.6	-116	-0.4
Enterprise-level agents	19,796	35.6	8,880	64.6
Enterprise-level general-purpose large models	28,368	42.3	86,342	58.5
Technical services and others	576	9.4	318	65.8
Total	251,611	26.4	95,424	50.0

Cost of Sales

The Group's cost of sales increased from RMB95.5 million for the six months ended June 30, 2025 to RMB702.3 million for the six months ended June 30, 2026, representing a period-on-period increase of 635.4%. The increase in costs was primarily attributable to the increase in computing service fees resulting from business expansion and revenue growth, which was broadly in line with the trend of revenue growth.

Capital Expenditure

The Group's capital expenditure increased from RMB53.7 million for the six months ended June 30, 2025 to RMB390.6 million for the six months ended June 30, 2026, representing a period-on-period increase of 627.4%. This increase was primarily due to the Company's completion of the acquisition of the entire equity interest in Beijing Hongzuan Technology Development Co., Ltd. (北京紅鑽科技發展有限公司) ("**Beijing Hongzuan**"), resulting in the addition of a significant amount of property, plant and equipment.

Other Income

The Group's other income increased from RMB4.6 million for the six months ended June 30, 2025 to RMB41.4 million for the six months ended June 30, 2026, representing a period-on-period increase of 800.0%. This was primarily due to the substantial increase in the scale of the Group's bank deposits following the Company's listing in January 2026, which resulted in a corresponding increase in interest income.

Research and Development Expenses

The Group's research and development expenses increased from RMB1,594.7 million for the six months ended June 30, 2025 to RMB2,131.2 million for the six months ended June 30, 2026, representing a period-on-period increase of 33.6%. This was primarily due to the Group's continued increase in research and development investment in model training to support the continuous breakthroughs in the upper bound of intelligence of the GLM series models.

Selling and Marketing Expenses

The Group's Selling and marketing expenses decreased from RMB208.6 million for the six months ended June 30, 2025 to RMB177.7 million for the six months ended June 30, 2026, representing a period-on-period decrease of 14.8%. This was primarily due to the Group's optimization of its allocation of marketing resources and improvement in the conversion efficiency of marketing investments.

General and Administrative Expenses

The Group's general and administrative expenses decreased from RMB185.2 million for the six months ended June 30, 2025 to RMB103.3 million for the six months ended June 30, 2026, representing a period-on-period decrease of 44.2%. This was primarily due to the decrease in professional service fees and equity-settled share-based compensation expenses during the reporting period, compared to the same period last year.

Impairment Losses on Financial Assets

The Group's impairment losses on financial assets increased from RMB10.9 million for the six months ended June 30, 2025 to RMB27.4 million for the six months ended June 30, 2026, representing a period-on-period increase of 151.4%. This was primarily due to the increase in the gross amount of trade and bills receivables resulting from the expansion of the business scale.

Finance Costs

The Group's finance costs increased from RMB53.3 million for the six months ended June 30, 2025 to RMB95.9 million for the six months ended June 30, 2026, representing a period-on-period increase of 79.9%. This was primarily due to the recognition of net foreign exchange losses and the increase in interest expenses on bank loans during the Reporting period.

Share of Profits Less Losses of Associates

The Group's share of profits less losses of associates increased from RMB14.1 million for the six months ended June 30, 2025 to RMB144.2 million for the six months ended June 30, 2026, representing a period-on-period increase of 922.7%. This was primarily due to the improvement in the operating and financial results of the associates during the Reporting Period.

Changes in Fair Value of Financial Instruments Measured at FVPL

The changes in fair value of financial instruments measured at fair value through profit or loss increased from RMB9.8 million for the six months ended June 30, 2025 to RMB50.4 million for the six months ended June 30, 2026, representing a period-on-period increase of 414.3%. This was primarily due to the increase in the valuation of the financial instruments held by the Group during the period, resulting in the recognition of fair value gains.

Changes in the Carrying Amount of Financial Instruments Issued to Investors

The carrying amount of financial instruments issued to investors resulted in a loss of RMB429.3 million for the six months ended June 30, 2025, which decreased to a loss of RMB22.1 million for the six months ended June 30, 2026, representing a period-on-period decrease of 94.9%. This was primarily due to the termination of financial instruments issued to investors and related redemption rights and their derecognition upon listing in January 2026, resulting in changes in carrying amount being recognized only for the short period prior to the listing during the Reporting Period.

Net Loss

The Group recorded a net loss of RMB2,072.0 million for the six months ended June 30, 2026, representing a decrease from a net loss of RMB2,357.9 million for the six months ended June 30, 2025. This was primarily due to an increase in gross profit during the reporting period and a significant decrease in losses arising from changes in the carrying amount of financial instruments issued to investors upon listing, which were converted into ordinary shares. The increase in research and development expenses partially offset the decrease.

Non-IFRS Measure

	For the six months ended June 30,	
	2026	2025
	RMB'000	RMB'000
	(Unaudited)	(Unaudited)
Loss for the Period	(2,071,992)	(2,357,852)
Add:		
– Equity-settled share-based compensation expenses	85,764	158,852
– Changes in the carrying amount of financial instruments issued to investors	22,090	429,295
– Listing expense	–	17,731
Adjusted net loss for the period (non-IFRS measure)	(1,964,138)	(1,751,974)

ANALYSIS OF KEY ITEMS IN FINANCIAL POSITION

Financial Position

Shareholders' equity increased from RMB-8,111.0 million as at December 31, 2025 to RMB4,429.6 million as at June 30, 2026, primarily due to the successful listing of the Group during the Reporting Period, pursuant to which the financial instruments issued to investors and related redemption rights had been terminated, and the related financial liabilities were transferred to the Company's equity.

Liquidity and Financial Resources

As at June 30, 2026, the Group held cash and cash equivalents of RMB3,993.7 million, an increase of RMB1,734.6 million from RMB2,259.1 million as at December 31, 2025. This fluctuation was mainly attributable to the receipt of proceeds from the listing and an increase in bank loans during the Reporting Period, while operating expenses consumed part of the funds.

Debt/Borrowings

As at June 30, 2026, the Group's total bank loans amounted to RMB2,224.8 million, an increase of RMB1,535.3 million from RMB689.5 million as at December 31, 2025, primarily due to funds borrowed to meet operational needs. All relevant loans are interest-bearing, denominated primarily in RMB and unsecured.

Lease Liabilities

As at June 30, 2026, the total lease liabilities recognized by the Group (including both current and non-current portions) amounted to RMB379.4 million, a decrease of RMB165.7 million from RMB545.1 million as at December 31, 2025. This fluctuation was primarily attributable to the payment of lease payments during the Reporting Period and changes in lease arrangements.

Gearing Ratio

The Group's gearing ratio (Gearing ratio is calculated by bank loans and lease liabilities divided by total equity as of the end of the period multiplied by 100%) was 58.8% as at June 30, 2026, compared to -15.2% as at December 31, 2025. This change was primarily attributable to the termination of the financial instruments issued to investors and related redemption rights upon listing, and the related financial liabilities were transferred to the Company's equity resulting in total equity changing from negative to positive.

Asset Pledges

The Group had no pledged assets for the six months ended June 30, 2026.

Capital Commitments

The Group had no outstanding capital commitment for the six months ended June 30, 2026.

Contingent Liabilities

The Group had no significant contingent liabilities for the six months ended June 30, 2026.

Significant Investments and Major Acquisitions and Disposals of Subsidiaries, Associates, and Joint Ventures

Pursuant to the resolution of the Board dated March 31, 2026 and the acquisition framework agreement dated April 14, 2026, the Company acquired 100% of the equity interest of Beijing Hongzuan at a total consideration of no more than RMB360.52 million. Upon completion of the acquisition, Beijing Hongzuan has been accounted for as a subsidiary of the Company. For details, please refer to the Company's announcement dated April 14, 2026.

Save as disclosed above, as of June 30, 2026, the Group did not make any significant investments or engage in any other major acquisitions or disposals of subsidiaries, associates, or joint ventures. The Group had no specific future plans for the acquisition of material capital assets as of the date of this announcement.

Employees and Remuneration Policy

As at June 30, 2026, the Group had 981 permanent employees (June 30, 2025: 883), with staffing levels subject to operational needs. The Group incurred total remuneration costs (including share-based payments) of RMB482.1 million during the Reporting Period (six months ended June 30, 2025: RMB499.6 million).

Talent is the Company's most core strategic resource. In light of the characteristics of the artificial intelligence industry, we have established a talent management system that is capability-oriented and incentive-driven, supporting the Company's continuous technological innovation and business expansion in the AI field.

In terms of remuneration strategy, we adopt a three-tier structure of “fixed remuneration + performance bonus + medium- and long-term incentives” to ensure that rewards are closely aligned with the Company’s performance and individual contributions. Through employee incentive plans and employee shareholding platforms, we closely align the interests of core management personnel, key technical personnel and business professionals with the Company’s long-term development. We regularly conduct industry remuneration benchmarking to maintain the market competitiveness of our remuneration levels and attract and retain top talent.

In terms of employee care, in addition to statutory social insurance and housing provident fund, we provide multiple layers of benefits, including supplementary commercial insurance, annual health check-ups, meal allowances and holiday benefits. We also support various cultural, recreational and sports activities, pay attention to employees’ physical and mental well-being, and foster a positive and uplifting working environment. The Company has established systematic internal training and knowledge-sharing mechanisms, encouraging cross-departmental collaboration and technical exchanges and providing ongoing support for employees’ career development.

As a leading enterprise in the large language model sector, we have established an AI-native working model as the fundamental mode of organizational operations. Built on our self-developed models and intelligent agent systems, we have positioned ourselves as a frontier environment for the large-scale validation of model capabilities in real-world production environments. In our daily R&D activities, standardized processes such as code generation, defect fixing and test writing are largely completed autonomously by AI, while employees focus on identifying the essence of problems, designing technical solutions and exploring innovative directions. This collaborative paradigm of “AI performs the work, while people make the decisions” is reshaping the Company’s productivity and pace of innovation.

Exposure to Exchange Rate Fluctuation

The Group’s foreign exchange risk arises from future commercial transactions and recognized assets and liabilities denominated in currencies other than the functional currency of the relevant Group entities. The Company’s functional currency is Renminbi. Most non-RMB assets and liabilities are cash and cash equivalents denominated in US dollars and Hong Kong dollars.

The Group is primarily exposed to risks arising from fluctuations in RMB/USD and RMB/HKD exchange rate. Currently, we have not conducted hedging activities designed or intended to manage foreign exchange rate risks. However, we manage foreign exchange risks by closely monitoring exchange rate movements, and will take necessary measures to mitigate such risks.

Principal Risks and Uncertainties

Our operations involve certain risks and uncertainties, including (i) risks related to our research and development; (ii) risks related to our commercialization; (iii) risks related to our operations; (iv) risks related to our intellectual property; (v) risks related to our financial position and additional funding requirements; and (vi) risks related to the jurisdictions in which we conduct business, including but not limited to:

- The AI industry represents a cutting-edge technology sector characterized by constant changes. The Group's future business operations, results of operations, financial position and competitive standing hinge on its capacity to continuously upgrade and iterate its models and technologies.
- We have made and expect to continue to make substantial investments in R&D to continuously upgrade and iterate our large models.
- The Group is exposed to risks relating to customer changes and slower-than-expected development of new customers.
- The development of AGI is still at an early stage and there are substantial uncertainties in the future realization of AGI.
- We need to maintain a sufficient and stable team of technical personnel over the long term to respond to market changes in a timely manner. Any adverse changes involving our key technical personnel may affect the development, advancement and commercialization of major R&D projects of the Group, which may in turn adversely affect the Group's operations.
- We are exposed to management risks arising from business expansion.
- The Group's assets and revenue have grown rapidly in scale, while the number of personnel has continued to increase, placing higher demands on the Group's management capabilities and risk control. The Group may, to a certain extent, face management risks arising from business expansion.
- We rely on third parties to provide computing resources to us, and any disruption of their services or fluctuation of prices could adversely affect our business, results of operations and financial condition.

- China’s AI industry is subject to an evolving and increasingly stringent regulatory environment. Future laws and regulations may impose additional requirements and other obligations that could adversely affect our business, results of operations, financial condition and prospects.
- We may not be able to obtain or maintain adequate intellectual property rights protection for our business, or the scope of such intellectual property rights protection may not be sufficiently broad.
- We are exposed to risks from intensifying market competition in the future. If the Group’s technological development and product iterations fail to remain at the leading level in the industry, the Group’s competitive position, market share and operating results may be adversely affected.
- The size of our addressable market and the demand for our solutions may not increase as rapidly as we anticipate due to a variety of factors, which would adversely affect our business, results of operations, financial condition and prospects.
- Any failure of our MaaS platform to perform as required could harm our business, results of operations, financial condition and prospects.
- We have a limited track record in the commercialization of our business. In addition, the large model market is subject to dual impacts from technological iteration and changes in regulatory policies, making it difficult to accurately predict development opportunities and the future market size.

The above list is not exhaustive. For further details, please refer to the “Risk Factors” section of our Prospectus.

CONSOLIDATED STATEMENT OF PROFIT OR LOSS AND OTHER COMPREHENSIVE INCOME

For the six months ended 30 June 2026 – unaudited

(Expressed in Renminbi (“RMB”))

		Six months ended 30 June	
		2026	2025
	Note	RMB'000	RMB'000
Revenue	3	953,892	190,877
Cost of revenue		<u>(702,281)</u>	<u>(95,453)</u>
Gross profit	3(b)	251,611	95,424
Other income		41,436	4,614
Selling and marketing expenses		(177,666)	(208,570)
General and administrative expenses		(103,346)	(185,165)
Research and development expenses		(2,131,171)	(1,594,661)
Impairment losses on financial assets		<u>(27,416)</u>	<u>(10,867)</u>
Loss from operations		(2,146,552)	(1,899,225)
Finance costs		(95,925)	(53,270)
Share of profits less losses of associates		144,152	14,147
Changes in fair value of financial instruments measured at fair value through profit or loss (“FVPL”)		50,385	9,791
Changes in the carrying amounts of financial instruments issued to investors		<u>(22,090)</u>	<u>(429,295)</u>
Loss before taxation		(2,070,030)	(2,357,852)
Income tax	4	<u>(1,962)</u>	–
Loss for the period		<u><u>(2,071,992)</u></u>	<u><u>(2,357,852)</u></u>

CONSOLIDATED STATEMENT OF PROFIT OR LOSS AND OTHER COMPREHENSIVE INCOME

For the six months ended 30 June 2026 – unaudited

(Expressed in Renminbi (“RMB”))

		Six months ended 30 June	
		2026	2025
	<i>Note</i>	RMB’000	RMB’000
Loss for the period		(2,071,992)	(2,357,852)
Other comprehensive income for the period (after tax):			
Item that may be reclassified to profit or loss:			
– Exchange differences on translation of financial statements into presentation currency		(3,049)	414
Total comprehensive income for the period		(2,075,041)	(2,357,438)
Loss attributable to:			
Equity shareholders of the Company		(2,071,326)	(2,351,173)
Non-controlling interests		(666)	(6,679)
		(2,071,992)	(2,357,852)
Total comprehensive income attributable to:			
Equity shareholders of the Company		(2,074,375)	(2,350,759)
Non-controlling interests		(666)	(6,679)
		(2,075,041)	(2,357,438)
Loss per share			
Basic and diluted (RMB)	5	(4.67)	(6.23)

CONSOLIDATED STATEMENT OF FINANCIAL POSITION

At 30 June 2026 – unaudited

(Expressed in RMB)

		30 June 2026 RMB'000	31 December 2025 RMB'000
	Note		
Non-current assets			
Property, plant and equipment	6	855,060	655,825
Intangible assets		67,801	50,963
Goodwill		283,815	39,379
Interests in associates		532,133	338,081
Other non-current assets		912,325	198,202
		<u>2,651,134</u>	<u>1,282,450</u>
Current assets			
Short-term investments measured at FVPL		506,139	377,287
Inventories and contract costs		87,100	125,817
Trade and bills receivables	7	782,944	303,208
Prepayments and other receivables	8	1,621,311	395,734
Contract assets		2,480	1,625
Time deposits		–	108,593
Cash at bank and on hand		3,993,722	2,259,147
		<u>6,993,696</u>	<u>3,571,411</u>
Current liabilities			
Trade and other payables	9	2,135,759	1,328,631
Contract liabilities		302,325	148,644
Bank loans		789,831	604,836
Lease liabilities		161,711	251,316
Financial instruments issued to investors		–	10,072,827
		<u>3,389,626</u>	<u>12,406,254</u>

CONSOLIDATED STATEMENT OF FINANCIAL POSITION

At 30 June 2026 – unaudited

(Expressed in RMB)

	30 June 2026 RMB'000	31 December 2025 RMB'000
Net current assets/(liabilities)	<u>3,604,070</u>	<u>(8,834,843)</u>
Total assets less current liabilities	<u>6,255,204</u>	<u>(7,552,393)</u>
Non-current liabilities		
Bank loans	1,434,958	84,616
Lease liabilities	217,699	293,827
Deferred income	<u>172,991</u>	<u>180,146</u>
	<u>1,825,648</u>	<u>558,589</u>
NET ASSETS/(LIABILITIES)	<u>4,429,556</u>	<u>(8,110,982)</u>
CAPITAL AND RESERVES		
Share capital	44,584	40,281
Reserves	<u>4,372,464</u>	<u>(8,132,783)</u>
Total equity/(total equity – deficit) attributable to equity shareholders of the Company	4,417,048	(8,092,502)
Non-controlling interests	<u>12,508</u>	<u>(18,480)</u>
TOTAL EQUITY/(TOTAL EQUITY – DEFICIT)	<u>4,429,556</u>	<u>(8,110,982)</u>

1 BASIS OF PREPARATION

Z.AI Co., Ltd. (formerly known as “**Knowledge Atlas Technology Joint Stock Company Limited**”) was established in the People’s Republic of China (the “**PRC**”) on 11 June 2019 as a limited liability company. On 26 March 2025, the Company was converted from a limited liability company into a joint stock limited liability company. The Company’s shares were listed on the Main Board of The Stock Exchange of Hong Kong Limited (the “**Stock Exchange**”) on 8 January 2026.

The Company and its subsidiaries (together, the “**Group**”) are principally engaged in the provision of large model-related services.

The interim financial statements have been prepared in accordance with the applicable disclosure provisions of the Rules Governing the Listing of Securities on the Stock Exchange, including compliance with International Accounting Standard (“**IAS**”) 34, Interim financial reporting, issued by the International Accounting Standards Board (the “**IASB**”). It was authorised for issue on 31 August 2026.

The interim financial statements have been prepared in accordance with the same accounting policies adopted in the 2025 annual financial statements, except for the accounting policy changes that are expected to be reflected in the 2026 annual financial statements. Details of any changes in accounting policies are set out in Note 2.

The preparation of the interim financial statements in conformity with IAS 34 requires management to make judgements, estimates, and assumptions that affect the application of policies and reported amounts of assets and liabilities, income and expenses on a year-to-date basis. Actual results may differ from these estimates.

The interim financial statements contain condensed consolidated financial statements and selected explanatory notes. The notes include an explanation of events and transactions that are significant to an understanding of the changes in financial position and performance of the Group since the 2025 annual financial statements. The condensed consolidated interim financial statements and the accompanying notes do not include all of the information required for a full set of financial statements prepared in accordance with IFRS Accounting Standards.

2 CHANGES IN ACCOUNTING POLICIES

The IASB has issued a number of new and revised IFRS Accounting Standards, that are first effective for the current accounting period. Of these, only the amendments to IFRS 9, *Financial instruments* and IFRS 7, *Financial instruments: Disclosures – Amendments to the classification and measurement of financial instruments*, are relevant to the Group's financial statements. The application of the amendments does not have a material impact on the Group's consolidated financial statements for the periods presented.

The Group has not applied any new standard or interpretation that is not yet effective for the current accounting period.

The amendments would also affect the disclosures to be included in the Group's annual financial statements for the period ending 31 December 2026 in respect of investments in equity instruments designated at FVOCI and financial instruments not measured at FVPL with specified contingent features. No additional disclosure has been included in the interim financial statements.

3 REVENUE AND SEGMENT REPORTING

(a) Revenue

The principal activities of the Group are the provision of large model-related services.

(i) Disaggregation of revenue

Disaggregation of revenue from contracts with customers by major service types and timing of revenue recognition are as follows:

	Six months ended 30 June	
	2026 RMB'000	2025 RMB'000
Revenue from contracts with customers within the scope of IFRS 15		
Disaggregated by major service types:		
– On-premise deployment	128,716	161,777
– Cloud-based deployment	825,176	29,100
	<u>953,892</u>	<u>190,877</u>
Timing of revenue recognition		
At a point in time	854,270	155,636
Over time	99,622	35,241
	<u>953,892</u>	<u>190,877</u>

(b) Segment reporting

The Group manages its businesses by service types. In a manner consistent with the way in which information is reported internally to the Group's most senior executive management for the purposes of resource allocation and performance assessment, the Group has presented the following two reportable segments. No operating segments have been aggregated to form the following reportable segments.

- **On-premise deployment:** this segment develops and provides customised large model-related services according to the customers' specific instructions and needs at the customers' infrastructure.
- **Cloud-based deployment:** this segment develops and provides cloud-based large model-related services to customers through cloud infrastructure.

(i) Segment results

For the purposes of assessing segment performance and allocating resources, the Group's most senior executive management monitors the results attributable to each reportable segment on the following bases:

Revenue and expenses are allocated to the reportable segments with reference to revenue generated by those segments and direct expenses incurred by those segments. The measure used for reporting segment result is gross profit. No inter-segment sales have occurred during the period. Assistance provided by one segment to another, including sharing of assets and technical know-how, is not measured.

The Group's other operating income and expenses, such as other income, selling and marketing expenses, general and administrative expenses, research and development expenses, impairment losses on financial assets, finance costs, and assets and liabilities are not measured under individual segments. Accordingly, neither information on segment assets and liabilities nor information concerning capital expenditure, interest income and interest expenses is presented.

Information regarding the Group's reportable segments as provided to the Group's most senior executive management for the purposes of resource allocation and assessment of segment performance during the period is set out below.

	Six months ended 30 June 2026		
	On-premise deployment RMB'000	Cloud-based deployment RMB'000	Total RMB'000
Segment revenue derived from external customers	128,716	825,176	953,892
Segment gross profit	48,740	202,871	251,611

	Six months ended 30 June 2025		
	On-premise deployment	Cloud-based deployment	Total
	<i>RMB'000</i>	<i>RMB'000</i>	<i>RMB'000</i>
Segment revenue derived from external customers	<u>161,777</u>	<u>29,100</u>	<u>190,877</u>
Segment gross profit	<u>95,540</u>	<u>(116)</u>	<u>95,424</u>

(ii) *Reconciliation of reportable segment profit or loss*

	Six months ended 30 June	
	2026	2025
	<i>RMB'000</i>	<i>RMB'000</i>
Reportable segment gross profit	251,611	95,424
Other income	41,436	4,614
Selling and marketing expenses	(177,666)	(208,570)
General and administrative expenses	(103,346)	(185,165)
Research and development expenses	(2,131,171)	(1,594,661)
Impairment losses on financial assets	(27,416)	(10,867)
Finance costs	(95,925)	(53,270)
Share of profits less losses of associates	144,152	14,147
Changes in fair value of financial instruments measured at FVPL	50,385	9,791
Changes in the carrying amounts of financial instruments issued to investors	<u>(22,090)</u>	<u>(429,295)</u>
Consolidated loss before taxation	<u>(2,070,030)</u>	<u>(2,357,852)</u>

4 INCOME TAX IN THE CONSOLIDATED STATEMENT OF PROFIT OR LOSS AND OTHER COMPREHENSIVE INCOME

Taxation in the consolidated statement of profit or loss represents:

	Six months ended 30 June	
	2026	2025
	<i>RMB'000</i>	<i>RMB'000</i>
Current taxation	1,962	—
Deferred taxation	—	—
	<u>1,962</u>	<u>—</u>

Entities of the Group established in the Chinese Mainland were subject to the PRC Corporate Income Tax rate of 25% during the Reporting Period (six months ended 30 June 2025: 25%).

Taxation for subsidiaries incorporated in other jurisdictions is calculated at the applicable income tax rates in the relevant jurisdictions.

Certain subsidiaries of the Group obtained the certificates of “High and New Technology Enterprise” from the tax authorities and were subject to a preferential tax rate of 15% during the six months ended 30 June 2026 and 2025.

According to the PRC income tax law and its relevant regulations, entities that qualified as small and low profit enterprise are entitled to a preferential income tax rate of 5% (for taxable income less than RMB3,000,000). Certain subsidiaries of the Group were qualified as small and low profit enterprise and entitled preferential income tax rate during the six months ended 30 June 2026 and 2025.

5 LOSS PER SHARE

Basic loss per share is calculated by dividing the loss attributable to ordinary equity shareholders of the Company by the weighted average number of ordinary shares in issue or deemed to be in issue during the Reporting Period.

The Company was converted into a joint stock company with limited liability on 26 March 2025. The Company's paid-in capital of RMB36,224,375 was converted into 36,224,375 shares of RMB1.00 each accordingly. For the purpose of computing basic and diluted loss per share, the weighted average number of ordinary shares deemed to be in issue before the Company's conversion into a joint stock company was determined assuming the conversion into joint stock company had occurred since 1 January 2025, at the exchange ratio established in the conversion in March 2025.

In addition, the weighted average number of shares during the Reporting Period has also been adjusted retrospectively for the impact of a share subdivision that became effective immediately prior to the completion of the listing of the Company's shares on 8 January 2026.

	Six months ended 30 June	
	2026	2025
Loss for the period attributable to ordinary equity shareholders of the Company (<i>RMB'000</i>) (<i>Note 5(a)</i>)	(2,020,788)	(993,276)
Weighted average number of ordinary shares (<i>'000</i>) (<i>Note 5(b)</i>)	432,359	159,522
Basic loss per share (<i>RMB</i>)	(4.67)	(6.23)

(a) Loss for the period attributable to ordinary equity shareholders of the Company

	Six months ended 30 June	
	2026	2025
	<i>RMB'000</i>	<i>RMB'000</i>
Loss for the period attributable to all equity shareholders of the Company	(2,071,326)	(2,351,173)
Allocation of loss for the period attributable to financial instruments issued to investors	50,538	1,357,897
Loss for the period attributable to ordinary equity shareholders of the Company	(2,020,788)	(993,276)

(b) Weighted average number of ordinary shares

	Six months ended 30 June	
	2026	2025
	<i>Number of shares '000</i>	<i>Number of shares '000</i>
Ordinary shares issued at 1 January	40,281	36,224
Effect of shares issued by initial public offering	40,361	–
Effect of ordinary shares deemed to be in issue	–	1,536
Effect of the financial instruments issued to investors	(1,081)	(21,808)
Effect of share subdivision	352,798	143,570
Weighted average number of ordinary shares at 30 June	432,359	159,522

(c) **Diluted loss per share**

The financial instruments issued to investors were not included in the calculation of diluted loss per share as their inclusion would have been anti-dilutive. Accordingly, diluted loss per share for the six months ended 30 June 2026 and 2025 are the same as basic loss per share.

6 PROPERTY, PLANT AND EQUIPMENT

Acquisitions of owned assets

In May 2026, the Company completed the acquisition of the entire share capital of Biejing Hongzuan for a total consideration of RMB360,226,000, comprising (1) the cash consideration for the acquired equity interest of RMB79,158,000 (*Note 9*); and (2) the sum of RMB281,068,000 to facilitate the repayment of the relevant indebtedness of Beijing Hongzuan. Full payment of the consideration had been completed as at 21 August 2026.

The Company made the election to apply the concentration test permitted under IFRS 3, *Business Combinations*, and concluded that the acquisition of Beijing Hongzuan is an asset acquisition rather than business combination. The identifiable assets and liabilities acquired from Beijing Hongzuan as at the completion date were as follows:

	Fair value as at completion date RMB'000
Identifiable assets and liabilities acquired:	
Property, plant and equipment	360,000
Prepayments and other receivables	869
Cash and cash equivalents acquired	49
Trade and other payables	(692)
Total consideration paid/payable	360,226

7 TRADE AND BILLS RECEIVABLES

	30 June 2026 RMB'000	31 December 2025 RMB'000
Trade receivables	838,139	339,198
Bills receivables	6,445	–
	844,584	339,198
Less: loss allowance	(61,640)	(35,990)
	782,944	303,208

At the end of the Reporting Period, the ageing analysis of trade and bills receivable (net of loss allowance), based on the invoice date, is as follows:

	30 June 2026 RMB'000	31 December 2025 RMB'000
Within 3 months	626,201	243,484
3 months to 6 months	23,697	19,831
6 months to 1 year	109,486	31,406
1 year to 2 years	23,538	8,415
2 years to 3 years	22	72
	782,944	303,208

The Group generally requires customers of on-premise deployment and related services to settle progress billings, and cloud-based deployment customers to either pay in advance or due within a period of 0–60 days from the date of invoice under usage-based and related services contracts and periodic billings under subscription-based contracts and related services.

8 PREPAYMENTS AND OTHER RECEIVABLES

	30 June 2026 RMB'000	31 December 2025 RMB'000
Deposits	78,404	70,538
Receivables from disposal of investments in equity securities measured at FVPL	–	7,098
Other receivables	110,379	52,029
	188,783	129,665
Less: loss allowance	(20,923)	(18,994)
	167,860	110,671
Prepayments for computing service fee and others	1,243,736	116,295
Input VAT deductible	209,715	168,768
	1,453,451	285,063
	1,621,311	395,734

9 TRADE AND OTHER PAYABLES

	30 June 2026 RMB'000	31 December 2025 RMB'000
Trade payables	241,589	297,139
Payables for computing service fees	1,427,415	727,348
Payables for marketing and promotion services	51,004	42,812
Payables of staff costs	122,923	146,507
Consideration payable for business combination and assets acquisition (Note 6 and 11)	229,809	–
Other payables and accruals	26,869	54,437
Financial liabilities measured at amortised cost	2,099,609	1,268,243
Other taxes payables	17,515	39,774
Provisions for warranties	18,635	20,614
	2,135,759	1,328,631

As at the end of the Reporting Period, the ageing analysis of trade payables, based on the invoice date, are as follows:

	30 June 2026 RMB'000	31 December 2025 RMB'000
Within 3 months	171,575	229,890
3 months to 6 months	15,254	46,448
6 months to 1 year	50,208	19,506
More than 1 year	4,552	1,295
	241,589	297,139

10 DIVIDENDS

The directors of the Company do not recommend the payment of an interim dividend for the six months ended 30 June 2026 (for the six months ended 30 June 2025: none).

11 BUSINESS COMBINATION

On 5 June 2026, the Group, via the Company, entered into a purchase agreement with the equity holders of Beijing XCore Sigma Intelligent Technology Co., Ltd. (北京中科加禾智能科技有限公司) (“XCore Sigma”), to acquire 60% equity interests in XCore Sigma with a total cash consideration of RMB291,916,000. In accordance with the purchase agreement, the consideration will be paid in four installments, and RMB141,264,000 had been paid in June 2026. XCore Sigma is principally engaged in the development of AI-infrastructure research and development. The acquisition was completed in June 2026.

The fair value of the identifiable assets and liabilities of XCore Sigma acquired is set out as follows:

	On acquisition date RMB'000
Identifiable assets and liabilities:	
Property, plant and equipment	1,309
Intangible assets	22,510
Deferred tax assets	3,377
Inventories	3,934
Trade and bills receivables	1,373
Prepayments and other receivables	209
Cash at bank and on hand	66,567
Trade and other payables	(6,558)
Contract liabilities	(9,906)
Lease liabilities	(304)
Deferred tax liabilities	(3,377)
Fair value of net identifiable assets acquired	79,134
Less: non-controlling interests	(31,654)
Add: goodwill	244,436
Total acquisition consideration	<u><u>291,916</u></u>

Analysis of net cash outflow of cash and cash equivalents in respect of acquisition of XCore Sigma as at the date of acquisition:

	<i>RMB'000</i>
Total cash consideration	291,916
Less: cash acquired as at the date of acquisition	(66,567)
Less: consideration payable (<i>Note 9</i>)	(150,651)
Net cash outflow for the acquisition of a subsidiary included in the condensed consolidated cash flow statement	74,698

12 NON-ADJUSTING EVENTS AFTER THE REPORTING PERIOD

On 13 July 2026, the Company completed a share placement and issued 19,780,000 ordinary shares at the price of HK\$1,588. The gross proceeds from the placing were approximately HK\$31,410,640,000 (equivalent to approximately RMB27,230,826,000).

OTHER INFORMATION

Interim Dividend

The Board does not recommend the payment of an interim dividend for the six months ended June 30, 2026 (for the six months ended June 30, 2025: Nil).

Purchase, Sale, or Redemption of the Company's Listed Securities or Sale of Treasury Shares

Repurchase Mandate

The Board of the Company has been granted a general mandate (the “**Repurchase Mandate**”) pursuant to a resolution passed by the shareholders on June 22, 2026, to repurchase shares on the open market from time to time. Under the Repurchase Mandate, the Company is authorized to repurchase up to 10% of the total number of the Company's H shares issued as at the date on which the foregoing resolution was passed at the 2025 annual general meeting.

H Shares of the Company were first listed on the Main Board of the Hong Kong Stock Exchange on January 8, 2026. For the period from the Listing Date up to June 30, 2026, neither the Company nor any of its subsidiaries has purchased, sold, or redeemed any of the Company's securities listed on the Stock Exchange (including the sale of any treasury shares).

The Company does not hold any treasury shares as at June 30, 2026 within the meaning of the Listing Rules.

Significant Events After the Reporting Period

Placing of new H Shares under general mandate

On July 9, 2026 (before trading hours), the Company and China International Capital Corporation Hong Kong Securities Limited (the “**Sole Placing Agent**”) entered into a placing agreement (the “**July 2026 Placing**”), pursuant to which the Company agrees to appoint the Sole Placing Agent and the Sole Placing Agent has agreed to act as the agent of the Company, to procure not less than six (6) Placees, who will be professional, institutional or other investors, and who and whose ultimate beneficial owners shall be the Independent Third Parties, on a best effort basis, to subscribe for up to 19,780,000 new H Shares (the “**Placing Shares**”) at the Placing Price of HK\$1,588.00 per Placing Share under the general mandate granted to the Directors at the annual general meeting of the Company held on June 22, 2026. The July 2026 Placing was completed on July 13, 2026. The aggregate net proceeds from the July 2026 Placing, after deducting commission and estimated expenses, amounted to approximately HK\$31,374.95 million. For details, please refer to the announcements of the Company dated July 9, 2026 and July 13, 2026.

Save as disclosed above and in this announcement, there have been no other material events affecting the Company from the end of the Reporting Period until the date of this announcement.

Compliance with the Corporate Governance Code

The Company recognizes the value and importance of achieving high standards of corporate governance to enhance corporate performance, transparency and accountability, and further secure the trust of shareholders and the public. The Company is committed to maintaining robust corporate governance standards, and believes such standards are vital to its development and the protection of shareholders’ interests. Corporate governance is the process by which the Board directs the management of the Group in the conduct of its affairs to ensure that its objectives are met.

From the Listing Date until the date of this announcement, the Company has strictly complied with all applicable provisions of the Corporate Governance Code. In addition, the Company will actively refer to the recommended best practices of the Corporate Governance Code, including establishing a sound internal control system, enhancing the timeliness and accuracy of information disclosure, safeguarding shareholders’ right to information and participation, and continuously optimizing the governance structure.

Compliance with the Model Code

Since its Listing, the Company has adopted the Model Code as its code of conduct governing the Company's securities transactions conducted by its Directors and employees who may have access to inside information concerning the Group or the Company's securities.

Following specific enquiries made to all directors, all Directors have confirmed that they have complied with the Model Code for the period from the Listing Date up to June 30, 2026.

Employees of the Group who may possess inside information of the Group shall comply with the Model Code. For the period from the Listing Date up to June 30, 2026, the Company has not identified any incidents of non-compliance with the Model Code by the relevant employees.

Continuous Disclosure Obligations under the Listing Rules

Save as disclosed in this announcement, the Company has no other disclosure obligations under Rules 13.20, 13.21, and 13.22 of the Listing Rules.

Adequacy of Public Float

Based on the information disclosed publicly and to the best knowledge of the Directors, from the Listing Date until the date of this announcement, the Company has continuously maintained the public float as required by the Listing Rules.

Audit Committee

The Audit Committee comprises Dr. Xie Deren, Dr. Yang Qiang, and Dr. Li Juanzi. Dr. Xie Deren serves as the Chairman of the Audit Committee. The Audit Committee has reviewed the Group's interim results for the six months ended June 30, 2026, as well as the unaudited condensed consolidated financial statements of the Group for the six months ended June 30, 2026, prepared in accordance with IFRS Accounting Standards. There has been no disagreement between the Board and the Audit Committee regarding the accounting treatment adopted by the Company for the Reporting Period.

Review of Interim Financial Information

The financial information set out in this interim results announcement represents an extract from the interim financial report for the six months ended June 30, 2026, which is unaudited, but has been reviewed by KPMG, in accordance with Hong Kong Standard on Review Engagements 2410 “*Review of interim financial information performed by the independent auditor of the entity*”, issued by the Hong Kong Institute of Certified Public Accountants, whose unmodified review report is included in the interim report to be sent to shareholders.

Publication of Interim Results and Interim Report

This results announcement has been published on the websites of the Stock Exchange (www.hkexnews.hk) and the Company (www.zhipuai.cn). The Company’s interim report for the six months ended June 30, 2026 will be published on the aforementioned websites of the Stock Exchange and the Company in due course and will be sent to the Company’s H Share shareholders in accordance with their chosen method of receiving corporate communications.

DEFINITIONS AND GLOSSARY OF TECHNICAL TERMS

In this announcement, unless the context otherwise requires, the following terms shall have the meanings set out below:

Definitions

“Audit Committee”	the Audit Committee of the Board
“Board”	the board of Directors of the Company
“Corporate Governance Code”	the Corporate Governance Code as set out in Appendix C1 to the Listing Rules
“China” or “PRC”	the People’s Republic of China, and solely for the purposes of this announcement, excluding Hong Kong, the Macau Special Administrative Region, and Taiwan
“Company”	Z.AI Co., Ltd. (北京智譜華章科技股份有限公司), a limited liability company established under the laws of the PRC on 11 June 2019 and converted into a joint stock company with limited liability on 26 March 2025, with its H Shares listed on the Stock Exchange (Stock Code: 2513)
“Director(s)”	the director(s) of the Company

“Global Offering”	the Hong Kong Public Offering and the International Offering as defined in the Prospectus
“Group” or “We”	the Company and its subsidiaries
“H Shares”	the overseas-listed shares of the Company with a par value of RMB0.10 each, traded in Hong Kong dollars and listed and traded on the Stock Exchange
“Hong Kong dollar(s)”	Hong Kong dollar(s), the lawful currency of Hong Kong
“Hong Kong”	the Hong Kong Special Administrative Region of the PRC
“IFRS Accounting Standards”	IFRS Accounting Standards issued by the International Accounting Standards Board
“Independent Third Party(ies)”	third parties independent of and not connected with the Company and its connected persons
“Listing”	the listing of the H Shares on the Main Board of the Stock Exchange on January 8, 2026
“Listing Date”	January 8, 2026, the date on which the H Shares were listed on the Main Board of the Stock Exchange
“Listing Rules”	the Rules Governing the Listing of Securities on The Stock Exchange of Hong Kong Limited
“Model Code”	the Model Code for Securities Transactions by Directors of Listed Issuers as set out in Appendix C3 to the Listing Rules
“Prospectus”	the Prospectus of the Company dated December 30, 2025 in relation to the Global Offering and the Listing
“Renminbi” or “RMB”	Renminbi, the lawful currency of the PRC
“Reporting Period”	the six months ended June 30, 2026
“Share(s)”	ordinary share(s) of RMB0.10 each in the share capital of the Company, including Unlisted Shares and H Shares
“Shareholder(s)”	holder(s) of the Share(s)

“Stock Exchange” or “Hong Kong Stock Exchange”	The Stock Exchange of Hong Kong Limited
“subsidiary(ies)”	has the meaning ascribed to it under the Listing Rules
“treasury shares”	has the meaning ascribed to it under the Listing Rules
“Unlisted Share(s)”	ordinary share(s) issued by the Company, with a par value of RMB0.10 each, which is/are not listed on any stock exchange
“US dollars” or “USD”	United States dollars, the lawful currency of the United States
“%”	percentage

Glossary of Technical Terms

Glossary	Definition
“Agent”	refers to an AI model capable of autonomously executing multi-step task chains, which can invoke tools on its own, handle intermediate failures and complete multi-step tasks without step-by-step human instructions
“Agentic Engineering”	refers to the capability stage in which a model progresses from simple code generation to the delivery of complete engineering projects, where the model needs to continuously understand objectives, invoke tools, handle exceptions, and adjust strategies based on environmental feedback to achieve end-to-end engineering delivery
“AGI Commercial Value Formula”	refers to the Group’s quantitative framework for the commercial value of AGI. The formula is: AGI commercial value $\Rightarrow$ (task value density $\times$ token consumption scale), where the upper bound of intelligence determines the task boundary, and the task boundary determines the value density in which tokens operate
“All-in-Infra strategy”	refers to the infrastructure strategy fully implemented by the Group since the beginning of 2026, which involves simultaneous investment in both training and inference, and incorporates domestic chips into primary inference computing power, in order to gain control over its own cost curve
“Computing Power Multiplier”	refers to the open platform and API revenue corresponding to every RMB1 of computing power invested (covering both training and inference), used to measure the model’s self-sustaining capability and the return efficiency of computing power investment
“Co-work”	refers to the fourth level in the model capability ladder, where the model delivers work outputs that can be reviewed by professionals, the model enters real-world workflows and is priced on a per-task-completed basis; the threshold is that reliability reaches a level where professionals are willing to review rather than redo the work

“Chat”	refers to the first level in the model capability ladder, where the model delivers a single-round response, the task is completed within one round, with the lowest value density but the broadest coverage
“Coding”	refers to the second level in the model capability ladder, where the model delivers a piece of runnable code, and an objective acceptance criterion (whether it can be executed successfully) emerges for the first time; the threshold for moving from Chat to Coding is verifiable results
“CyberGym”	refers to a benchmark testing platform for evaluating the capabilities of AI models in the field of cybersecurity
“ExploitGym”	refers to a benchmark testing platform for evaluating the vulnerability exploitation capabilities of AI models
“Scaling deep”	refers to the scaling path that goes deep into post-training, reinforcement learning, long-horizon trajectories and task feedback, as opposed to Scaling up (scaling up model size, context and computing power)
“Scaling up”	refers to the path of enhancing model capabilities by scaling up model size, context length, computing power and system capacity
“GLM-5.3-Flash (Ox-Alpha)”	refers to the Group’s next-generation lightweight flagship model, testing codename Ox-Alpha, built on an entirely new architecture with 320B total parameters, 18B activated parameters and 45 layers, adopting a hybrid architecture of sparse attention and linear attention, achieving high-performance inference at extremely low cost, with its design goal being the cost frontier
“GLM-5.3”	refers to the Group’s flagship large model, which shares the same architecture, total parameters and activated parameters as GLM-5.2, with the only variable being the scale of post-training; its end-to-end completion rate improves by over 50% compared with GLM-5.2, representing the current upper bound of model capabilities

“Mid-training”	refers to the training stage where the Group separately allocates computing power between pre-training and post-training, used to enhance knowledge density, extend effective context length, and establish foundational patterns for tool invocation and long-horizon tasks
“Critic-based SAO”	refers to the Group’s self-developed critic-based single-trajectory advantage optimization, which evolved from GRPO exploration and addresses the challenges of long trajectories in reward allocation, feedback utilisation and continual learning
“OPD”	refers to the training approach that the Group has extended from supervised fine-tuning (SFT) in the algorithm dimension, in full known as Offline Policy Distillation
“Index Share”	refers to the Group’s self-developed technology used to reduce the overhead of sparse attention over long contexts
“MTP”	refers to Multi-Token Prediction technology, used to optimise inference decoding efficiency
“LayerSplit”	refers to the Group’s self-developed technology used to optimise KV Cache management and cluster network efficiency
“ZCube”	refers to the Group’s self-developed technology, which works in synergy with LayerSplit to optimise KV Cache management and cluster network efficiency
“mHC (manifold-constrained Hyperconnection)”	refers to the architectural innovation adopted by GLM-5.3-Flash, in full known as manifold-constrained Hyperconnection, used to enhance the model’s scaling capability
“Encode-Prefill-Decode disaggregated architecture”	refers to the inference architecture adopted by the Group at the production-grade cluster level, which separates multimodal encoding, prompt prefill and token-by-token decoding into independently schedulable and independently scalable worker pools, to optimise inference performance on domestic chips

“Loop Transformer”	refers to a model architecture that increases effective computational depth by recurrently reusing layers without significantly increasing the number of parameters, shifting computing power from “memorisation” to “thinking”, suitable for high-complexity tasks requiring long-horizon reasoning
“Model-vs-Model self-play pipeline”	refers to the Group’s self-developed synthetic data generation mechanism: one model generates solutions and content, while another model reviews, tests and performs execution-based verification; only data that passes adversarial scrutiny enters the training process, and is then used for pre-training, mid-training and post-training after four layers of filtering (rule-based validation, execution-based verification, model review, and manual spot checks)
“Fully Self Training”	refers to the Group’s technical vision that the next generation of GLM models will be trained in environments built by the previous generation of GLM models, forming a closed loop of Recursive Self-Improvement, encompassing three dimensions: self-generated data, self-built environments, and self-optimising infrastructure
“Recursive Self-Improvement”	refers to the recursive closed loop formed by the model’s participation in its own improvement, including generating environments with models to train the next generation of models, and optimising the inference engine that serves the model itself
“Security Disclosure Ledger (CVD)”	refers to the Group’s public security disclosure mechanism, which records the entire process of vulnerabilities from discovery, confirmation to remediation, and only discloses vulnerabilities for which disclosure has been completed and which have been fixed by the vendors; for vulnerabilities still in the coordinated disclosure stage, hash values are published for subsequent verification
“OpenVuln”	refers to the public security audit results platform under the Group’s “Open-source Shield” initiative, used to disclose to the open-source community the results of ongoing security audits of key open-source projects

“ZCode”	refers to the Group’s coding agent product, in which the Harness undertakes context management, tool invocation, task scheduling, caching and result verification, improving the model’s performance in real engineering environments, and providing code audit functionality
“GLM Coding Plan”	refers to the coding agent subscription service opened by the Group to all users in August 2026, covering Lite, Pro, Max and Team editions, and is one of the earliest large model services in China to deliver coding agent capabilities on a subscription basis
“Capability Ladder”	refers to the Group’s framework for describing the model capability evolution path, which proceeds in order: Chat → Coding → Agent → Co-work → Autonomous AI; each level upwards has specific technical thresholds, and only after these thresholds are crossed can the business model of the next level be unlocked
“Autonomous AI”	refers to the fifth level in the capability ladder, where the model delivers continuously operating capabilities and undertakes long-term objectives under unattended conditions; the threshold is that the model is able to autonomously judge the quality of task completion
“Intelligence Index”	refers to the comprehensive model intelligence score compiled by Artificial Analysis, used to measure the overall capability level of models
“Pareto Frontier”	refers to the set of models that are simultaneously optimal across the two dimensions of cost and intelligence, i.e., the frontier formed by models with the highest intelligence at a given cost, or the lowest cost at a given level of intelligence
“Task Value Density”	refers to the economic output corresponding to each unit of intelligence consumed (i.e., each unit of tokens); the value gap produced by the same scale of tokens used for different tasks (e.g., casual chat versus fixing vulnerabilities in production environments) can span several orders of magnitude

For ease of reference, the names of Chinese laws and regulations, government authorities, institutions, natural persons or other entities are listed in both Chinese and English in this announcement. In the event of any ambiguity, the Chinese version shall prevail.

This announcement contains certain forward-looking statements. These forward-looking statements are made based on information currently available to the Group or the current beliefs, expectations and assumptions of the Board. These forward-looking statements are subject to risks, uncertainties and other factors beyond the Company's control, which may cause actual results or performance to differ materially from those expressed or implied in such forward-looking statements. Given these risks and uncertainties, the forward-looking statements contained in this announcement should not be regarded as representations by the Board or the Company that the plans and objectives will be achieved, and shareholders and investors of the Company should not place undue reliance on such statements.

Unless otherwise specified, all charts, rankings and data presented in this announcement that reference third-party organizations, publicly available leaderboards, benchmarking platforms or industry research are quoted from publicly available sources and are provided for reference and illustrative purposes only. The Company was not involved in the compilation, calculation or verification of such data, rankings or benchmark results, and makes no representation, warranty or assumes any responsibility as to their methodology, sample scope, accuracy, completeness or timeliness.

By Order of the Board

Z.AI Co., Ltd.

Dr. LIU Debing

Executive Director and Chairman of the Board

Hong Kong, August 31, 2026

As at the date of this announcement, the Board comprises: (i) Dr. Liu Debing, Dr. Zhang Peng and Ms. Zhang Xiaohan as executive Directors; (ii) Dr. Li Juanzi, Mr. Li Jiaqing and Mr. Wang Meng as non-executive Directors; and (iii) Dr. Yang Qiang, Dr. Xie Deren and Mr. Xu Wenming as independent non-executive Directors.