

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.



UNISOUND AI TECHNOLOGY CO., LTD.

雲知聲智能科技股份有限公司

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

(Stock Code: 9678)

VOLUNTARY ANNOUNCEMENT RELEASE OF U2-FLASH, A NEW-GENERATION HIGH-PERFORMANCE CORE MODEL

This announcement is made by Unisound AI Technology Co., Ltd. (the “**Company**”) on a voluntary basis to inform the shareholders and potential investors of the Company of its latest technological developments.

The Company officially released U2-Flash, a new-generation high-density intelligent model built on reinforced post-training of the U2 general foundation model. Compared with the previous-generation U2 model, U2-Flash achieves significant improvements in task completion quality and overall efficiency: on the DeepSWE v1.1 benchmark, which represents coding capability, U2-Flash doubled its score compared with the previous-generation model, surpassing models such as GLM5.3-Flash and DeepSeek-V4-Pro-0813 with a score of 64.6; its TerminalBench3.0 score increased substantially to 24.3, exceeding trillion-parameter-level models such as K3. Its SWE-Bench Pro score reached 61.6, an increase of 10.5 points over the previous generation. Meanwhile, the model achieves comprehensive optimization in end-to-end execution efficiency and inference cost. The number of iteration steps for Agent tasks is reduced by 20%–30%, the task execution cycle is shortened by 35%, and Token consumption is reduced by 20%–30%.

U2-Flash adopts a sparse Mixture-of-Experts (MoE) architecture with a total of approximately 266B parameters and activates approximately 10B parameters per inference. Its average time to first token is kept within 3 seconds, with peak output throughput of up to 300 Tokens/s. This model unifies multi-domain capabilities such as programming, Agent, mathematical reasoning, and instruction following within a single set of weights, achieving flagship-level task completion quality at an activation scale far smaller than comparable dense models. The model is equipped with mechanisms for implicit thinking and continuous-state reasoning, and encapsulates reasoning intensity into a four-level controllable interface. At high-intensity levels, it outputs a complete and readable chain of reasoning, ensuring the reasoning process is verifiable and traceable. In addition, U2-Flash has completed systematic adaptation to mainstream domestic computing power platforms, providing more flexible computing options for large-scale deployment.

In terms of training methodology, U2-Flash establishes an autonomous closed-loop mechanism where the model deeply participates in its own training, representing the Company's first practical step towards Recursive Self-Improvement (RSI): the model can participate in task generation, trajectory analysis, and error correction with resampling, autonomously constructing a high-quality SWE task set of nearly 100,000 tasks. Through innovations such as asynchronous Agent RL, online policy distillation from multiple self-trained teacher models, and adaptive dynamic task generation, the number of effective training trajectories has increased by approximately 60%, training steps have been reduced by approximately 55%, and the training system can perform autonomous inspection and repair. All autonomous adjustments are conducted within a manually set sandbox environment and according to predefined validation standards, with complete and roll-backable records maintained.

The release of U2-Flash marks the transition of the Company's large model technology from the stage of "human-led model training" to a new stage of "model participation in its own evolution". Through the RSI mechanism, the model can continuously self-validate, self-correct, and self-enhance in real-world tasks, laying the foundation for the long-term, compounding improvement of the model's capabilities. Looking ahead, the Company will adhere to a technical path that prioritizes validation, ensures full traceability, and maintains clear boundaries. While ensuring safety and controllability, it will steadily expand the scope of the model's participation in its own capability growth, promote the large-scale implementation of high-density intelligence in more real-world production scenarios, and create long-term value for shareholders.

By Order of the Board
Unisound AI Technology Co., Ltd.

Dr. Huang Wei
Executive Director and Chief Executive Officer

Beijing, the PRC
September 15, 2026

As of the date of this announcement, the board of directors of the Company comprises: (i) Dr. Liang Jia'en, Dr. Huang Wei, Dr. Kang Heng, Dr. Li Xiaohan and Dr. Liu Shengping as executive directors; and (ii) Mr. Hu Jianjun, Dr. Fan Jian, Dr. Jin Huihua and Dr. Zhang Kun as independent non-executive directors.