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-1118, A BRAND-NEW UPGRADED VERSION OF THE LARGE 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-1118. This model is a brand-new upgraded version of U2-Flash, achieving comprehensive leaps in three major capabilities of complex coding, autonomous Agent execution and multimodal understanding. In multiple authoritative evaluations, U2-Flash-1118 has demonstrated leading performance, achieving excellent results in the three major domains of complex coding and Agent execution, native visual understanding, and scientific and mathematical reasoning. The specific evaluation results are as follows:

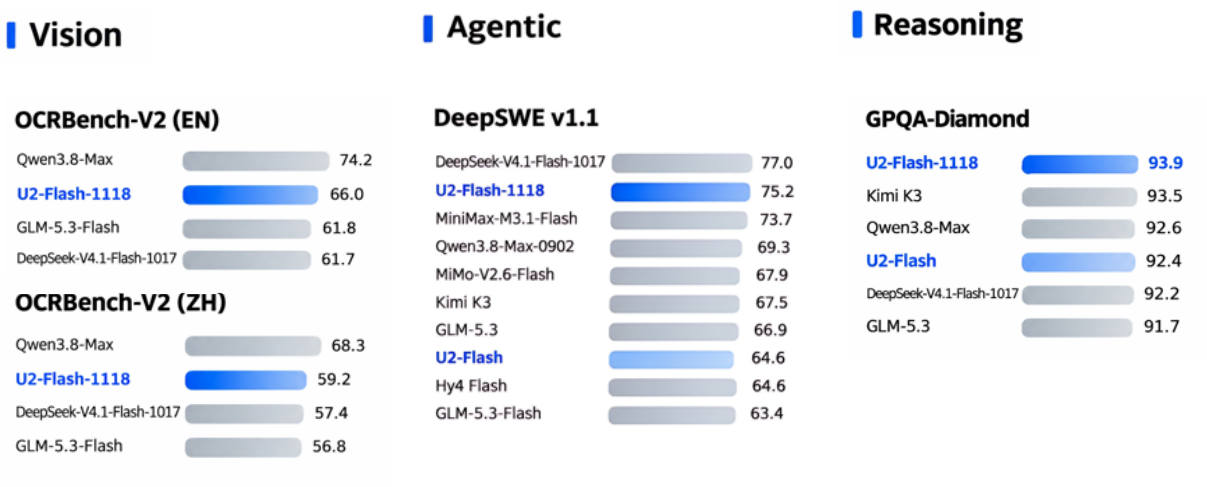


Figure: evaluation results across the three major domains

i. In terms of native visual understanding, the model ranked among the top two in all three visual evaluations evaluated. Specifically, its scores on both English and Chinese evaluations of OCRBench-V2 outperformed DeepSeek-V4.1-Flash-1017 and GLM-5.3-Flash, demonstrating outstanding capabilities in graphic detail recognition, spatial relationship perception and Chinese and English document layout understanding.

ii. In terms of complex coding and Agent execution, the model outperformed U2-Flash in all four engineering evaluations evaluated. Specifically, it ranked second in DeepSWE v1.1 evaluation, surpassing MiniMax-M3.1-Flash and Qwen3.8-Max-0902, demonstrating excellent capabilities in parsing large code repositories, decomposing complex engineering requirements, multi-round code modification and terminal task execution.

iii. In terms of scientific and mathematical reasoning, the model ranked first and second in the two reasoning capability evaluations evaluated, respectively. Specifically, it ranked first in GPQA-Diamond evaluation and outperformed Kimi K3 and Qwen3.8-Max, demonstrating excellent capabilities in high-difficulty scientific question-answering and competition-level mathematical reasoning.

While maintaining the parameter size of its original language base model, U2-Flash-1118 has enhanced its native multimodal understanding capability by adding a vision module with approximately 1B parameters, and has further improved its complex task execution capability and Token usage efficiency through post-training and collaborative optimization with Agent Harness. Its capability upgrades are mainly in the following three major aspects:

(1) **Enhanced multimodal understanding capability:** The model introduces native image input capability, enabling direct understanding of UI screenshots, design drafts and document layouts, and supporting image-text Q&A, interface understanding and screenshot-based assisted development.

(2) **Optimized execution of complex tasks:** Through the Recursive Self-Improvement (RSI) mechanism, task execution feedback is converted into targeted training tasks, and through error correction training, independent testing and capability regression verification, the model and the Agent execution framework are continuously optimized, and can autonomously advance the complete development process from requirement understanding, issue diagnosis, code modification to test verification.

(3) **Improved Token efficiency:** In terms of Agent execution efficiency, repetitive exploration, redundant tool calls and context transfer during the task execution process are reduced through execution trajectory optimization, key action refinement, dynamic budget allocation and collaborative optimization with Agent Harness. Under the same task set and consistent evaluation conditions, compared to the original U2-Flash version, total Token consumption is reduced by approximately 37.0% and the number of model calls is reduced by approximately 44.8%, further lowering the task execution cost while maintaining the quality of task completion.

By enhancing its multimodal understanding, improving complex task execution, and optimizing Agent execution efficiency, U2-Flash-1118 further boosts the model's comprehensive capability in practical application scenarios, offering more efficient model services to developers and enterprise users.

By Order of the Board
Unisound AI Technology Co., Ltd.
Dr. Huang Wei
Executive Director and Chief Executive Officer

Beijing, the PRC
November 18, 2026

As at 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.