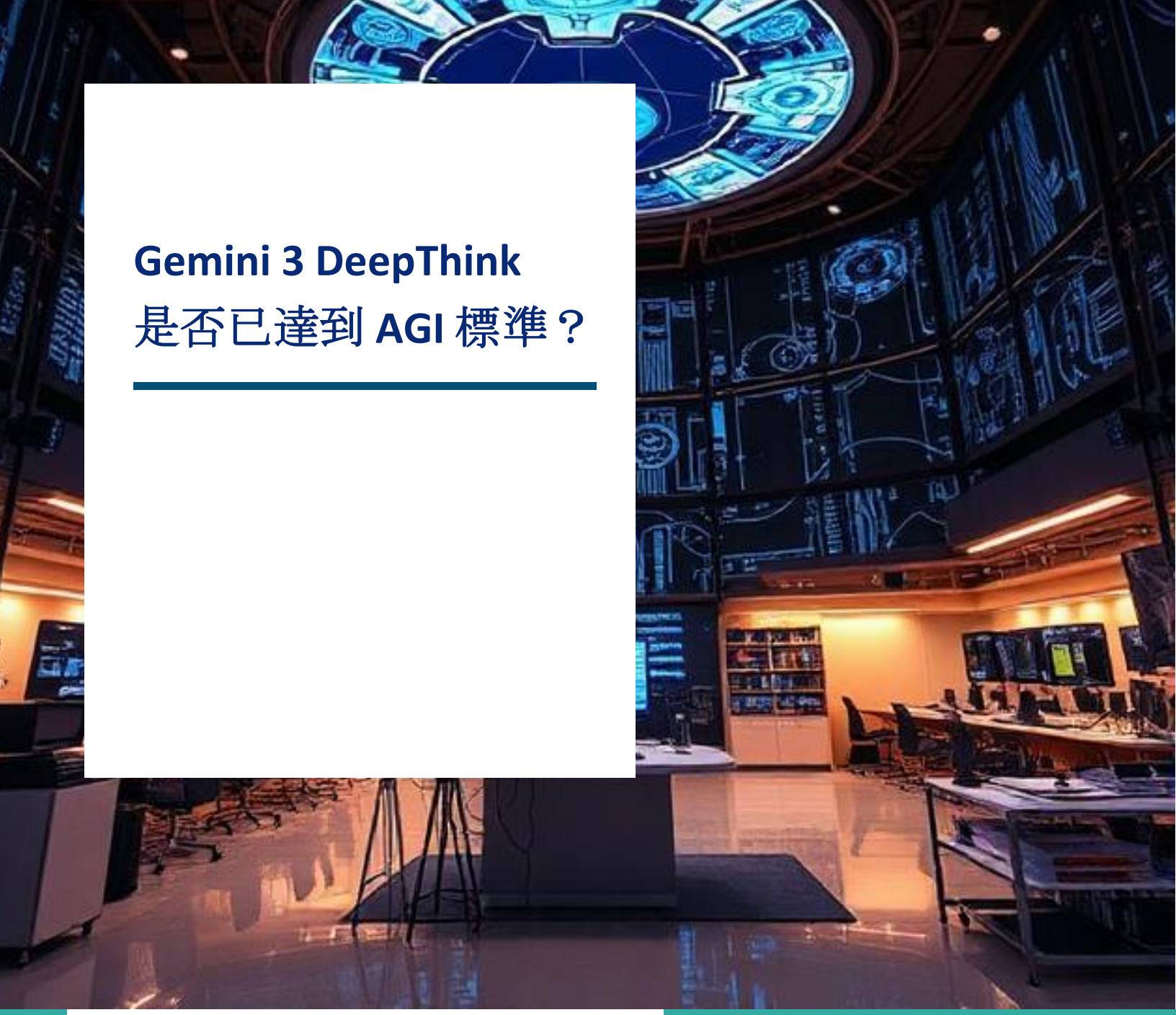

Gemini 3 DeepThink
是否已達到 AGI 標準？

2026年2月19日

人工智能中心
https://www.aicenters.info/
生成：Claude
編輯、修正：屈思宏

Gemini 3 DeepThink 是否已達到 AGI 標準？

Gemini 3 Deep Think 在 ARC-AGI-2 取得 84.6%，推理能力大幅提升，並在程式設計與奧林匹亞筆試達金牌水準，部分表現超越人類平均，被視為最接近 AGI 的系統之一。然而，多數專家仍認為尚未達到 AGI。原因包括：推理成本遠高於人類、缺乏持續自主學習與互動式推理驗證，且 ARC 基金會明言 AGI 問題仍未解決。整體而言，它是邁向 AGI 的重要里程碑，但尚非終點。

一、前言：什麼是 AGI？

人工通用智能（Artificial General Intelligence，AGI）至今仍無統一定義，但學術界普遍認為，AGI 應具備以下特質：能在任意領域自主學習新技能、以與人類相當甚至更高的效率解決前所未見的問題、具備靈活的抽象推理能力，而非依賴記憶化的訓練資料。這一高門檻正是 Gemini 3 Deep Think 所試圖挑戰的。

二、Gemini 3 Deep Think 的核心表現

2026年2月12日，Google DeepMind 宣布對 Gemini 3 Deep Think 進行重大升級，定位為專攻科學、研究與工程難題的推理模式。該模式在多項指標上創下新紀錄：在 ARC-AGI-2 基準測試中取得前所未有的 84.6% 成績（由 ARC Prize Foundation 驗證）；在 Codeforces 競技程式設計平台上達到 3455 Elo 分數；並在 2025 年國際數學奧林匹亞、國際物理奧林匹亞及國際化學奧林匹亞的筆試部分均達到金牌水準。[1]

在「人類最後一次考試」（Humanity's Last Exam）這一專為測試頂尖前沿模型極限而設計的基準測試中，Gemini 3 Deep Think 在不使用外部工具的情況下取得 48.4% 的成績。[2]

與競爭對手相比，Gemini 3 Deep Think 在 ARC-AGI-2 以 84.6% 顯著領先，勝過 Claude Opus 4.6 Thinking Max 的 68.8% 及 GPT-5.2 Thinking xhigh 的 52.9%。[3]

三、ARC-AGI-2：最接近 AGI 測量的基準？

ARC-AGI-2 是目前被認為最接近用來測量「通用推理能力」的測試標準之一。它由 François Chollet 在 2019 年設計。這個測試的核心想法是：題目對人類來說相對容易理解，但對人工智能而言卻非常困難，藉此檢驗 AI 是否真的具備像人類一樣的推理能力。經驗證，ARC-AGI-2 中 100% 的任務均可由至少兩名非專業人類測試者解決，人類平均每題僅需 2.7 分鐘，整體成功率約為 75%。[4]

Gemini 3 Deep Think 的 84.6% 成績是業界的重大躍進——相較之下，早期的 AI 模型往往難以突破 20%；而人類平均分數約為 60%，意味著該模型在此測試中已超越人類平均水準。[5]

四、反方論點：高分不等於 AGI

儘管成績令人矚目，多數研究者仍持保留態度。

ARC Prize Foundation 的立場：ARC Prize Foundation 明確表示，AGI 問題尚未解決，仍需要新的想法，例如如何分離知識與推理；AI 推理系統儘管表現出色，仍存在許多 AGI 所必需的缺陷與低效問題。[6]

高成本問題：ARC-AGI-2 不僅評估準確率，也衡量效率，因為智能不只是解決問題，而是高效地解決問題。人類解決每題的成本約為 17 美元，部分 AI 系統卻需要數百美元——這一差距才是真正的信號。[7]

定義本身有爭議：François Chollet 等專家強調，ARC-AGI 旨在揭示 AI 的不足之處；即便在 GPT-5.2 等模型上取得進展，流動性概括（fluid generalization）仍超出當前架構的能力範圍。[8]

ARC-AGI-3 正在路上：ARC Prize Foundation 計劃在 2026 年初推出 ARC-AGI-3，這次將從靜態網格推理轉型為互動式推理，要求具備探索、規劃、持久記憶與目標推斷能力——這是 Gemini 3 Deep Think 尚未被測試的維度。[6]

五、Google 自身的定位

值得注意的是，Google 在推出 Gemini 3 時的措辭非常謹慎。Google DeepMind CEO Demis Hassabis 描述這是「通往 AGI 道路上的重要一步」，而非宣稱 AGI 已然實現。[9] 這種自我節制的表述，反映出即使是開發者本身也未將其定性為 AGI。

六、綜合評估

評估面向	結論
抽象推理（ARC-AGI-2）	超越人類平均，但未達人類滿分（100%）
知識深度（奧林匹亞金牌）	超人類水準
效率	每題成本遠高於人類
自主學習與持續學習	尚無法驗證
互動式推理（ARC-AGI-3 標準）	尚未測試
業界/學界共識	尚未達到 AGI

結論

Gemini 3 Deep Think 在多項核心推理基準上取得了劃時代的成績，某些面向甚至已超越人類平均水準，是迄今最接近 AGI 定義的 AI 系統之一。然而，目前尚不能認定其已達到 AGI 標準。主要原因有三：其一，ARC Prize Foundation 明確表示 AGI 問題尚未解決；其二，模型的推理成本依然遠超人類；其三，AGI 所需的持續自主學習、互動環境下的目標推斷等能力，仍缺乏充分驗證。Gemini 3 Deep Think 代表的是通往 AGI 道路上的一個里程碑，而非終點。

主要資料來源

[1] Gemini 3 Deep Think: Advancing science, research and engineering, Feb 12, 2026: https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-deep-think/

[2] Google's new Gemini 3 Deep Think update pushes the boundaries of AI reasoning, Feb 12, 2026: https://chromeunboxed.com/googles-new-gemini-3-deep-think-update-pushes-the-boundaries-of-ai-reasoning/

[3] Google Releases Gemini 3 Deep Think, Tops ARC-AGI 2 Benchmark With 84.6%, Feb 12, 2026: https://officechai.com/ai/gemini-3-deep-think-benchmarks-arc-agi/

[4] ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems, Jan 15, 2026: https://arxiv.org/html/2505.11831v2

[5] Is This AGI? Google's Gemini 3 Deep Think Shatters Humanity's Last Exam And Hits 84.6% On ARC-AGI-2 Performance Today, Feb 12, 2026: https://www.marktechpost.com/2026/02/12/is-this-agi-googles-gemini-3-deep-think-shatters-humanitys-last-exam-and-hits-84-6-on-arc-agi-2-performance-today/

[6] ARC Prize 2025 Results & Analysis, Dec 5, 2025: https://arcprize.org/blog/arc-prize-2025-results-analysis

[7] Best LLM for reasoning in 2026: ARC-AGI-2 benchmark results, Jan 29, 2026: https://www.bracai.eu/post/arc-agi-2-benchmark

[8] GPT-5.2 & ARC-AGI-2: A Benchmark Analysis of AI Reasoning: https://intuitionlabs.ai/articles/gpt-5-2-arc-agi-2-benchmark

[9] A new era of intelligence with Gemini 3, Nov 18, 2025: https://blog.google/products-and-platforms/products/gemini/gemini-3/