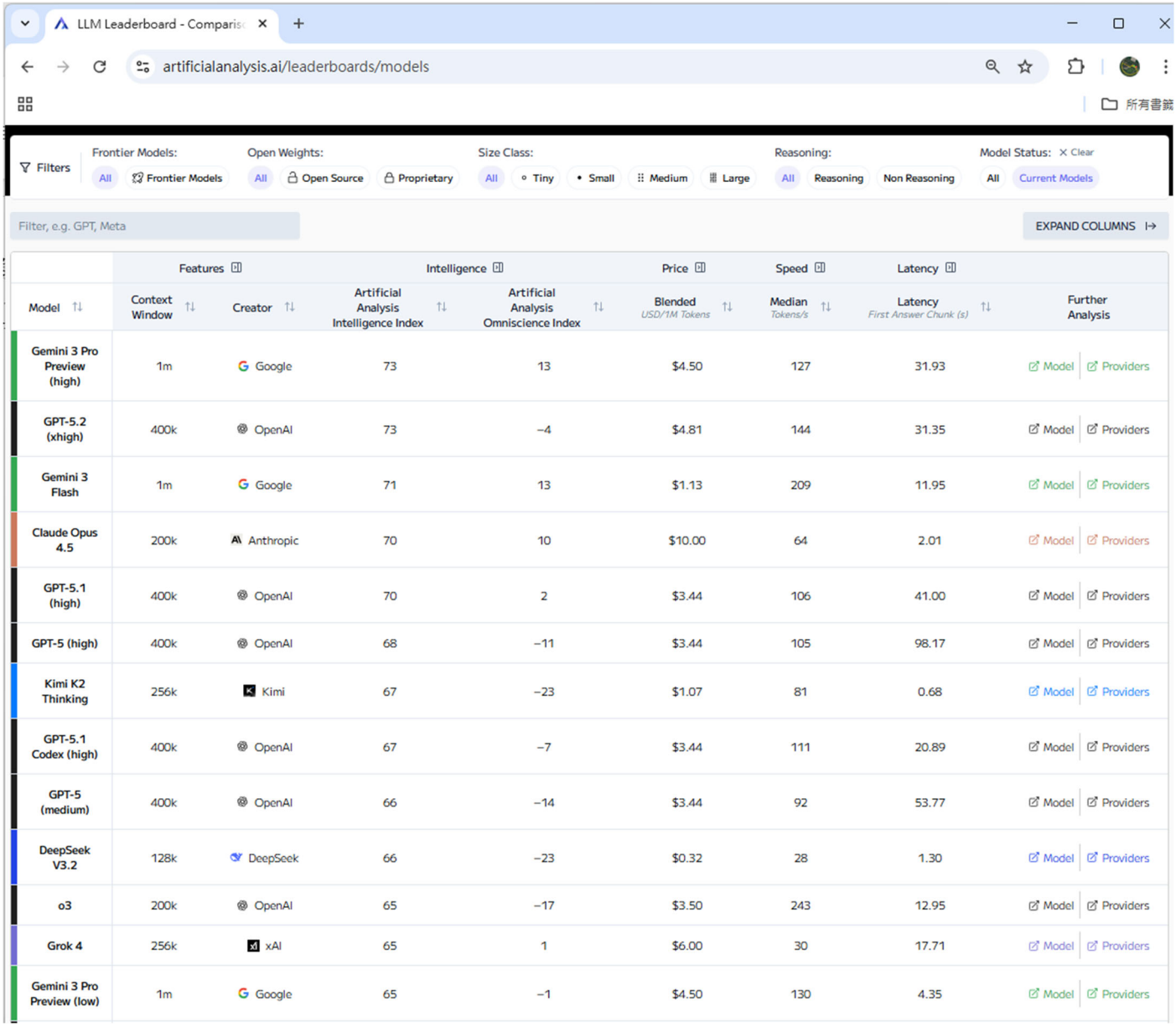
LLM 模型效能評比

資料來源：AI Analysis　資料時間：2025/12/19 09:00
編撰：屏東大學 周國華老師 (與 ChatGPT 共筆)

AI Analysis Leaderboard:



	Features		Intelligence		Price	Speed	Latency	
Model	Context Window	Creator	Artificial Analysis Intelligence Index	Artificial Analysis Omniscience Index	Blended USD/1M Tokens	Median Tokens/s	Latency First Answer Chunk (s)	Further Analysis
Gemini 3 Pro Preview (high)	1m	Google	73	13	$4.50	127	31.93	Model Providers
GPT-5.2 (xhigh)	400k	OpenAI	73	−4	$4.81	144	31.35	Model Providers
Gemini 3 Flash	1m	Google	71	13	$1.13	209	11.95	Model Providers
Claude Opus 4.5	200k	Anthropic	70	10	$10.00	64	2.01	Model Providers
GPT-5.1 (high)	400k	OpenAI	70	2	$3.44	106	41.00	Model Providers
GPT-5 (high)	400k	OpenAI	68	−11	$3.44	105	98.17	Model Providers
Kimi K2 Thinking	256k	Kimi	67	−23	$1.07	81	0.68	Model Providers
GPT-5.1 Codex (high)	400k	OpenAI	67	−7	$3.44	111	20.89	Model Providers
GPT-5 (medium)	400k	OpenAI	66	−14	$3.44	92	53.77	Model Providers
DeepSeek V3.2	128k	DeepSeek	66	−23	$0.32	28	1.30	Model Providers
o3	200k	OpenAI	65	−17	$3.50	243	12.95	Model Providers
Grok 4	256k	xAI	65	1	$6.00	30	17.71	Model Providers
Gemini 3 Pro Preview (low)	1m	Google	65	−1	$4.50	130	4.35	Model Providers

以下依 ArtificialAnalysis.ai 的 LLM Leaderboard 介面，逐一解說畫面中出現的評比欄位與篩選項目。

一、上方篩選（Filters）的意義

❑ Frontier Models

- **Frontier Models**：當前「最前沿」的大型模型（如 GPT-5、Claude Opus、Gemini Pro）
- **All**：包含非前沿與較小模型

👉 用途：
如果你只想看「SOTA 等級(= State of the art 當前最高水準)」模型，就勾 Frontier Models。

2️⃣ Open Weights

- **Open Source / Open Weights**：模型權重可取得（如 DeepSeek、部分 Kimi）
- **Proprietary**：封閉商用模型（OpenAI、Anthropic、Google）

👉 用途：

- 教學、研究、私有部署 → 看 Open Weights
- 商業 API 穩定度 → 看 Proprietary

3️⃣ Size Class

- **Tiny / Small / Medium / Large**：模型規模分類（不是參數數字，而是效能等級）

👉 用途：
在「成本 vs 能力」取捨時快速篩選。

4️⃣ Reasoning

- **Reasoning**：強化推理能力（如 GPT-5.2、o3、Kimi Thinking）
- **Non-Reasoning**：偏生成 / 反應型模型

👉 用途：

- 邏輯推理、數學、法規分析 → Reasoning
- 摘要、翻譯、客服 → Non-Reasoning 即可

Model Status

- Current Models：目前可用
- 排除已淘汰或 preview 結束的模型

二、主要評比欄位（表格核心）

🔹 Model

模型名稱與版本

- GPT-5.2 (xhigh)
- Claude Opus 4.5
- Gemini 3 Pro Preview

👉 **版本非常重要**，不同 preview / high / medium 差異很大。

🔹 Context Window

可處理的 **最大上下文長度**

- 128k / 200k / 400k / 1m tokens

👉 教學重點可這樣說：

Context window ≈「一次能讀多厚的財報、法規或論文」

你這行背景（XBRL、IFRS、ESG）會**非常在意這個指標**。

🔹 Creator

模型開發商：

- OpenAI、Google、Anthropic、DeepSeek、xAI、Kimi

三、Intelligence（智能指標，最常被誤解）

🧠 Artificial Analysis Intelligence Index

- 綜合「推理、理解、生成」能力
- 類似 **整體智力分數**
- 數字越高 ≠ 每個任務都最好，但「平均能力」高

例如：

- 73：頂級前沿
- 65～68：一線實用
- <60：偏輕量或特定用途

🔮 Artificial Analysis Omniscience Index

- 偏重 **知識覆蓋面與廣度（Omniscience 意指「全知全能」、「無所不知」、「博聞廣識」）**
- 不等於「準確率」，而是「知道多少東西」

⚠️ 很重要的一點：

- **負數是正常的**
- 表示該模型「不追求全知型回答」，而偏向謹慎或推理

👉 對會計／法規分析：

- Omniscience 高 ≠ 合規好
- Reasoning + 低 hallucination 才重要

四、Price（成本）

💰 Blended USD / 1M Tokens

- 每 100 萬 tokens 的平均成本
- Blended = input + output 的平均
- 註：Token = AI 讀與寫文字時的「最小單位」，例如：畢氏定理「$a^2 + b^2 = c^2$」≈ 8 tokens（依模型略有差異），簡單英文「Accounting is fun.」 約 4 tokens。

例：

- DeepSeek V3.2：$0.32（超便宜）
- Claude Opus：$10（昂貴）
- GPT-5 系列：約 $3～5

👉 教學可這樣比喻：

像是「每處理一百萬字的會計資料要多少錢」

五、Speed & Latency（效能體感）

⚡ Speed – Median Tokens/s

- 每秒生成幾個 token（速度）
- 數字越高 → 回答「打字」越快

⏱️ Latency – First Answer Chunk (s)

- **第一次回應出現的時間**
- 對互動式應用非常重要
- 註：Chunk 是話語組成部分，組塊

例如：

- 0.68s：幾乎即時（Kimi K2）
- 30s+：適合離線分析，不適合即時對話

👉 教學可用：

- Speed = 打字速度
- Latency = 想多久才開口說第一句話

六、Further Analysis

🔗 Model / Providers

- **Model**：模型詳細評測頁
- **Providers**：有哪些平台提供此模型 API

👉 實務選型一定會點這裡。

七、給你的「專業級解讀建議」（重點）

以你在 **AIS / XBRL / ESG / 教學** 的背景，我會這樣教學生讀這張表：

✔ 不要只看 Intelligence 排名

而是三件事一起看：

1. **Context window**
2. **Reasoning or not**
3. **Latency + Price**

✔ 教學示範可這樣分組

- **法規／準則分析**：GPT-5.2、Claude Opus、o3
- **大量文本摘要**：Gemini 3 Pro、GPT-5 medium
- **低成本實驗 / 作業**：DeepSeek、Kimi

✔ 這張表是「模型工程視角」

不是「誰最聰明排行榜」，而是：

在限制條件下，誰最適合某種任務