資料來源：ChatGPT　編輯：屏東大學 周國華老師

以下是對 **MMLU（Massive Multitask Language Understanding，大規模多任務語言理解測試）** 的清楚解析與教學式說明，可用於課堂、報告或研究參考。

☑ 什麼是 MMLU？

MMLU 是一個用來評估大型語言模型（如 GPT、Claude、Llama 等）**在跨領域知識與推理能力**上的基準測試（benchmark）。

它測試 AI 是否具備「類人類的知識理解力」，涵蓋 **57 個領域、4 個難度層級、13,908 題選擇題**，包括：

- 初等教育（elementary）
- 高中（high school）
- 大學（college）
- 專業或進階（professional level）

☑ MMLU 測驗內容涵蓋哪些領域？

類別	範例領域
STEM	數學、物理、化學、生物、計算機科學
社會科學	經濟學、法學、心理學、社會學
人文	歷史、哲學、文學、宗教研究
醫學與專業領域	醫學、護理、營養學、法律、商業倫理

☑ 題型長什麼樣？

MMLU 主要是選擇題，格式如下：

Question:
What is the derivative of sin(x)?

A) cos(x)
B) -sin(x)
C) -cos(x)
D) sin(x)

模型必須選出正確選項，不允許 Chain-of-Thought 思考過程公開（通常採「直接回答模式」）。

☑ MMLU 為什麼重要？

重要性	說明
☑ 測試模型是否具有真實知識（而非只是語言模仿）	
☑ 評估模型跨學科能力（Multi-task）	
☑ 作為 GPT-4、Claude、PaLM、LLaMA 等模型的對照指標	
☑ 是目前國際最常用的 LLM Benchmark 之一	

☑ 主流模型在 MMLU 的表現（2024–2025 常見數據）

模型	MMLU 分數（%）
GPT-4o / GPT-5*	86–89%
Claude 3.5	約 86%
GPT-3.5	約 70%
LLaMA 2 (70B)	約 53–56%
人類平均（大學生）	約 89%

*GPT-5 的評測數據因尚未公開，通常推估介於 88–90%。

☑ 如何測試模型的 MMLU？

常見方式包括：

☑ 方式 1：OpenAI API（測評模式）

```
from openai import OpenAI
client = OpenAI()

response = client.chat.completions.create(
    model="gpt-4o",
    messages=[
        {"role": "system", "content": "Answer with only the correct
option letter."},
        {"role": "user", "content": "What is the capital of
France?\nA) Berlin\nB) Madrid\nC) Paris\nD) Rome"}
    ]
)
print(response.choices[0].message.content)
```

☑ 方式 2：Hugging Face + 🤗 Evaluation Library

```
from lm_eval import evaluator

results = evaluator.simple_evaluate(
    model="hf-causal",
    model_args="pretrained=meta-llama/Llama-2-7b-hf",
    tasks=["mmlu"],
    batch_size=8,
)
print(results["mmlu"])
```

☑ 總結懶人包

關鍵詞	重點
MMLU 是什麼？	評估模型知識與理解力的 57 科目大測驗
題型？	多選題（A/B/C/D）

關鍵詞	重點
為何重要？	測 AI 是否真的理解知識，而非僅背句子
GPT-4/5 的分數？	約 86－90%
是否可教學？	非常適合作為「AI 智能測驗」教材或競賽題庫來源

如果您需要：

✔ MMLU 類似題庫（中文/英文版）
✔ 給學生的 AI 模擬測驗系統
✔ Python / Colab 自動評分程式
✔ MMLU 與台灣會計 / ESG / XBRL 題目融合

我可以幫您製作，您要我開始嗎？