# What is the difference between LLM and SLM?

APMIC Founder & CEO Jerry Wu

APMIC

APMIC

# Accelerate Private Machine Intelligence Company

**Leading provider of accelerate enterprise-grade Private AI solutions**

# APMIC - 協助企業的每個人都成為AI開發者

APMIC成立於2017年

- 由Google機器學習開發專家(GDE) Jerry成立
- 專注於自然語言理解 (Natural Language Understanding)
- 提供大語言模型即服務 (LLM as a Service)

APMIC 團隊擁有 7 年的語言模型應用服務經驗

HUGGING FACE 開源大語言模型榜單,自研發模型台灣唯二入榜

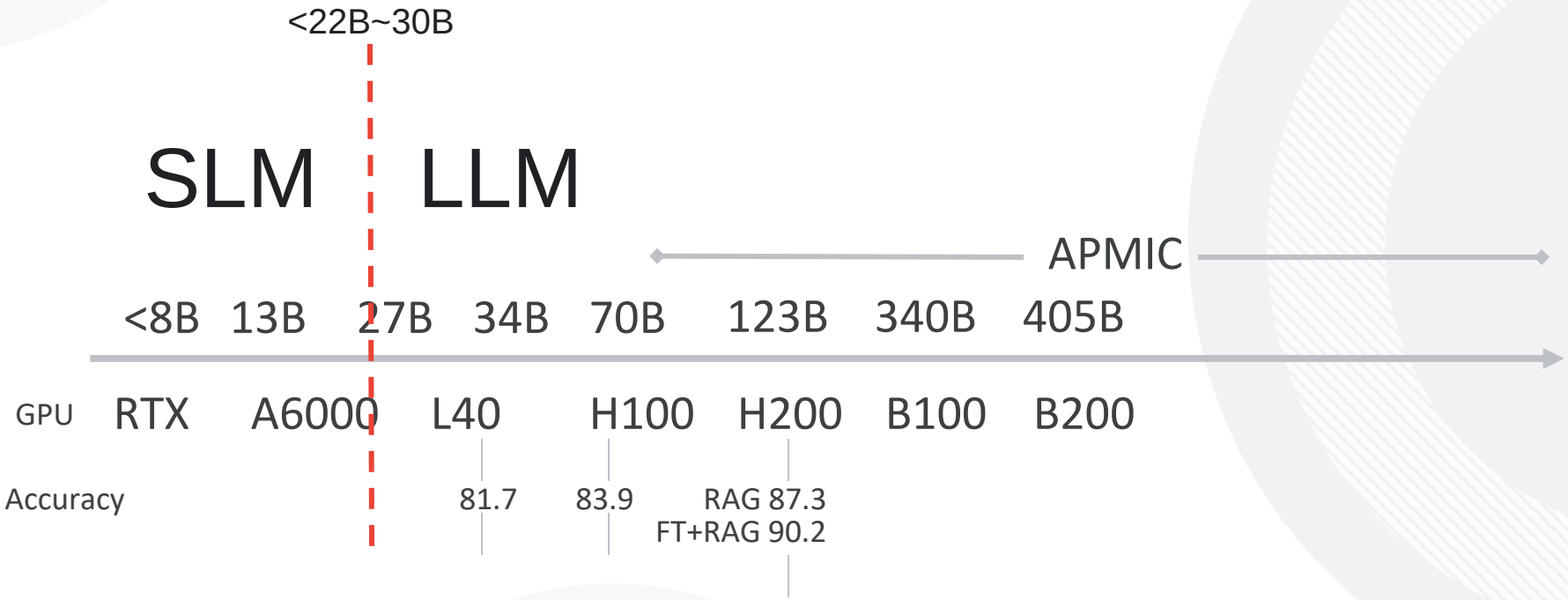
全球排名 Top 100 (2024.1)

國立成功大學 National Cheng Kung University
國立臺灣科技大學 NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY
國立陽明交通大學 NATIONAL YANG MING CHIAO TUNG UNIVERSITY
國立臺灣大學 National Taiwan University

# CaiGunn(人人都可以成為AI開發者)

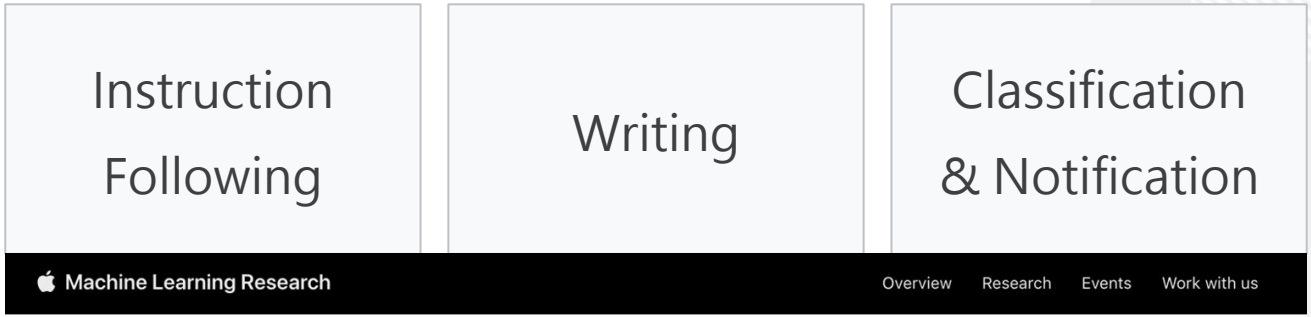

全系列支援，免費玩：
Gemini 1.5、Gemma 27B、Llama3.1 405B、TaME等模型的平台
目前破千個使用者

# From SLM to LLM



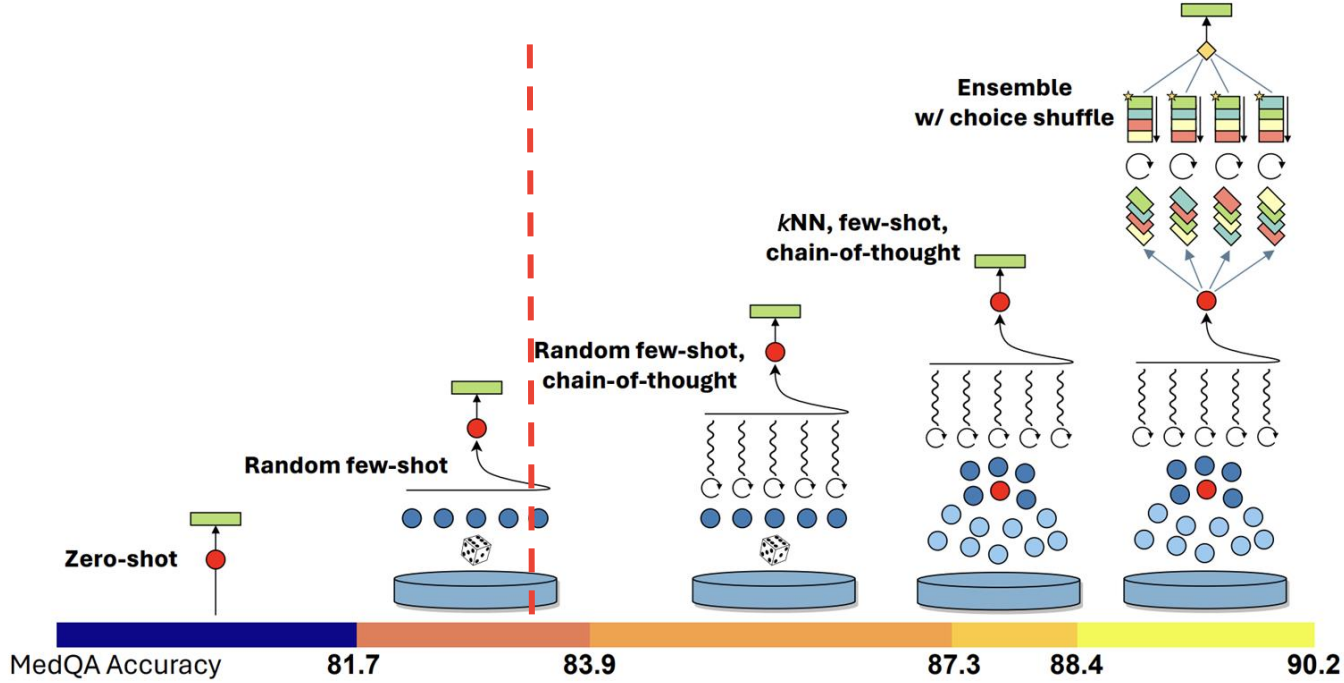
# Small Language Models

Apple Foundation Models has analyzed the strengths of models smaller than 1-10B.



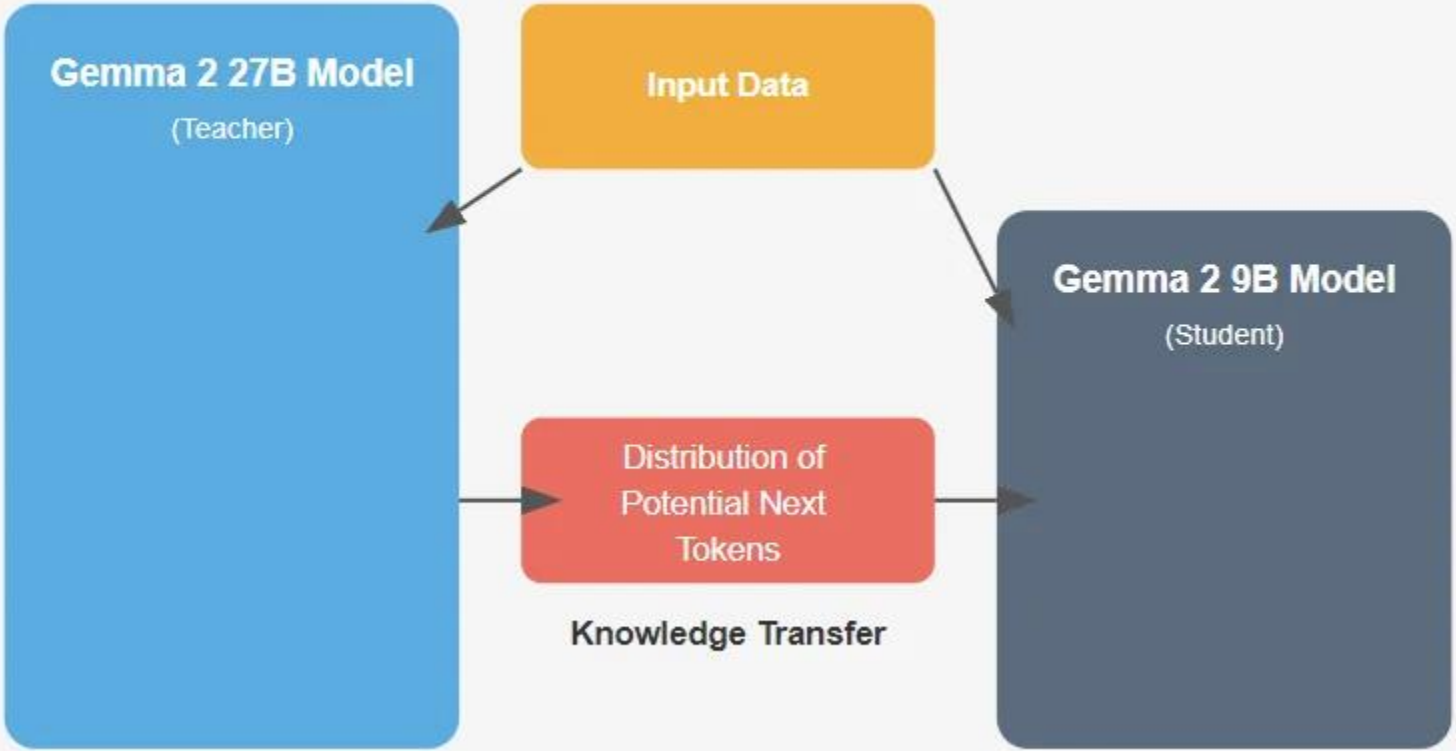
# Large language model

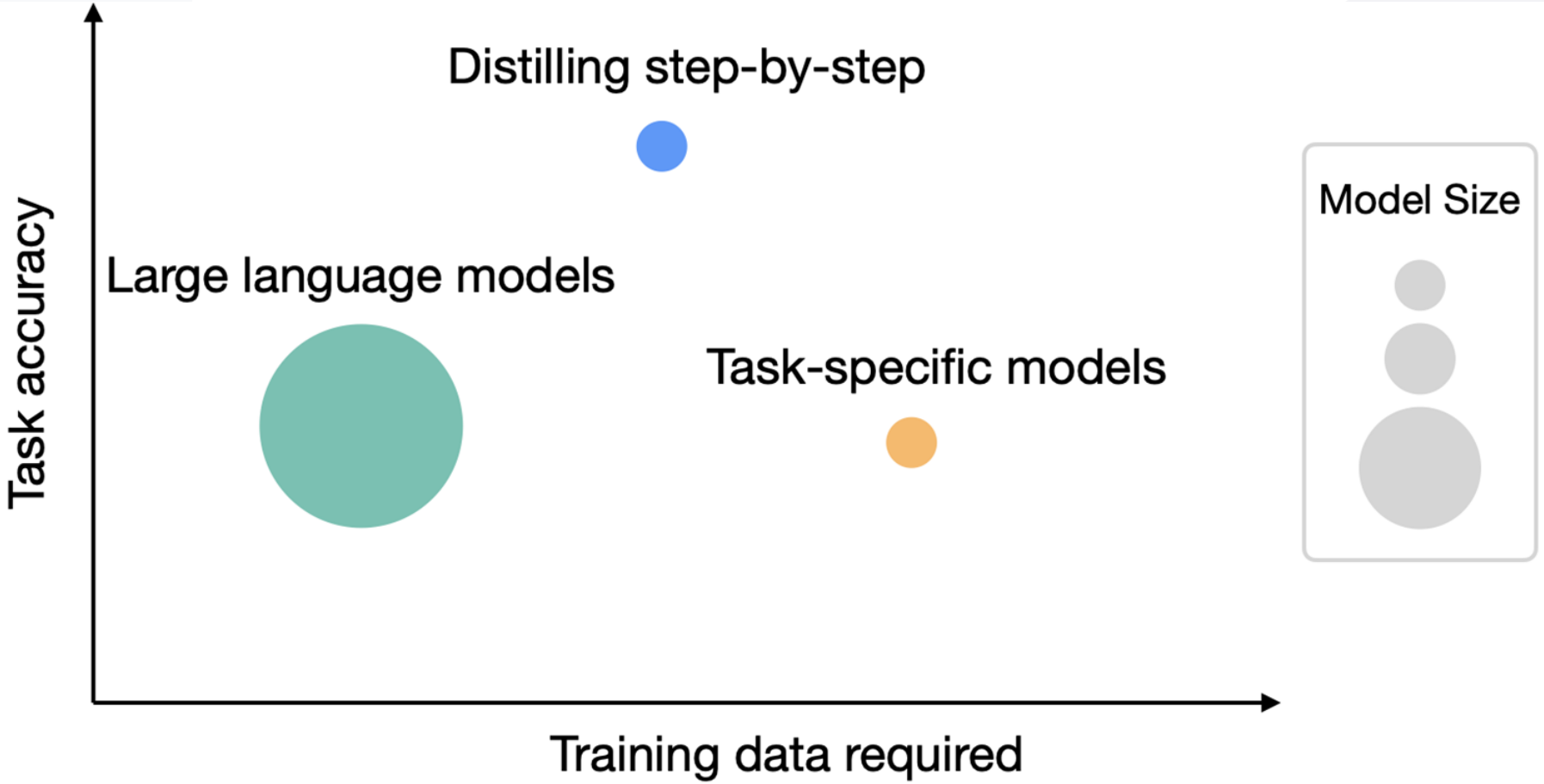
根據Google Brain、Google DeepMind、Microsoft等團隊指出純LLM RAG極限在90%



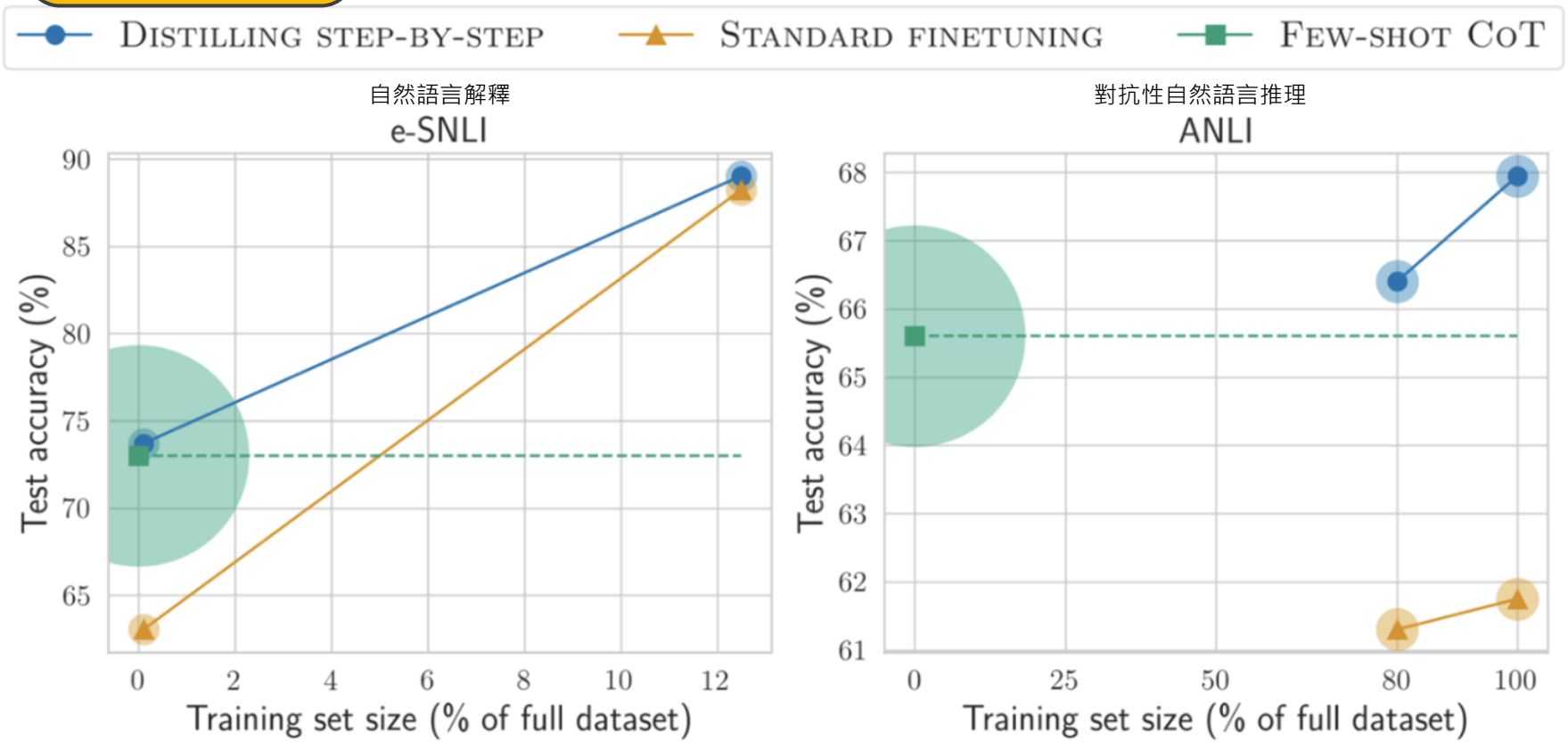
- 從大模型轉到小模型(KDFT)
- 落地後Gemma的案例估算

# Knowledge Distillation in Gemma 2





From LLM to SLM

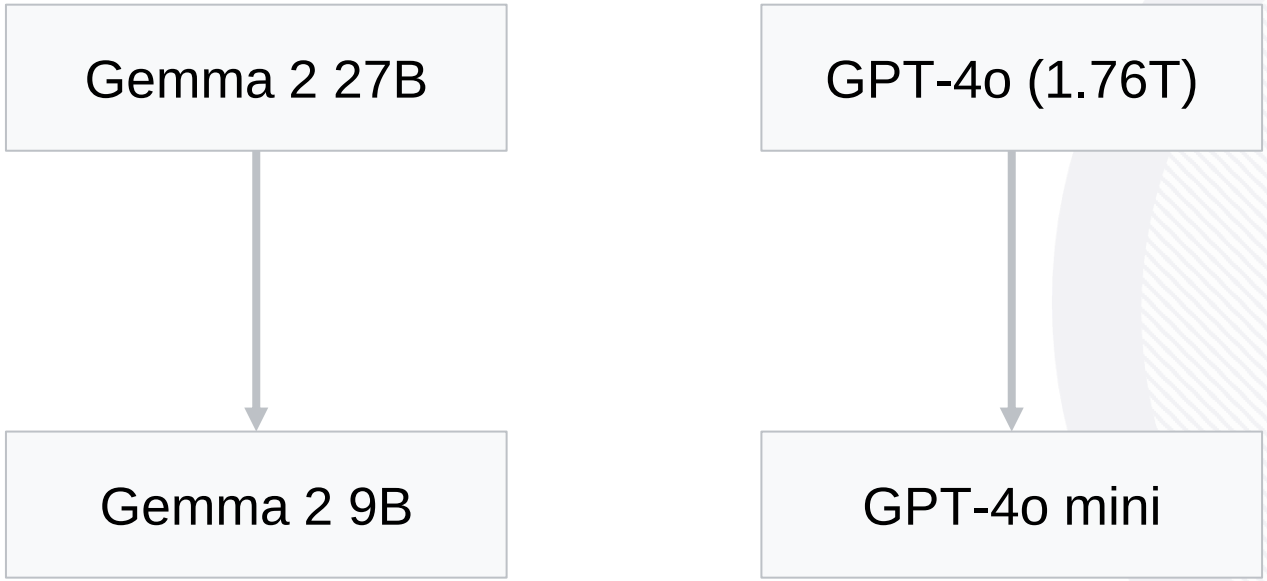



Even with a 75% reduction in data, we can still achieve a 5-10% increase in accuracy.

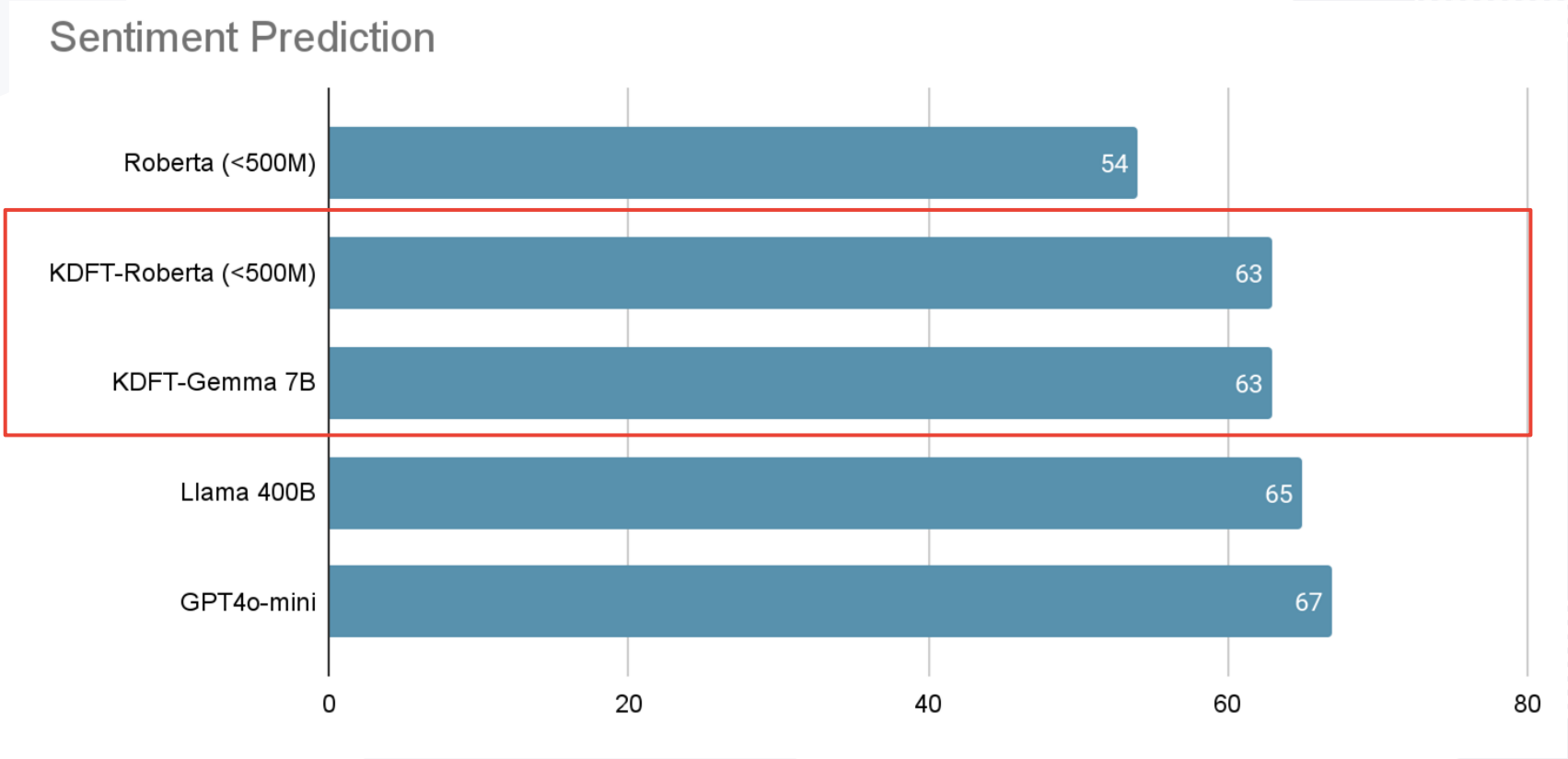
# Start big, then distillation to small

| LLM | MI350(288GB) |
| --- | --- |
| 70B(>150) | MI350 * 1 |
| 123B(>480) | MI350 * 2 |
| 340B(>640) | MI350 * 3 |
| 405B(>960) | MI350 * 4 |

## From LLM to SLM

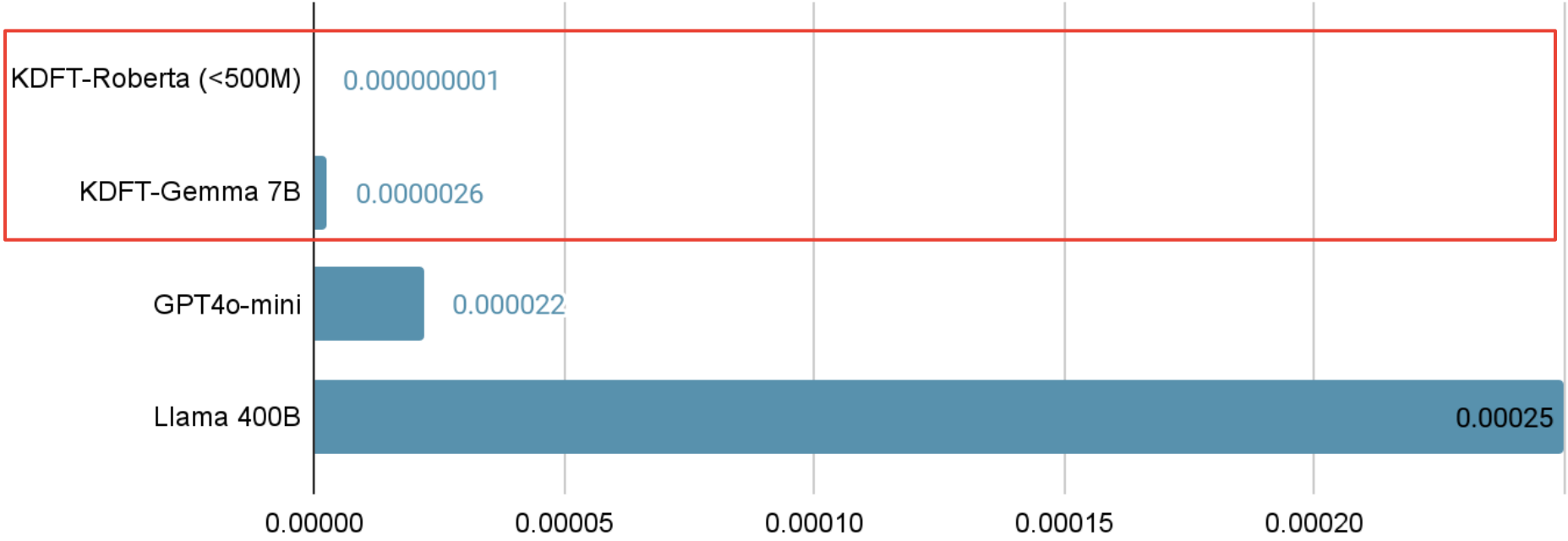


Case Study



Case Study

## Sentiment Prediction

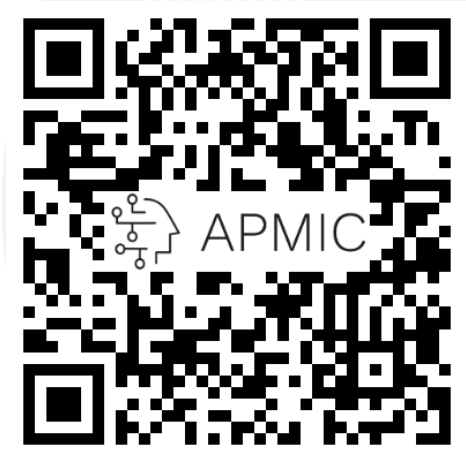


# Thank you very much !

APMIC

**Accelerate Private Machine Intelligence Company**

**Leading provider of accelerate enterprise-grade Private AI solutions**




jerry@ap-mic.com