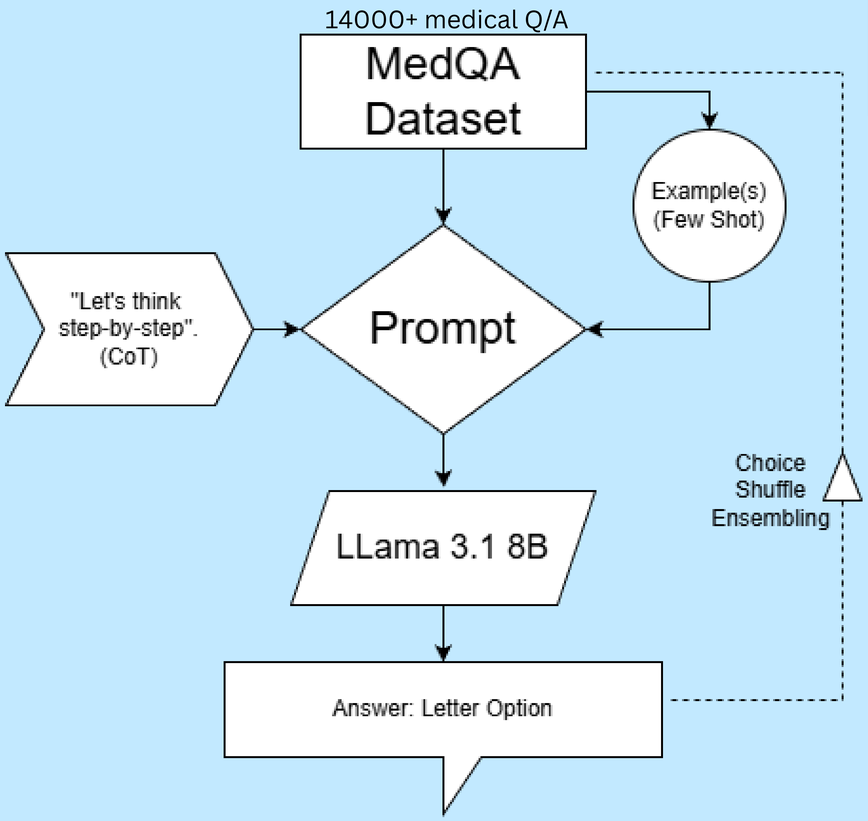
Background

MedPrompt showed that prompting techniques make general-purpose LLMs like GPT-4 excel at specialized problems without fine-tuning.
But what about smaller, cheaper models for us peasants?

Methods



- Dynamic, similarity-based example retrieval also explored, but not shown in flowchart for brevity

Extra Results

- Prompting helps get **64% accuracy** vs GPT3.5's 60.2 and **MedPaLM's 67.2%**.
- Chain-of-Thought Prompting alone gives **largest performance gain** (+18.6%).
- Small models are brittle: **examples can mislead** and degrade accuracy.

Even small language models **can rival** fine-tuned specialist models simply by using **effective prompting**.

Prompting gains for large models?

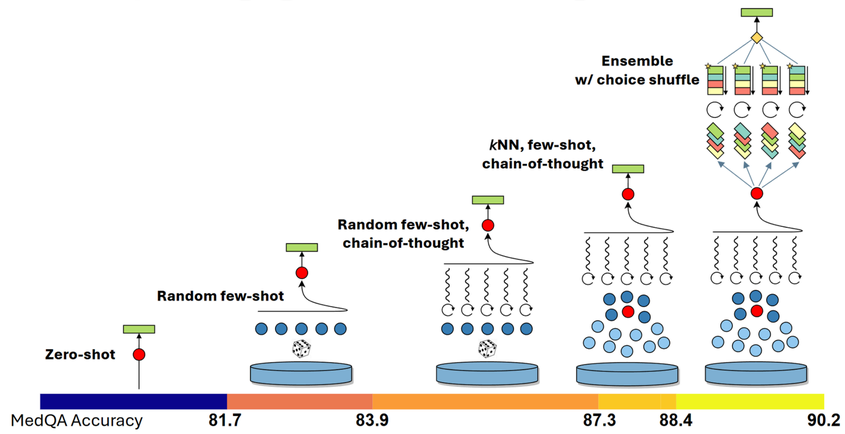


Figure from Microsoft MedPrompt paper

Prompting gains for small models?

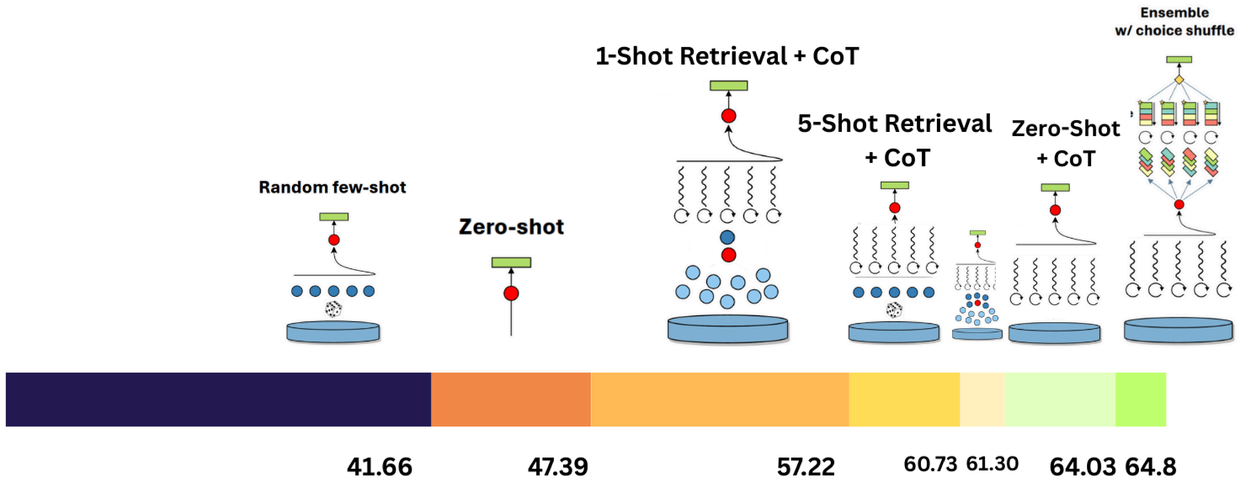


Figure representing our original work

Take a picture to read the full paper