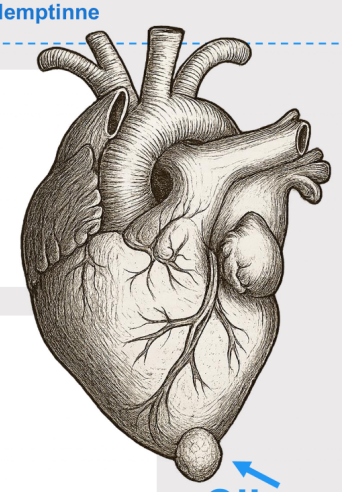
Pattern Recognition Or Medical Knowledge? The Problem With Multiple-choice Questions In Medicine

Maxime Griot, Jean Vanderdonckt, Demet Yuksel And Coralie Hemptinne

INTRODUCTION

- **Multiple choice questions** are the primary evaluation for **medical capabilities** of **Large Language Models (LLMs)**
- **LLMs** obtain **expert level** scores (>90%) on the **USMLE**
- Evaluations are used as **arguments** for **clinical implementation** of these models



Glianorex

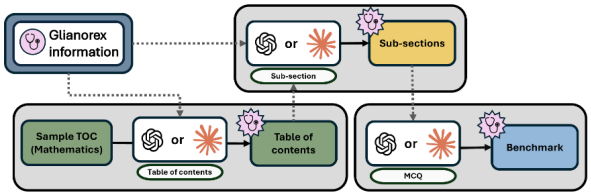
QUESTIONS

- **Are multiple-choice questions a reliable evaluation methodology?**
- **LLMs** obtain **expert level** scores on the **USMLE,** why and how?
- To what extent does the score reflect **test-taking skills** versus **content mastery**?

METHODS

Goal: Isolate test taking skills from memorization using a fictional organ

Dataset construction pipeline

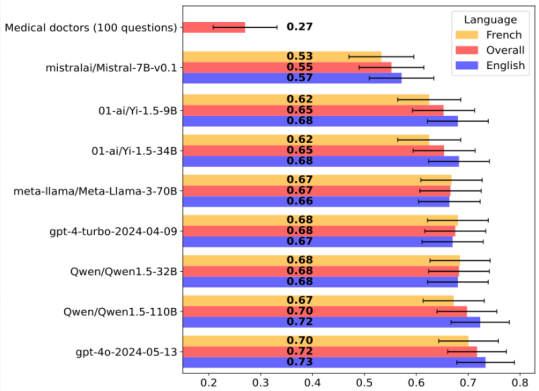


Benchmark use

- Literature on a **fictional organ**, **4 textbooks** (2 in English, 2 in French)
- **488 question-answer pairs** per language derived from the textbooks
- Baseline of **2 medical doctors**
- Evaluation in zero-shot of **15 models** including **GPT-4o**, **Claude 3.5 Sonnet**, and **DeepSeek V3 0324**
- Interpretability study comparing **zero-shot**, **chain-of-thought**, **with** and **without questions**

RESULTS

Overall performance in zero-shot



- **Models** achieve high scores averaging **64%** while **physicians** obtain only **27%**
- **Performance** on the **Glianorex** is **correlated** with performance on **MedQA-USMLE** ($R^2 = 0.5952$)
- **Performance** is correlated with **model size** and **release date**

Interpretability analysis of DeepSeek V3

	zero-shot	CoT	AO + zero-shot	AO + CoT
English	0.705	**0.717**	0.457	0.473
French	**0.697**	0.689	0.473	0.438
Average	0.701	**0.703**	0.465	0.455

- **46.5%** correct **without the question stem**
- **High agreement** between **zero-shot** and **CoT** ($\kappa = 0.681$ in English and 0.653 in French)
- Model decisions in MCQs are not based on answer length

Explicit test taking strategies

Test construction

D stands out as the most "constructed" correct answer in a medical context, resembling how hypothetical disorders are framed in exams. (While all options contain questionable terms, D is the most logically structured and aligns best with how exam questions are typically designed - tying a novel mechanism to a targeted treatment.)

The model exploits answer patterns even without access to the question

Specificity

In exams, highly specific answers ("biotransplant," "modulators," "re-equilibration") are often correct when other choices are generic.

DeepSeek picks up on unintended cues in test design

TAKE HOME

- **Test taking skills** seem to play a **much larger role** for **LLMs** compared to humans
- **Multiple Choice Questions** likely **overestimate** the actual **medical capabilities** of models
- **Alternative methodologies** should be used for "in-vitro evaluations" such as **key-features**, **open-ended tasks** using LLM-as-a-judge with expert sampling
- **Clinicians should be involved** in the development process, and clinical trials are necessary to demonstrate medical capabilities

UCLouvain

Cliniques universitaires SAINT-LUC UCLouvain BRUXELLES

ACL 2025 VIENNA

maxime.griot@uclouvain.be