Ordering Imaging Studies via Language Model Alignment with the ACR Appropriateness Criteria

Michael S. Yao, Allison Chae, Charles E. Kahn, Jr., Walter R. Witschey,
James C. Gee, Hersh Sagreiya, Osbert Bastani

Penn Medicine
Penn Engineering
University of Pennsylvania

Introduction

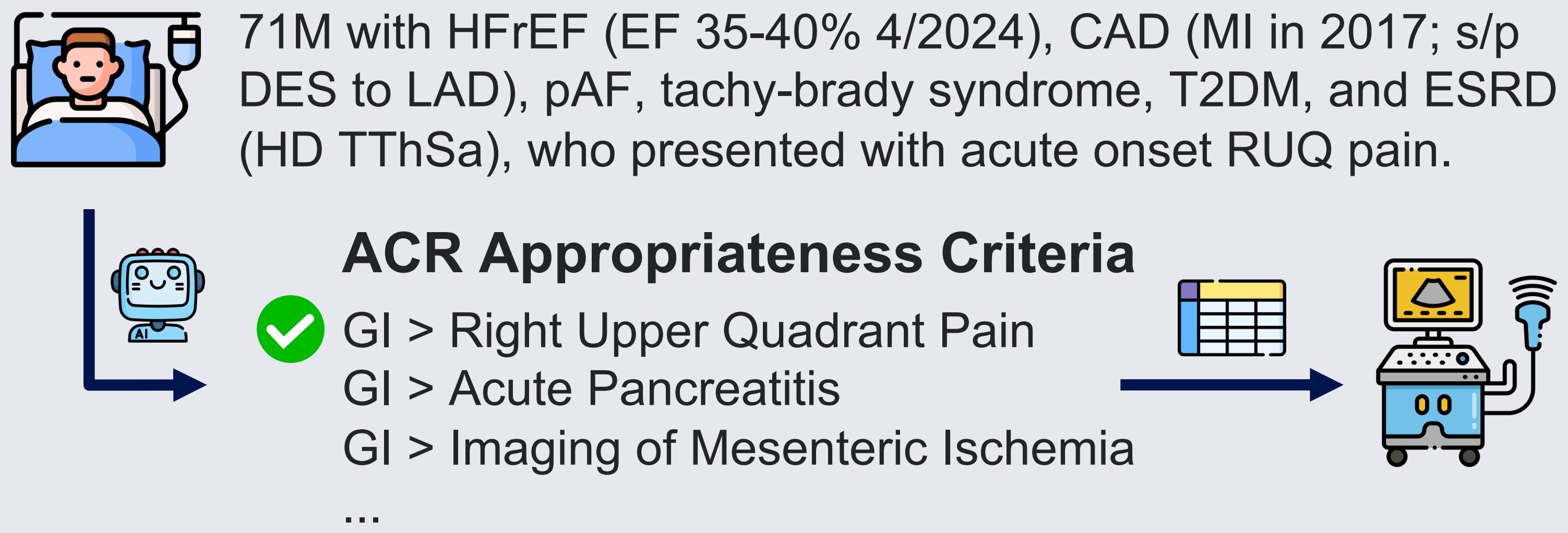
Ordering **diagnostic imaging** studies in the emergency room (ER) is challenging, and it is hard for ER doctors to order the most appropriate study for patients. This results in wasted resources, exposes patients to extra radiation, and increases healthcare costs.

To address this, the American College of Radiology (ACR) created medical guidelines called the **ACR Appropriateness Criteria (ACR AC)** to help doctors choose the right imaging tests. However, many doctors don't use these guidelines because they are complex and hundreds of pages long.

Can LLMs help ER doctors order imaging studies that are better aligned with the ACR AC guidelines?

Key Idea

Traditional LLM-based approaches directly try to predict the most appropriate imaging study given a patient description. Instead, we (1) use LLMs to predict the **most relevant ACR AC Topic**, which is a title to an ACR AC guideline. We then (2) look up the most appropriate imaging study according to the guidelines themselves (without using LLMs).



71M with HFrEF (EF 35-40% 4/2024), CAD (MI in 2017; s/p DES to LAD), pAF, tachy-brady syndrome, T2DM, and ESRD (HD TThSa), who presented with acute onset RUQ pain.

ACR Appropriateness Criteria

✓ GI > Right Upper Quadrant Pain
GI > Acute Pancreatitis
GI > Imaging of Mesenteric Ischemia
...

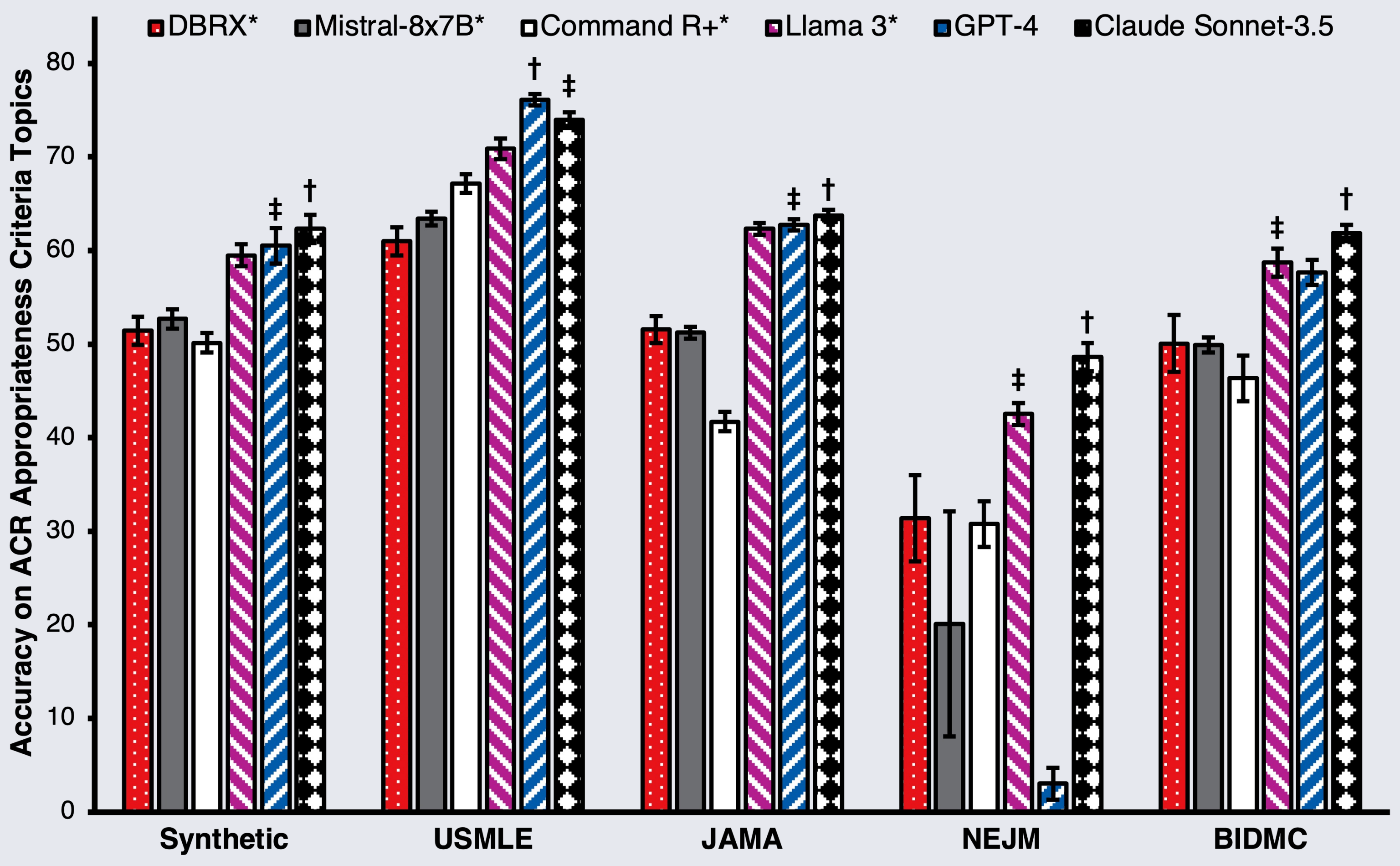
RadCases Dataset

To evaluate LLMs, we constructed and annotated a dataset of **1500+** patient "one-liners" labelled by the ACR AC Topic that is most relevant to each patient. These one-liners come from:

1. **Synthetic** cases generated by ChatGPT;
2. **USMLE** cases from prior medical board exams;
3. Patient cases published in **JAMA** and **NEJM**; and
4. Real patient cases from the **BIDMC** medical center

Which LLM is the best?

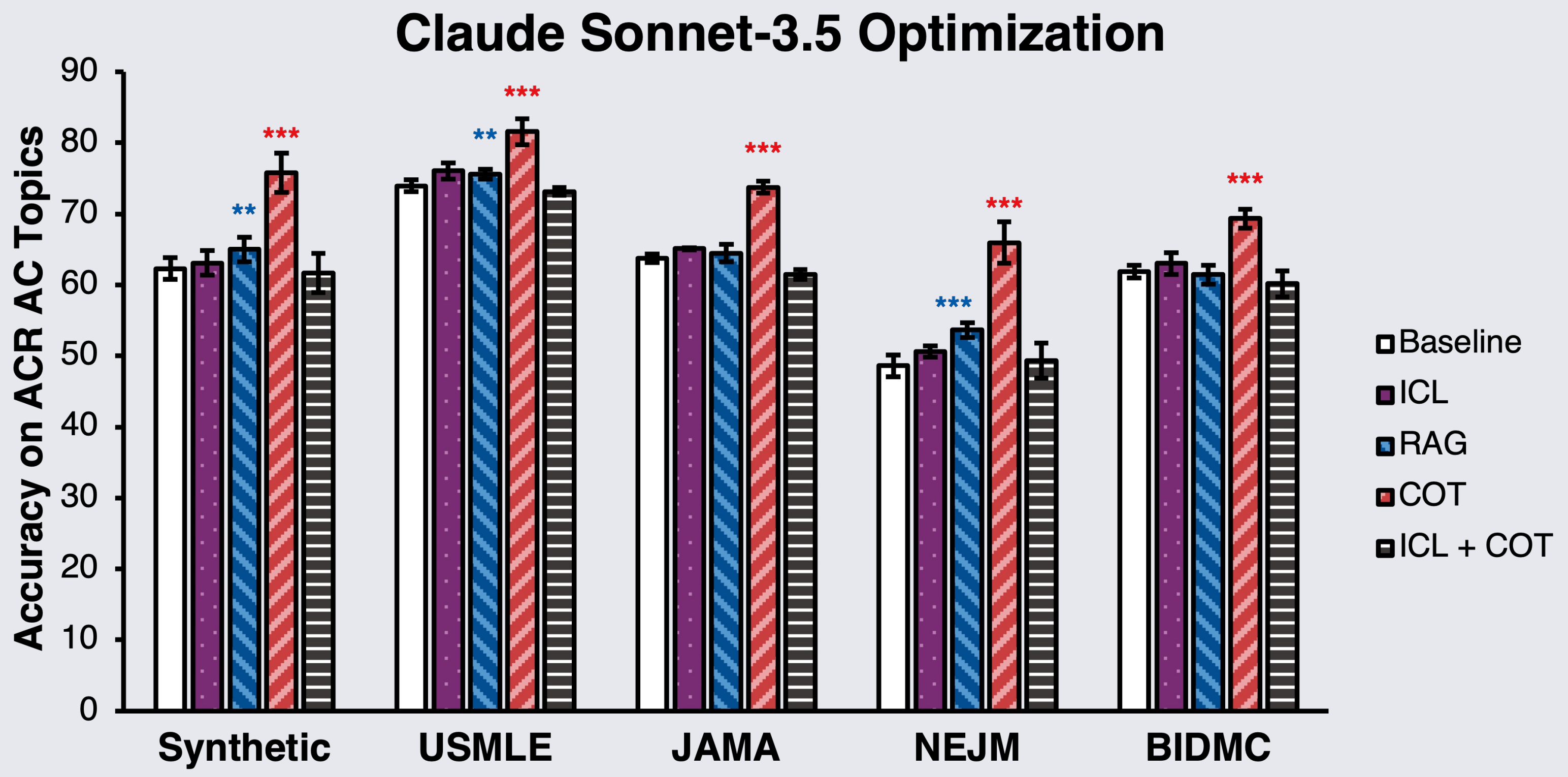
Claude Sonnet-3.5 performed the best. Meta's **Llama 3 (70B)** was the best performing open-source model.



How can we improve LLMs?

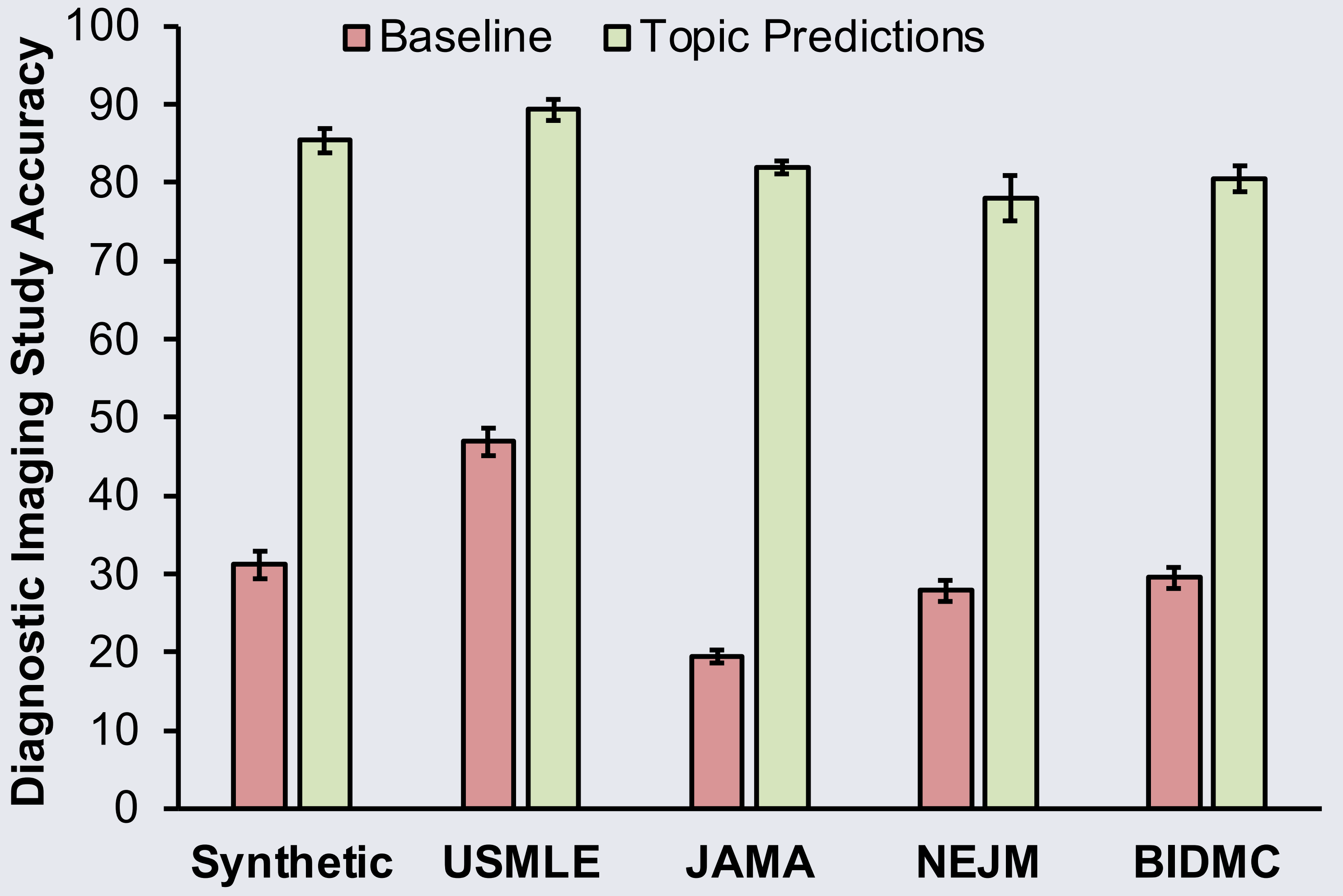
We found that **chain-of-thought (COT)** reasoning improved Claude's performance the most. In-context learning (ICL) and retrieval-augmented generation (RAG) did not offer significant improvements in model performance.

Interestingly, none of these optimization strategies (including model fine-tuning) improved the performance of Llama 3!



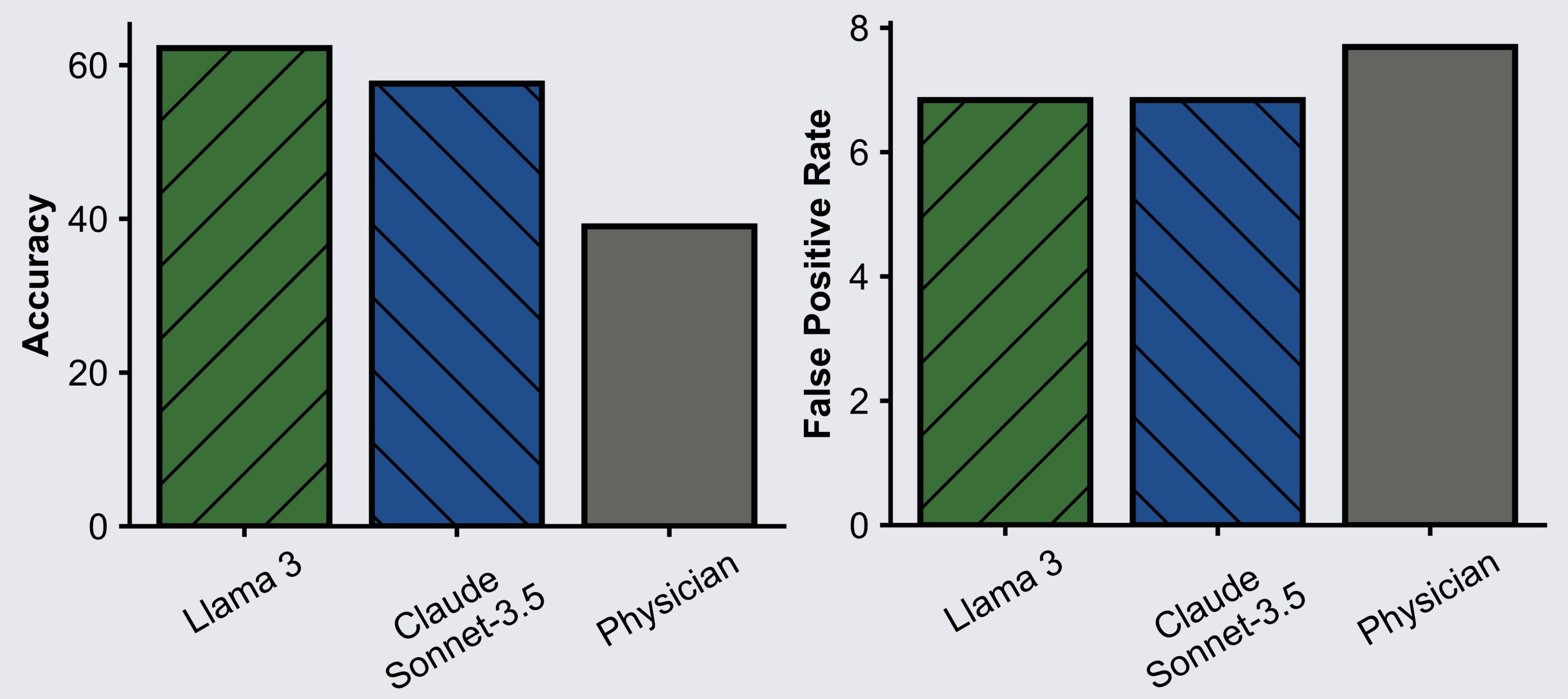
Does predicting ACR AC Topics help?

Yes! The accuracy of LLMs in predicting the most appropriate imaging study increases by up to **60%**!



How do LLMs perform in practice?

Our results showed that LLMs were **non-inferior to clinicians** in a retrospective study using real patient data.



In a prospective study with US med students and EM residents, accuracy of diagnostic image ordering significantly improved when they had access to imaging recommendations from LLMs.

In conclusion, we found that LLMs can provide clinical recommendations aligned with medical guidelines. How can we use similar techniques in other healthcare settings?