The safe use of LLMs for screening in systematic reviews

The promise of LLMs

LLMs have appeared useful for various tasks, but how can we **safely** use them to save work while screening?

They are temperamental, and they return text, not probabilites

Next token probabilities

LLMs are **probabilistic models**, which calculate the **probable next token** given a **prompt**.

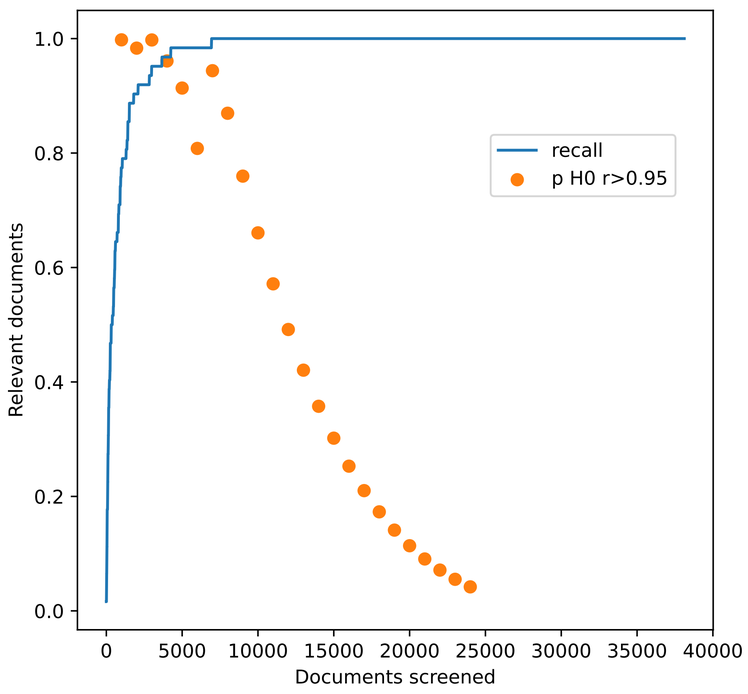
Wang et al. (2024) showed that we can ask a yes/no question and get a probability-like score by subtracting the probability that the next token is no, from the probability it is yes.

$$S(d,t) = P(yes|d,t) - P(no|d,t)$$

But the paper did not contain a way to use these scores to save work in a way that would reliably satisfy our need in systematic reviews for high recall

Prioritised screening

However, if we have probabistic scores for documents, then we can simply do prioritised screening *with* a reliable **stopping criteria** (Callaghan and Müller-Hansen, 2020)



Bibliography

Callaghan, M. and Müller-Hansen, F. (2020). Statistical Stopping Criteria for Automated Screening in Systematic Reviews. *Systematic Reviews*.

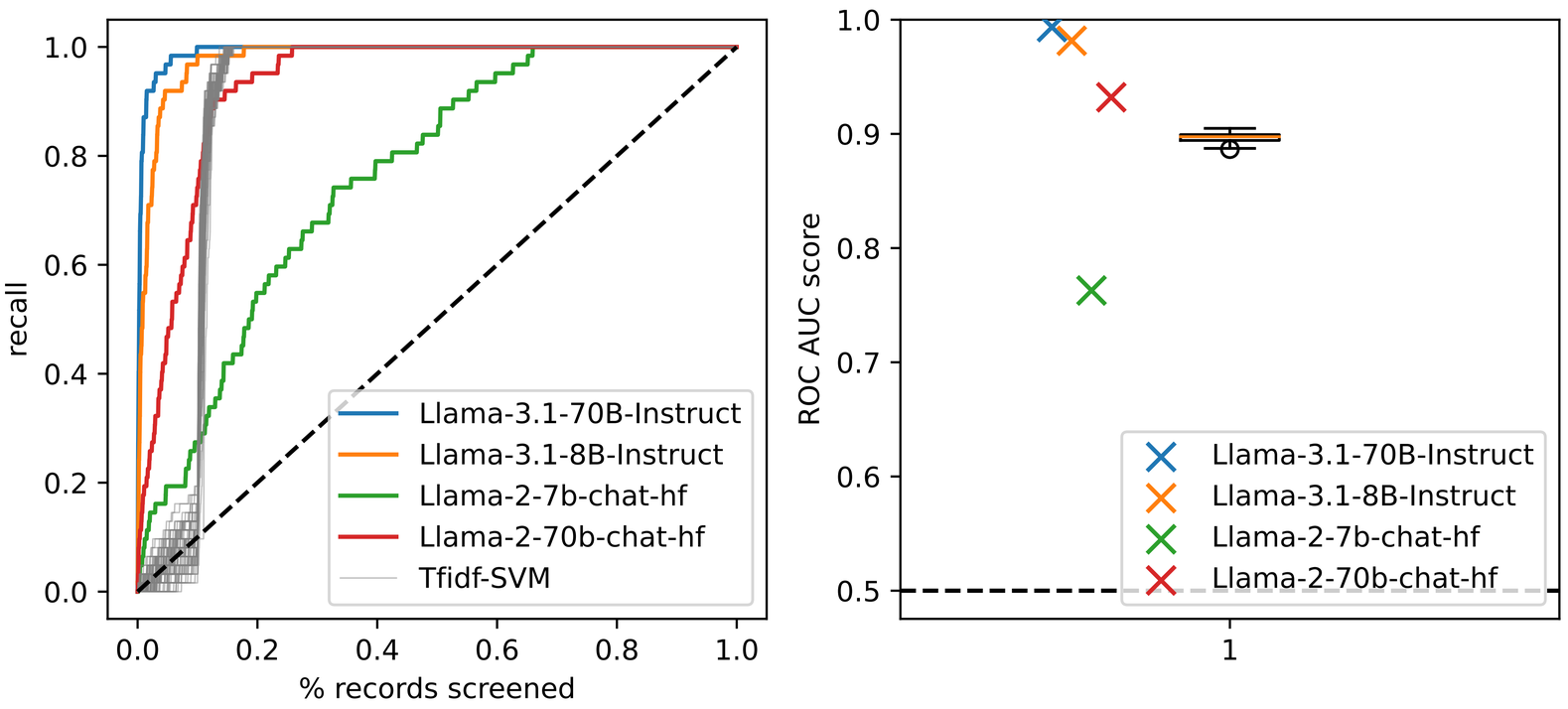
De Bruin, J., Ma, Y., Ferdinands, G., Teijema, J., and Van de Schoot, R. (2023). SYNERGY - Open machine learning dataset on study selection in systematic reviews.

Wang, S., Scells, H., Zhuang, S., Potthast, M., Koopman, B., and Zuccon, G. (2024). Zero-shot Generative Large Language Models for Systematic Review Screening Automation.

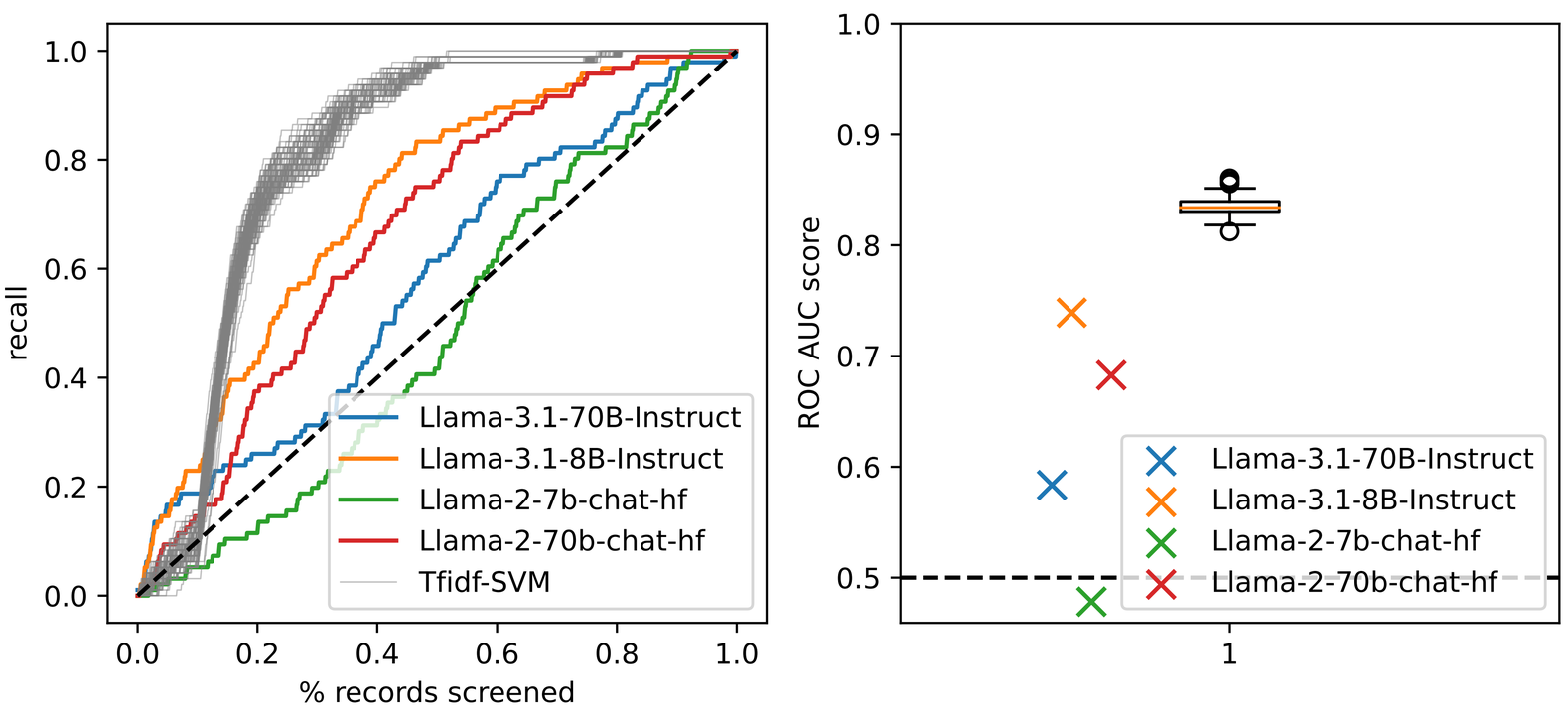
Results

Using the synergy dataset (De Bruin et al., 2023), we compared rankings from LLM screening with rankings generated in a traditional "active learning" pipeline with SVMs

Brouwer_2019 (N=38114, p=0.2%)
Psychological theories of depressive relapse and recurrence: A systematic review and meta-analysis of prospective studies



Jeyaraman_2020 (N=1175, p=8.2%)
Does the Source of Mesenchymal Stem Cell Have an Effect in the Management of Osteoarthritis of the Knee? Meta-Analysis of Randomized Controlled Trials



The best models mostly, but not always, outperform the baseline

Results

Llama 3.1 works better than 2 Wang et al. (2024), and larger models work better than smaller

Conclusion

Performance with 0 human input is impressive. Combining approaches, or using human labels to provide in-context learning could be promising.

For more info on stopping criteria see `mcallaghan.github.io/buscar-app`

Max Callaghan
callaghan@mcc-berlin.net, @MaxCallaghan5