

## LLM-prioritised screening

Max Callaghan



July 17, 2024

## We can ask LLMs whether a study should be included in a systematic review

Result: <s>[INST] «SYS» You are a systematic review helper tasked with finding out whether a study is relevant to the review Drug Class Review: Pharmacologic Treatments for Attention Deficit Hyperactivity Disorder

Answer 'yes' if the study is relevant, or 'no' if not «/SYS»

Study: Diuretics and beta-blockers do not have adverse effects at 1 year on plasma lipid and lipoprotein profiles in men with hypertension. Department of Veterans Affairs Cooperative Study Group on Antihypertensive Agents.

Concern based on the reported short-term adverse effects of antihypertensive agents on plasma lipid and lipoprotein profiles (PLPPs) has complicated the therapy for hypertension.

Should the study be included? Answer yes or no. [/INST] No, this study should not be included in the drug class review for Pharmacologic Treatments for Attention Deficit Hyperactivity Disorder. The study is focused on the effects of diuretics and beta-blockers on plasma lipid and lipoprotein profiles in men with hypertension, and does not directly relate to the treatment of Attention Deficit Hyperactivity Disorder (ADHD). The review is focused on pharmacologic treatments for ADHD, and the study does not address any of the included drug classes or outcomes of interest for the review.

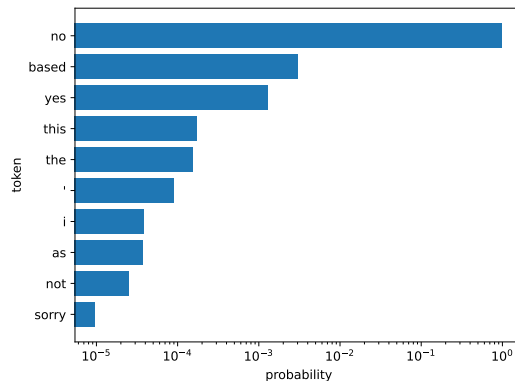
## We can ask LLMs whether a study should be included in a systematic review

Result: <s>[INST] «SYS» You are a systematic review helper tasked with finding out whether a study is relevant to the review Drug Class Review: Pharmacologic Treatments for Attention Deficit Hyperactivity Disorder

Answer 'yes' if the study is relevant, or 'no' if not «/SYS»  
Study: Diuretics and beta-blockers do not have adverse effects at 1 year on plasma lipid and lipoprotein profiles in men with hypertension. Department of Veterans Affairs Cooperative Study Group on Antihypertensive Agents.

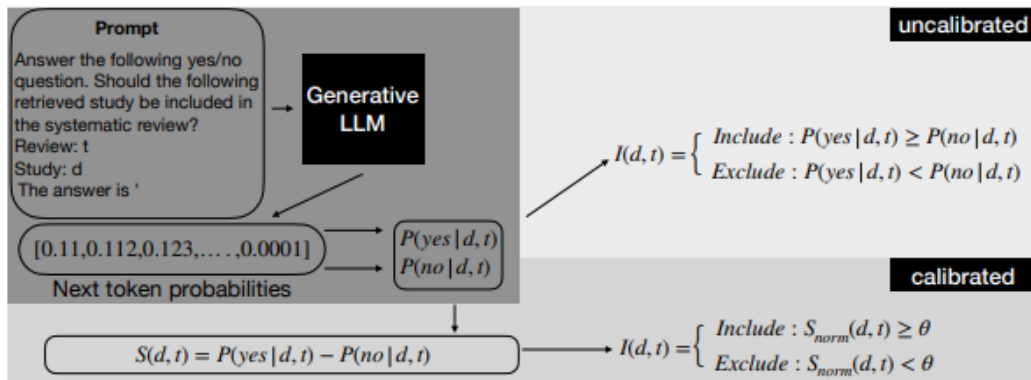
Concern based on the reported short-term adverse effects of antihypertensive agents on plasma lipid and lipoprotein profiles (PLPPs) has complicated the therapy for hypertension.

Should the study be included? Answer yes or no. [/INST] No, this study should not be included in the drug class review for Pharmacologic Treatments for Attention Deficit Hyperactivity Disorder. The study is focused on the effects of diuretics and beta-blockers on plasma lipid and lipoprotein profiles in men with hypertension, and does not directly relate to the treatment of Attention Deficit Hyperactivity Disorder (ADHD). The review is focused on pharmacologic treatments for ADHD, and the study does not address any of the included drug classes or outcomes of interest for the review.



# Zero-shot Generative Large Language Models for Systematic Review Screening Automation

? propose a method to extract probability-like scores from LLMs for inclusion/exclusion decisions in a systematic review



Results are mixed, because the probabilities are poorly calibrated, and their attempts to re-invent stopping criteria do not makes sense

	Model	Setting	P	R	B-AC	F3	Suc	WSS
CLEF-2017	BioBERT	Unc	0.06	0.95	0.61	0.30	0.74	0.26
		Cal(0.95)	0.06	0.92	0.64	0.31	0.50*	0.34*
		Cal(1)	0.06	0.97	0.60	0.29	0.82	0.23
	7b-ins	Unc	0.08	0.87	0.72	0.35	0.26	0.56
		Cal(0.95)	0.06*	0.92*	0.69*	0.32	0.52	0.44
		Cal(1)	0.05*	<b>0.99*</b>	0.60*	0.28	0.96	0.20
	13b-ins	Unc	0.19	0.41	0.66	0.31	0.04	0.91
		Cal(0.95)	0.06*	0.93	0.59*	0.28	0.50*	0.25*
		Cal(1)	0.05*	0.98	0.53*	0.26	0.88*	0.08*
	Ensemb	Unc	<b>0.31</b>	0.13	0.56	0.13	0.00	<b>0.98</b>
		Cal(0.95)	0.08	0.93*	<b>0.72</b>	<b>0.35*</b>	0.52*	0.50*
		Cal(1)	0.06	0.97*	0.63	0.30	<b>0.90*</b>	0.29*
CLEF-2018	BioBERT	Unc	0.06	0.97	0.59	0.29	0.87	0.19
		Cal(0.95)	0.07	0.91*	0.63	0.30	0.57*	0.33*
		Cal(1)	0.06	0.97	0.59	0.29	0.87	0.21
	7b-ins	Unc	0.09	0.88	0.75	0.37	0.27	0.59
		Cal(0.95)	0.08*	0.94*	0.71*	0.35*	0.50	0.46
		Cal(1)	0.06*	<b>0.99*</b>	0.62*	0.30	<b>1.00</b>	0.24
	13b-ins	Unc	0.26	0.36	0.66	0.30	0.00	0.94
		Cal(0.95)	0.06	0.94*	0.59*	0.29	0.47*	0.22*
		Cal(1)	0.05	0.97	0.53*	0.27	0.80*	0.08*

## Implementation is relatively straightforward

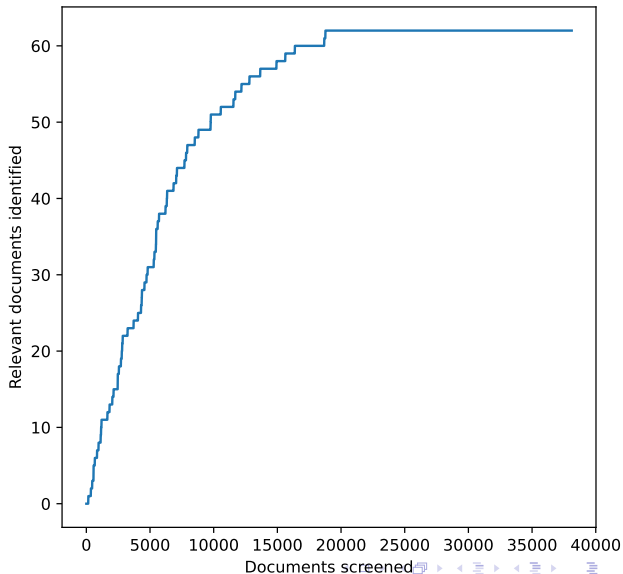
```
1 def binary_probs(tokenizer, model, prompt, no_words=['no'], yes_words=['yes'], return_all=False):
2     device = 'cuda' if torch.cuda.is_available() else 'cpu'
3     encoded_text = tokenizer(prompt, return_tensors="pt").to(device)
4     #1. step to get the logits of the next token
5     with torch.inference_mode():
6         outputs = model(**encoded_text)
7
8     next_token_logits = outputs.logits[0, -1, :]
9
10    # 2. step to convert the logits to probabilities
11    next_token_probs = torch.softmax(next_token_logits, -1)
12
13    topk_next_tokens= torch.topk(next_token_probs, 50)
14    tokens = [tokenizer.decode(x).strip().lower() for x in topk_next_tokens.indices]
15    p = topk_next_tokens.values
16
17    df = pd.DataFrame.from_dict({'t': tokens, 'p': p.cpu()})
18    y = df[df['t'].isin(yes_words)]['p'].sum()
19    n = df[df['t'].isin(no_words)]['p'].sum()
20
21    if return_all:
22        return df.groupby('t').sum().reset_index().sort_values('p', ascending=False).
23        reset_index(drop=True)
24    return y-n, y+n
```

## Implementation is relatively straightforward

```
1 prompt = Template('''<s>[INST] <<SYS>>
2 You are a systematic review helper tasked with finding out whether a study is relevant to
   the review $t
3
4 Answer 'yes' if the study is relevant, or 'no' if not
5 <</SYS>>
6
7 Study: $s
8
9 Should the study be included? Answer yes or no. [/INST] ''')
10
11 prompt.substitute({'t': review, 's': study_title}),
```

## Result #1: It works!

- The method produces a ranking that identifies all relevant documents before all documents have been screened

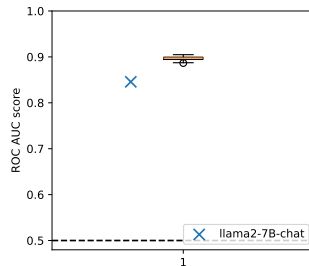
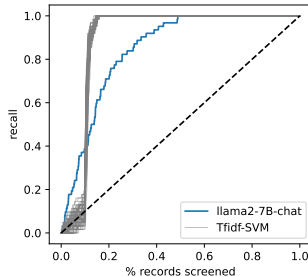




## Result #1: It works!

- The method produces a ranking that identifies all relevant documents before all documents have been screened
- But it doesn't perform as well as simple tfidf + SVM

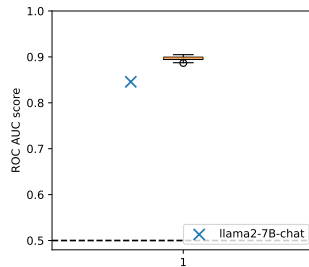
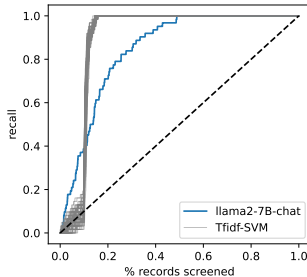
Brouwer\_2019 (N=38114, p=0.2%)  
Psychological theories of depressive relapse and recurrence: A systematic review and meta-analysis of prospective studies



## Result #1: It works!

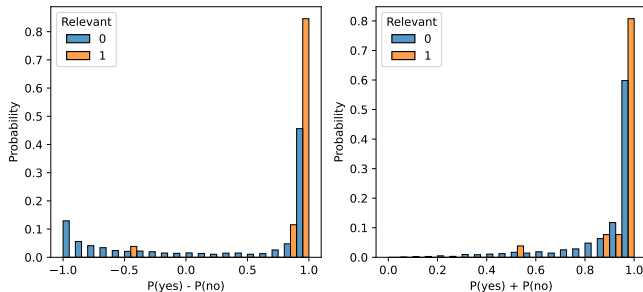
- The method produces a ranking that identifies all relevant documents before all documents have been screened
- But it doesn't perform as well as simple tfidf + SVM
- To compare the quality of rankings, we can use ROC AUC scores

Brouwer\_2019 (N=38114, p=0.2%)  
Psychological theories of depressive relapse and recurrence: A systematic review and meta-analysis of prospective studies



This method simply produces a ranking of documents, but is this ranking better than what we can produce with traditional active learning approaches?

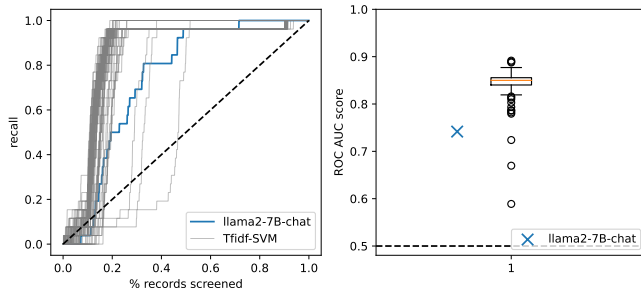
Appenzeller-Herzog\_2019 (N=2873, p=0.9%)  
Comparative effectiveness of common therapies for Wilson disease: A systematic review and meta-analysis of controlled studies



- We generated LLM inclusion scores for each document in a set of reviews in the Synergy dataset

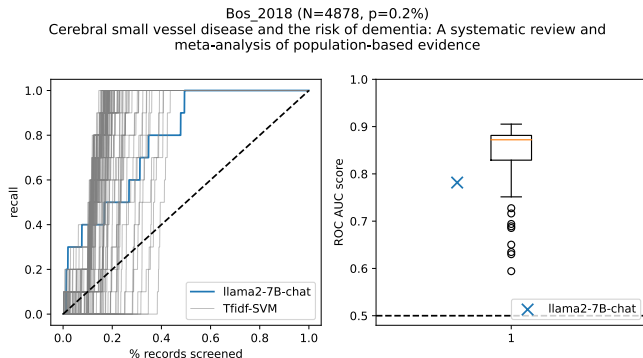
This method simply produces a ranking of documents, but is this ranking better than what we can produce with traditional active learning approaches?

Appenzeller-Herzog\_2019 (N=2873, p=0.9%)  
Comparative effectiveness of common therapies for Wilson disease: A systematic review and meta-analysis of controlled studies



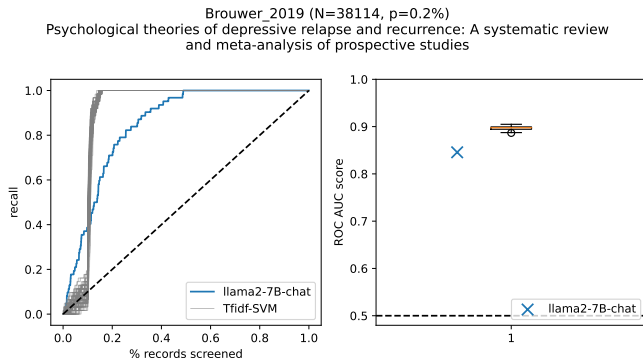
- We generated LLM inclusion scores for each document in a set of reviews in the Synergy dataset
- We compare what happens to screening in descending order of this ranking to 100 active learning runs with a default configuration of SVM with Tfidf

This method simply produces a ranking of documents, but is this ranking better than what we can produce with traditional active learning approaches?



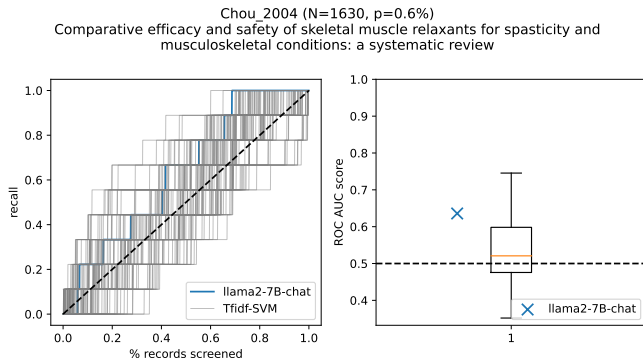
- We generated LLM inclusion scores for each document in a set of reviews in the Synergy dataset
- We compare what happens to screening in descending order of this ranking to 100 active learning runs with a default configuration of SVM with Tfidf
- Initial results are not very promising

This method simply produces a ranking of documents, but is this ranking better than what we can produce with traditional active learning approaches?



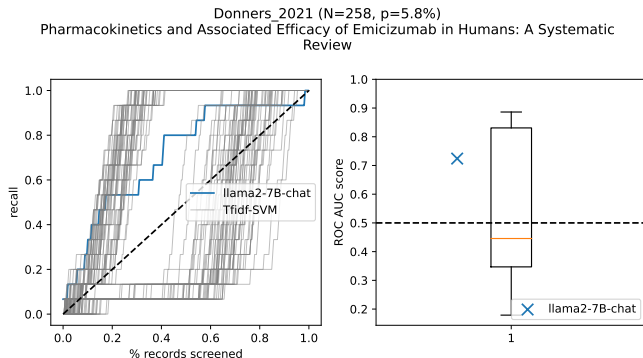
- We generated LLM inclusion scores for each document in a set of reviews in the Synergy dataset
- We compare what happens to screening in descending order of this ranking to 100 active learning runs with a default configuration of SVM with Tfidf
- Initial results are not very promising

This method simply produces a ranking of documents, but is this ranking better than what we can produce with traditional active learning approaches?



- We generated LLM inclusion scores for each document in a set of reviews in the Synergy dataset
- We compare what happens to screening in descending order of this ranking to 100 active learning runs with a default configuration of SVM with Tfidf
- Initial results are not very promising
- But LLMs can do well where traditional methods struggle

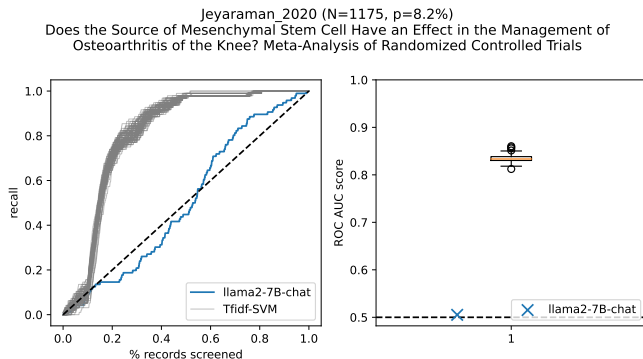
This method simply produces a ranking of documents, but is this ranking better than what we can produce with traditional active learning approaches?



- We generated LLM inclusion scores for each document in a set of reviews in the Synergy dataset
- We compare what happens to screening in descending order of this ranking to 100 active learning runs with a default configuration of SVM with Tfidf
- Initial results are not very promising
- But LLMs can do well where traditional methods struggle

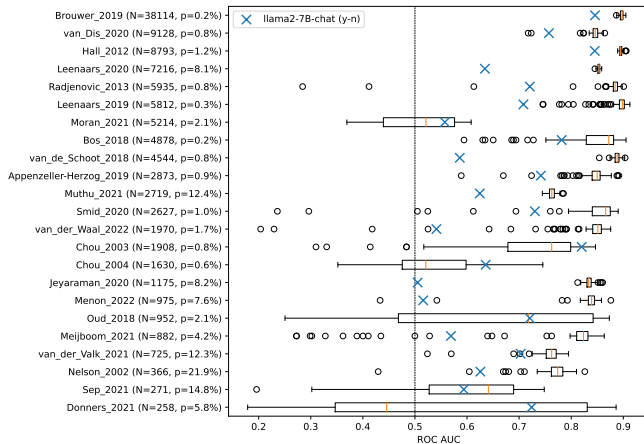


This method simply produces a ranking of documents, but is this ranking better than what we can produce with traditional active learning approaches?



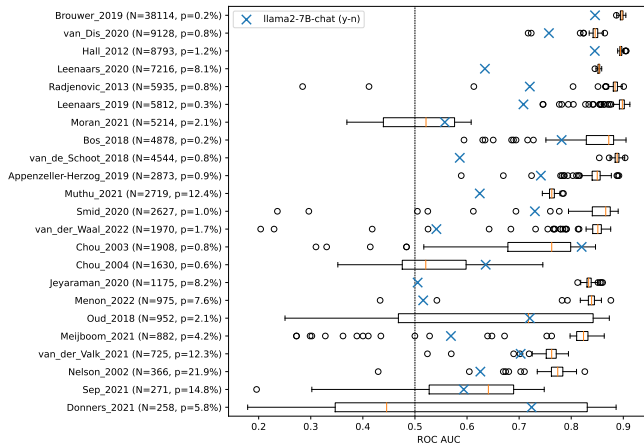
- We generated LLM inclusion scores for each document in a set of reviews in the Synergy dataset
- We compare what happens to screening in descending order of this ranking to 100 active learning runs with a default configuration of SVM with Tfidf
- Initial results are not very promising
- But LLMs can do well where traditional methods struggle
- Sometimes they are no use at all

## All results



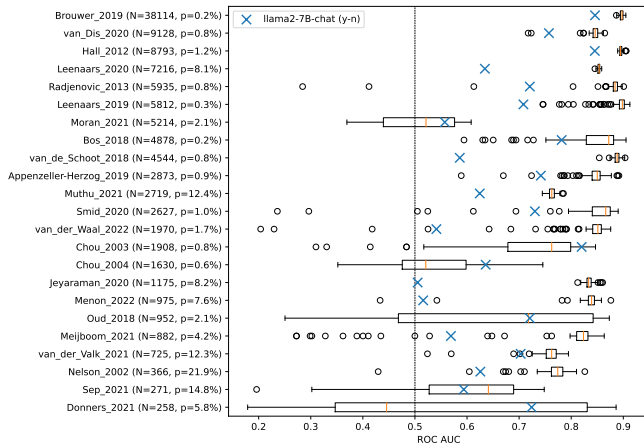
- LLMs and basic active learning pipelines seem to have different weaknesses

# All results



- LLMs and basic active learning pipelines seem to have different weaknesses
- Combining both could improve general performance

## All results



- LLMs and basic active learning pipelines seem to have different weaknesses
- Combining both could improve general performance
- LLMs seem most useful for smaller datasets (where active learning has little time to learn)

## Some things still to explore

- Prompting strategies (inclusion criteria)

## Some things still to explore

- Prompting strategies (inclusion criteria)
- Bigger/different models

## Some things still to explore

- Prompting strategies (inclusion criteria)
- Bigger/different models
- Combining LLMs with traditional approaches

## Some things still to explore

- Prompting strategies (inclusion criteria)
- Bigger/different models
- Combining LLMs with traditional approaches
- Updating prompts based on user feedback



# Conclusions

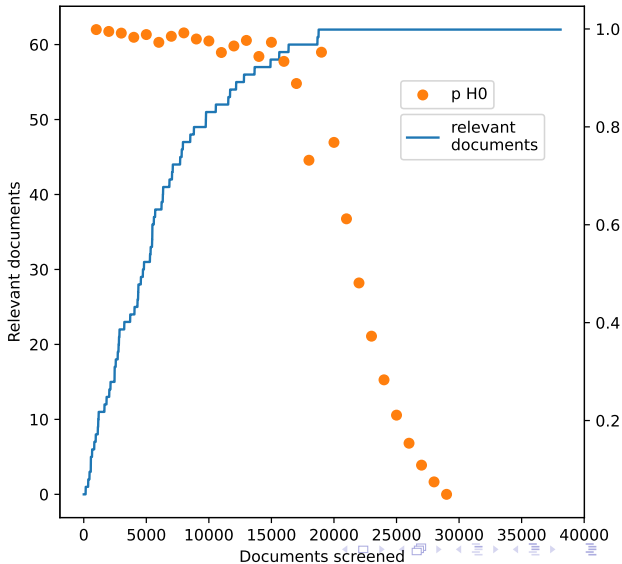
- LLMs are neither a quick fix or a silver bullet

# Conclusions

- LLMs are neither a quick fix or a silver bullet
- Evaluation is vital

## Conclusions

- LLMs are neither a quick fix or a silver bullet
- Evaluation is vital
- We can't forget the need for appropriate stopping criteria ?



## Conclusions

- LLMs are neither a quick fix or a silver bullet
- Evaluation is vital
- We can't forget the need for appropriate stopping criteria ?
- [mcallaghan.github.io/buscar-app](https://mcallaghan.github.io/buscar-app)

Thanks!

---

callaghan@mcc-berlin.net

Twitter: @MaxCallaghan5



