

Perspectives on Large Language Models for Relevance Judgment

Guglielmo Faggioli, Laura Dietz, **Charles Clarke**, Gianluca Demartini,
Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas,
Martin Potthast, Benno Stein, Henning Wachsmuth

with thanks to
Christine Bauer, Ben Carterette, Nicola Ferro, and Norbert Fuhr

and with special thanks to
Ian Soboroff

Dagstuhl Seminar 23031

Frontiers of Information Access Experimentation for Research and Education

January 15–20, 2023



This Dagstuhl seminar addressed “technology-enhanced information access and specifically focus on developing more responsible experimental practices leading to more valid results, both for research as well as scientific education.”

We are a subset of these people, who were interested in: The impact of Large Language Models (LLMs) on Information Retrieval evaluation.

We asked ChatGPT if it could help with IR evaluation



Dear LLM, can you assist us with relevance judgments?



Yes, I can certainly provide assistance with relevance judgments. Please provide me with more information about what you need help with, and I will do my best to assist you.

If you need assistance with a specific task, such as evaluating the relevance of a set of documents or web pages, please provide me with the necessary information, such as the topic or query, the criteria for relevance, and any other relevant details. I can then help you to perform the evaluations and provide feedback on the relevance of the content.

Are we on the verge of being able to delegate relevance judgments to machines—either fully or partially—by employing large language models?

Reminder: The Cranfield Paradigm (Cyril Cleverdon)

The Cranfield paradigm depends on a test collection that consists of:

1. A document corpus
2. A set of information needs or topics
3. **Human** relevance judgments for documents on the topics



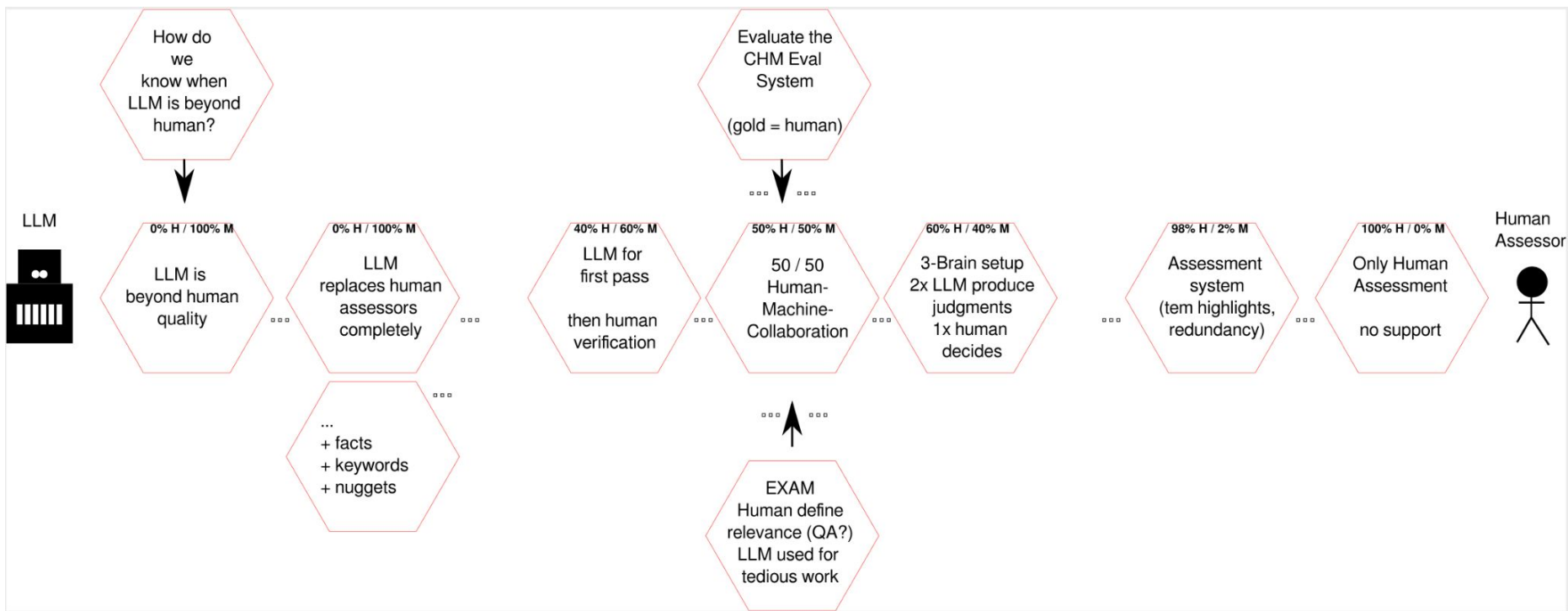
Humans are a critical (but expensive) component of the Cranfield Paradigm.

For example, for the 50 topics in the TREC-8 Ad Hoc track, 129 participating systems led to more than 86,000 pooled documents to judge, requiring more than 700 **human** assessor hours at a cost of about USD 15,000.

We considered past attempts to reduce the cost of relevance assessment



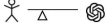

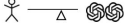
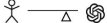


- Judging text nuggets instead of documents
- Employing crowdworkers
- Cleverly selecting which documents to judge
- Constructing test collections from Wikipedia
- Classifiers and active learning
- ...

We constructed a spectrum of human - machine collaboration



We constructed a spectrum of human - machine collaboration

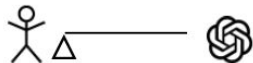
~~machine~~
LLM

Collaboration Integration	Task Organization
<hr/>	
Human Judgment	
	Humans do all judgments manually without any kind of support.
	Humans have full control of judging but are supported by text highlighting, document clustering, etc.
AI Assistance	
	Humans judge documents while having access to LLM-generated summaries.
	Balanced competence partitioning. Humans and LLMs focus on (sub-)tasks they are good at.
Human Verification	
	Two LLMs each generate a judgment, and humans select the better one.
	An LLM produces a judgment (and an explanation) that humans can accept or reject.
	LLMs are considered crowdworkers with varied specific characteristics, but supervised / controlled by humans.
Fully Automated	
	Fully automatic judgments.

Collaboration Integration

Task Organization

Human Judgment



Humans do all judgments manually without any kind of support.



Humans have full control of judging but are supported by text highlighting, document clustering, etc.

AI Assistance



Humans judge documents while having access to LLM-generated summaries.



Balanced competence partitioning. Humans and LLMs focus on (sub-)tasks they are good at.

Human Verification

clustering, etc.

AI Assistance

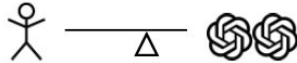


Humans judge documents while having access to LLM-generated summaries.

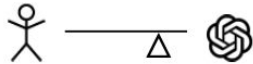


Balanced competence partitioning. Humans and LLMs focus on (sub-)tasks they are good at.

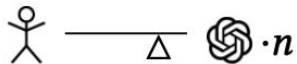
Human Verification



Two LLMs each generate a judgment, and humans select the better one.



An LLM produces a judgment (and an explanation) that humans can accept or reject.



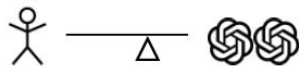
LLMs are considered crowdworkers with varied specific characteristics, but supervised / controlled by humans.

Fully Automated

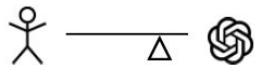


Balanced competence partitioning. Humans and LLMs focus on (sub-)tasks they are good at.

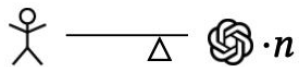
Human Verification



Two LLMs each generate a judgment, and humans select the better one.

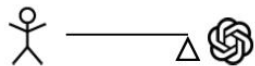


An LLM produces a judgment (and an explanation) that humans can accept or reject.



LLMs are considered crowdworkers with varied specific characteristics, but supervised / controlled by humans.

Fully Automated




Fully automatic judgments.

Issues and opportunities

1. LLM Judgment Cost and Quality
2. Human Verification
 - a. Using Multiple LLMs as Assessors
 - b. Truthfulness & Misinformation
 - c. Bias
 - d. Faithful Reasoning
 - e. Explaining Relevance to LLMs
3. Circularity of LLM-based Evaluation of LLM-based Systems
4. Beyond humanity



Type of Assessor	Cost	Type of Judgment			
		Preference	Binary	Graded	Explained
User	free	⊕	⊕	⊕	⊙
Expert	expensive	⊕⊕	⊕⊕	⊕	⊕
Crowdworker	cheap	⊙	⊕	⊕	⊙
LLM 	very cheap	⊕	⊕	⊙	⊕

Legend: ⊕⊕ can judge, ⊕ depends, ⊙ unknown

We ran some experiments

- Collections:
 - TREC 2021 Deep Learning track
 - TREC-8 adhoc track
- LLM's:
 - GPT-3.5 text-davinci-003 accessed through API
 - ChatGPT(-3) accessed manually through web interface
 - YouChat accessed manually through web interface
- Approaches:
 - Small numbers of query-document pairs entered through web interface
 - Larger numbers of query-document pairs entered through API
- Judgment types:
 - Binary
 - 4-point graded scale

We ran some experiments - TREC-8 adhoc track

Table 3: Judgment agreement on TREC-8 between TREC assessors and the LLMs; 1000 topic–document pairs for GPT-3.5 and 100 for each grade (relevant, non-relevant) for YouChat.

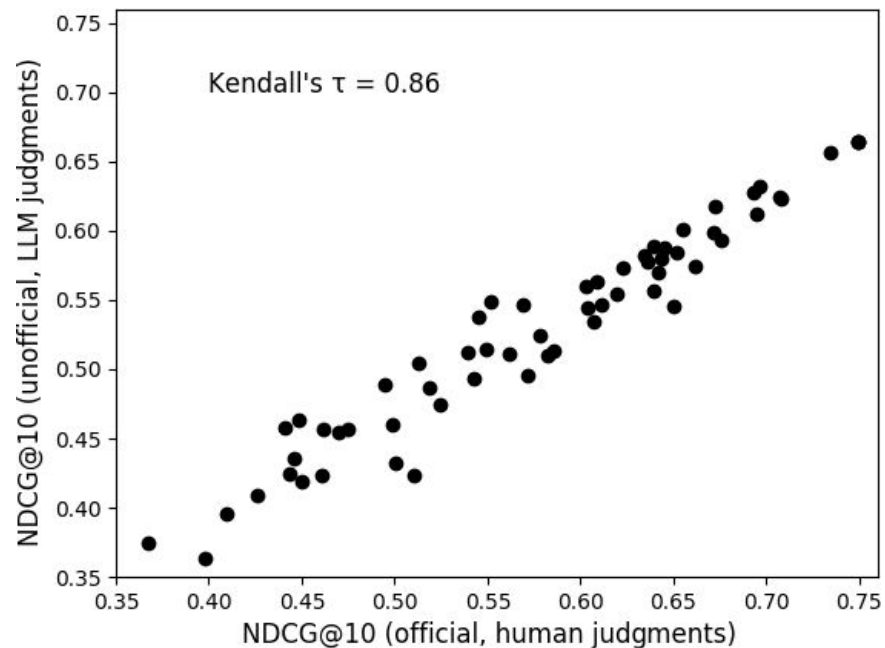
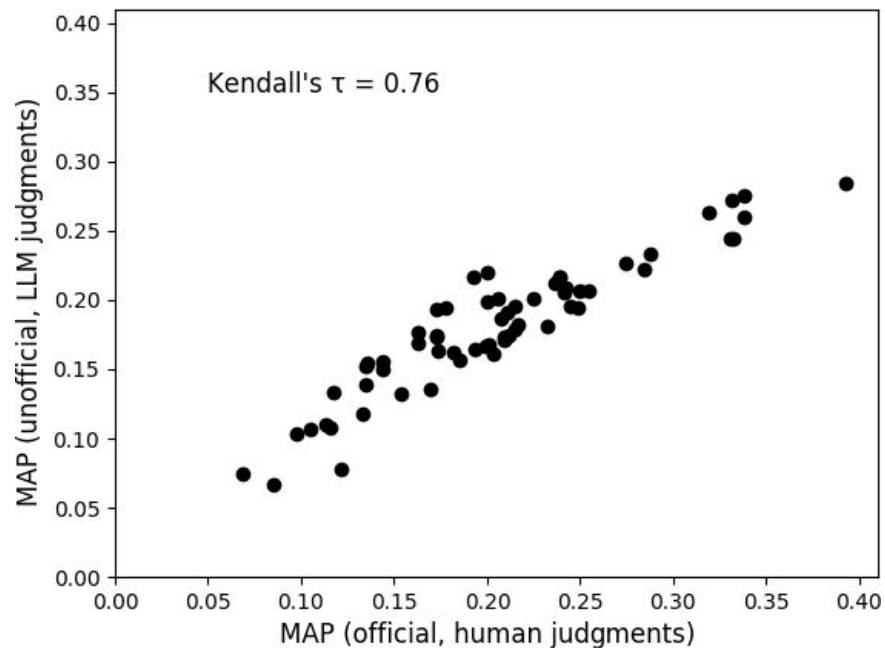
LLM	Prediction	TREC-8 Assessors		Cohen’s κ
		Relevant	Non-relevant	
GPT-3.5	Relevant	237	48	0.38
	Non-relevant	263	452	
YouChat	Relevant	33	26	0.07
	Non-relevant	67	74	

We ran some experiments - TREC 2021 Deep Learning

Table 4: Judgment agreement on TREC-DL 2021 between TREC assessors and the LLMs; 100 question–passage pairs for each grade from 3 (highly relevant) to 0 (non-relevant).

LLM	Prediction	TREC-DL 2021 Assessors				Cohen's κ
		3	2	1	0	
GPT-3.5	Relevant	89	65	48	16	0.40
	Non-relevant	11	35	52	84	
YouChat	Relevant	96	93	79	42	0.49
	Non-relevant	4	7	21	58	

We ran some experiments - TREC 2021 Deep Learning



We ran some experiments- Some observations

- Human assessors may use subtle details to distinguish ‘somewhat relevant’ from ‘probably non-relevant documents’ in the binary case that are not captured by the LLMs.
- Cohen’s κ indicates ‘fair’ level of agreement on LLM rejudging of TREC Deep Learning 2021.
- Measures computed under the LLM TREC Deep Learning judgments are less sensitive than measures computed under human judgments.

Three perspectives emerged from our discussion

1. In Favor of Using LLMs for Judgments
2. Against Using LLMs for Judgments
3. A Compromise

In Favor of Using LLMs for Judgments



- **Scalable.** With no boredom or tiredness. As many judgments as you want, whenever you want them, at a relatively low cost.
 - **Multilingual.** Capable of code switching.
 - **Consistent?** May assess conceptually similar documents the same way?
 - **Explainable.** LLMs are able to produce a natural language explanation why a certain document is relevant or not to a topic.
-
- *Human assessors may still serve as a quality control mechanism for the LLM.*

Against Using LLMs for Judgments



- **LLMs are not people.** IR measures must be grounded in a human judgment. Ultimately, only the searcher tell which results are useful.
- We do not know if LLM judgments are systematically different from human judgments. Are the judgments made by LLMs are grounded in reality?
- Circularity of the evaluation reduces IR to a problem of efficiency.
- Generative AI models can be used to generate fake photos, videos, and other misinformation. LLMs are affected by bias, stereotypical associations, and adverse sentiments towards specific groups, which could be reflected in judgments.
- *Is it ethical to employ generative models in deployed systems **at all**?*

A Compromise



- LLMs for relevance assessment shows promise and deserves further study.
- While not replacing human assessment, LLMs could be used to evaluate early prototypes of approaches, for initial judgments for novel tasks, and for large-scale training.
- “AI Assistance” is probably the most credible path for LLMs to be incorporated during evaluation.
- We need research on innovative ways to use LLMs for assistance during the judgment process and how to leverage humans for verifying the LLMs’ suggestions.
- We should explore how the performance of human assessors changes, when they are, e.g., shown rationales or chain-of-thoughts generated by LLMs.



Thank you!
Questions?

Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini,
Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas,
Martin Potthast, Benno Stein, Henning Wachsmuth

