Who determines what is relevant? Humans or AI? Why not both!

A Spectrum of Human–AI Collaboration in Assessing Relevance

ACM Reference Format:
. 2024. Who determines what is relevant? Humans or AI? Why not both!: A Spectrum of Human–AI Collaboration in Assessing Relevance. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 JUDGING WHAT IS RELEVANT

To measure progress on better methods for web search, question answering, conversational agents, or retrieval from knowledge bases, it is essential to know which responses are relevant to a user's information need. Such judgments of what is relevant are traditionally obtained by asking human assessors.

With the latest improvements on auto-regressive large language models (LLMs) like chatGPT, researchers started to experiment with the idea of replacing human relevance assessment by LLMs [9]. The approach is simple: just ask an LLM chatbot, whether a response is relevant for an information need, and it does provide an "opinion".

In recent empirical studies on web search [3] but also in programming [7], human–computer interaction [5], or protein function prediction [10], it has been shown that LLM-generated opinions often agree with the assessment of humans. Some people already believe that the decision on what is relevant can be outsourced to "AI" in the form of LLMs, without any involvement of humans.

However, as we will argue next, there are severe issues with such a fully automated judgment approach—and these issues cannot be overcome by a technical solution. Rather than continuing with the ongoing quest to study where and how AI can replace humans, we suggest to examine forms of Human–AI collaboration for which we lay out a spectrum in this article.

2 WHY NOT JUST USE LLMS?

There are a number of issues that arise when we let LLMs judge the quality of search results or system-provided answers.

Judgment Bias towards a particular LLM. If we use a particular LLM to create relevance judgments to measure system quality, it would likely favour responses from systems that use the same or a similar LLM for response generation. Such a bias in the gold standard benchmark can lead to wrong findings when comparing multiple systems for quality.

Bias towards User Groups. Bender et al. [1] highlight the severe risk of LLMs to bias against underrepresented user groups. Such bias will likely be reflected in the relevance decisions made by the LLMs. Before using this technology, the computing community should develop approaches to quantify model bias and to understand possible ways of making LLMs more resilient when trained on biased data.

Resilience against Misinformation. Some information on the Web may seem topically relevant, but may be factually incorrect and hence should not be perpetuated. For example, on an information request like "do lemons cure cancer?" a system response may discuss factually incorrect information about healing cancer with lemons. While on topic, such potentially harmful responses should not be presented to a user. Factuality is already difficult for humans to assess correctly and without additional resilience mechanisms in place against misinformation, an LLM is unlikely to make correct relevance decisions in such situations.

LLM-based LLM Training. In a world where LLMs are used both for judging relevance and for generating responses, the issue of concept-drift also arises. Rather soon, a lot of web content will be LLM-generated. At the same time, new LLMs may be trained using large amounts of web content. This would lead to a cyclic learning problem, where possibly various LLMs agree on a definition of relevance that may not make sense to human end users.

Judging vs. Predicting. When a strong LLM is used to create relevance judgments for training a system to produce relevant responses, another question arises: Why not directly use the judging LLM to produce the response? There could be arguments with respect to reduced model size or improved response times, but still the trained system may not be able to surpass the quality of the judging LLM.

Truthfulness and Hallucinations. A well-known issue of LLMs is that they tend to generate text that contains inaccurate or false information (i.e., confabulation or hallucination). Responses are often presented in such an affirmative manner that makes it difficult for humans to detect errors. While chain-of-thought reasoning [8] or reinforcement from human feedback [11] can reduce the issue, it remains unclear to which extent the problem can be avoided.

LLM Relevance Judgments for Training only. Even when LLM-generated relevance judgments are only used to train a system—but not to evaluate it—many of the above issues still hold. Following the "garbage in / garbage out" mantra, issues arising from biased judgments, misinformation, and hallucinations will affect the quality of the end user-facing system.

3 LLMS ARE THE NEW CROWDWORKERS

It is yet to be understood what the benefits and risks associated with LLM technology are—especially when it comes to creating gold standards. A rather similar debate was spawned more than ten years ago when a lot of data annotations started to rely on crowdworkers instead of trained editors—with a substantial decrease in annotation quality somewhat compensated by a huge increase in annotated

Conference'17, July 2017, Washington, DC, USA
© 2024 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

data. Quality-assurance methods for crowdworkers were developed to obtain reliable labels [2]. With LLMs, history may repeat itself: a huge increase in available relevance assessment data at a possibly decreased quality. However, the specific extent of the deterioration is still unclear and requires further study.

A related idea is to allow LLMs to learn by “observing” human relevance assessors or by following an active learning paradigm [13]. Starting from generated relevance assessments that a human rates [17], the LLM could learn to provide better assessments. We believe that humans working with LLMs is not only an option, but is likely unavoidable as shown by recent results indicating that a large proportion of crowdworkers already make use of LLMs to increase their productivity [15].

4 A SPECTRUM OF HUMAN–LLM / AI COLLABORATION

Rather than exploring options for LLMs to replace humans, or reasons why LLMs should not be used, in this viewpoint article we discuss a spectrum of options to combine human and machine intelligence in a complementary and collaborative fashion.

The spectrum outlines different levels of collaboration. At one end, humans make judgments manually, while at the other end, LLMs replace humans completely. In between, LLMs assist humans with various degrees of interdependence or humans provide feedback to decision-making LLMs. A summary of our proposed levels of human–machine collaboration is shown in Table 1. In the following, we discuss each level in detail.

Human Judgment. On one extreme, humans make all relevance judgments manually without being influenced by an LLM. The relevance assessment interface only supports well-understood automatic features that do not require any form of automatic training / feedback. For instance, humans may decide which keywords should be highlighted during assessment, they may limit viewing a certain data subset, or they may sort the data in certain ways that influence their decision. This end of the spectrum thus represents the status quo practiced in the field of information retrieval and natural language processing, where humans are considered to be the only reliable arbiter.

Model in the Loop. A decision could be made easier for a human with an advanced level of automatic support, with the goal to save time and improve consistency in human judgments. For example, an LLM may generate a summary of a to-be-judged document, the human assessor then bases their relevance judgment on this compressed representation making the task quicker. Another approach could be to manually define information nuggets that are relevant [12] and to then train an LLM to automatically determine how many test nuggets are contained in the retrieved results (e.g., via a QA system). We hope to see more research on helpful sub-tasks that can be taken over by LLMs, such as highlighting of relevant passages and rationale generation.

An important open question is: How to employ LLMs and other AI tools to assist human assessors in devising more reliable and faster relevance judgments?

Table 1: Collaboration perspective: spectrum of possibilities for collaborative 🯅 human – ֎ machine task organization to make (relevance) decisions. The △ indicates where on the spectrum each possibility falls.

Collaboration Balance	Task Allocation
Human Judgment	
🯅 △—————— ֎	Humans manually decide (about relevance) without any kind of AI support.
🯅 —△————— ֎	Humans have full control of deciding but are supported by machine-based text highlighting, data clustering, etc.
Model in the Loop	
🯅 ——△———— ֎	Humans decide based on LLM-generated summaries needed for the decision.
🯅 ———△——— ֎	Balanced competence partitioning. Humans and LLMs focus on decisions they are good at.
Human in the Loop	
🯅 ————△—— ֎֎	Two (or more) LLMs each generate a decision, and a human selects the better one.
🯅 ————△—— ֎	An LLM makes a decision (and an explanation for it) that a human can accept / reject.
🯅 ————△—— ֎·*n*	LLMs are considered crowdworkers—varied by specific characteristics—, aggregated and controlled by a human.
Fully Automated	
🯅 ——————△֎	Fully automatic decision without humans.

Human in the Loop. Automated judgments could be produced by an LLM and then verified by humans. For instance, a first-pass automatic relevance judgment could come with a generated natural language rationale based on which a human accepts or rejects the judgment, or, following the “preference testing” paradigm [16], two or more LLMs each could generate a judgment while a human will select the best one. In such cases, a human might possibly only intervene in case of disagreements between the LLMs, thus increasing scalability. The purpose of this scenario is to simplify the decision for a human in most cases, and to use humans for difficult decisions or in situations where the LLMs generate a low confidence decision.

Many issues identified in the field of explainability in machine learning apply to this scenario, such as the human tendency to over-rely on machines, or to be unable to relate an LLM's decision to its generated rationale [4]. Thus, important open questions are: What are sub-tasks of the decision making process that require human input (e.g., prompt engineering [14]) and for what tasks should humans *not* be replaced by machines?

Fully Automated. If LLMs were able to make reliably judge relevance, they could completely replace humans in judging relevance. Indeed, a recent study showed a good correlation between LLMs' relevance judgments and human assessors [3], both, for an agreement on every judgment decision as well as to the correlation of leaderboards that rank systems by quality obtained with either set of judgments. Automatic relevance judgments might even surpass those of humans in terms of quality. However, it is not entirely clear how to detect such super-human performance.

An important open question is: In which cases can human relevance judgments be replaced entirely by LLMs?

A central aspect to be investigated is where on this four-level human–machine collaboration spectrum one can obtain relevance decisions that are most cost-efficient, fast, fair, and high in quality. In other words: how can one achieve ideal *competence partitioning* [6], where humans would perform tasks that humans are good at, while machines perform tasks that machines are good at.

5 VIEWPOINT

We believe that our current understanding is not sufficient to let LLMs perform relevance judgments without human intervention. Furthermore, we wish for more research on amplifying rather than replacing human intelligence using LLMs for judging the relevance of system responses, especially with respect to "model in the loop" and "human in the loop" scenarios.

To this end, we proposed a spectrum of possible ways in which we can balance human and artificial intelligence to increase the efficiency, effectiveness, and fairness in decision making processes like relevance assessment.

REFERENCES

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.). ACM, 610–623. https://doi.org/10.1145/3442188.3445922

[2] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Comput. Surv.* 51, 1 (2018), 7:1–7:40. https://doi.org/10.1145/3148148

[3] Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on Large Language Models for Relevance Judgment. *Proceeding of the 2023 SIGIR Conference on the Theory of Information Retreival (ICTIR)* (2023).

[4] Raymond Fok and Daniel S Weld. 2023. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. *arXiv preprint arXiv:2305.07722* (2023).

[5] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[6] PA Hancock. 2013. Task partitioning effects in semi-automated human–machine system performance. *Ergonomics* 56, 9 (2013), 1387–1399. https://doi.org/10.1080/00140139.2013.816374

[7] Ali Kashefi and Tapan Mukerji. 2023. Chatgpt for programming numerical methods. *arXiv preprint arXiv:2303.12093* (2023).

[8] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful Chain-of-Thought Reasoning. *CoRR* abs/2301.13379 (2023). https://doi.org/10.48550/arXiv.2301.13379 arXiv:2301.13379

[9] Sean MacAvaney and Luca Soldaini. 2023. One-Shot Labeling for Automatic Relevance Estimation. In *SIGIR '23: The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. https://doi.org/10.1145/3539618.3592032

[10] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* (2023), 1–8.

[11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[12] David Sander and Laura Dietz. 2021. EXAM: How to Evaluate Retrieve-and-Generate Systems for Users Who Do Not (Yet) Know What They Want. In *Proceedings of the 2nd International Conference on Design of Experimental Search & Information REtrieval Systems (DESIRES)*. https://www.cs.unh.edu/~dietz/papers/sander2021exam.pdf

[13] Seungmin Seo, Donghyun Kim, Youbin Ahn, and Kyong-Ho Lee. 2022. Active Learning on Pre-trained Language Model with Task-Independent Triplet Loss. In *Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 11276–11284. https://ojs.aaai.org/index.php/AAAI/article/view/21378

[14] Taylor Sorensen, Joshua Robinson, Christopher Michael Rytting, Alexander Glenn Shaw, Kyle Jeffrey Rogers, Alexia Pauline Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 819–862. https://doi.org/10.18653/v1/2022.acl-long.60

[15] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. *CoRR* abs/2306.07899 (2023). https://doi.org/10.48550/arXiv.2306.07899 arXiv:2306.07899

[16] Jennifer Windsor, Laura M Piché, and Peggy A Locke. 1994. Preference testing: A comparison of two presentation methods. *Research in developmental disabilities* 15, 6 (1994), 439–455. https://doi.org/10.1016/0891-4222(94)90028-0

[17] Jiechen Xu, Lei Han, Shazia Sadiq, and Gianluca Demartini. 2023. On the role of human and machine metadata in relevance judgment tasks. *Information Processing & Management* 60, 2 (2023), 103177.