Enhancing Human Annotation: Leveraging Large Language Models and Efficient Batch Processing

Oleg Zendel
RMIT University
Melbourne, VIC, Australia

J. Shane Culpepper
The University of Queensland
Brisbane, QLD, Australia

Falk Scholer
RMIT University
Melbourne, VIC, Australia

Paul Thomas
Microsoft
Adelaide, SA, Australia

ABSTRACT

Large language models (LLMs) are capable of assessing document and query characteristics, including relevance, and are now being used for a variety of different classification labeling tasks as well. This study explores how to use LLMs to classify an *information need*, often represented as a user query. In particular, our goal is to classify the cognitive complexity of the search task for a given "backstory". Using 180 TREC topics and backstories, we show that GPT-based LLMs agree with human experts as much as other human experts. We also show that batching and ordering can significantly impact the accuracy of GPT-3.5, but rarely alter the quality of GPT-4 predictions. This study provides insights into the efficacy of large language models for annotation tasks normally completed by humans, and offers recommendations for other similar applications.

CCS CONCEPTS

• **Information systems** → **Retrieval tasks and goals**; **Task models**; • **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

Cognitive task complexity, large language models, search task classification

ACM Reference Format:
Oleg Zendel, J. Shane Culpepper, Falk Scholer, and Paul Thomas. 2024. Enhancing Human Annotation: Leveraging Large Language Models and Efficient Batch Processing. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24), March 10–14, 2024, Sheffield, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3627508.3638322

1 INTRODUCTION

Knowing the cognitive complexity of a particular information need allows a search engine to answer queries differently based on the user's needs. For example, when a user searches for information on a complex topic, such as understanding *how Google works*, the ideal Search Engine Result Page (SERP) would provide a concise overview along with relevant resource links. Conversely, when a query represents a straightforward topic such as *how did Julius Caesar die?*, an effective SERP could just answer the question directly. In both cases, the capability of a search engine to determine query complexity improves the user search experience.

Large Language Models (LLMs), exemplified by OpenAI's GPT models, demonstrate remarkable proficiency across many different Natural Language Processing (NLP) tasks such as text generation, summarization, and classification. However, to the best of our knowledge, they have not been explored for classifying cognitive complexity. This study asks: *can an LLM determine the cognitive complexity of an information need?*

By adopting a comparative approach, our research evaluates the agreement between expert human annotators and LLMs for the cognitive complexity categorization task. Our initial findings suggest that LLMs can achieve a proficiency level that is comparable to expert human annotators for this task. This exploration not only shows the potential of using LLMs for intricate annotation tasks within Information Retrieval (IR) and psychology, but also demonstrates a cost-efficient solution for a task that normally requires human experts, and is an important step towards democratizing ML research. By alleviating financial and logistical burdens associated with labor-intensive annotation processes, LLMs enable us to do more with less – allowing more inclusive and collaborative research endeavors to be undertaken. In subsequent sections we describe our methodology, datasets, results, implications, and potential limitations – offering a comprehensive perspective on using LLMs to determine the cognitive complexity of a task in IR.

2 RELATED WORK

2.1 Annotating Information Needs

The impact of search task complexity on information-seeking behavior and search engine usage is well-documented [11]. This extends to the complexity of the required information, the diversity of consulted sources, and the number of steps required in the search process [8, 24]. Several models have been proposed to categorize these tasks. Initially, Byström and Järvelin [7] introduced a five-level taxonomy, ranging from automatable tasks to a genuine decision-making task. This taxonomy was later refined by Bell and Ruthven [6] to a three-level model, primarily based on the ability of humans to distinguish between the levels and the clarity of the tasks. Wu

CHIIR '24, March 10–14, 2024, Sheffield, United Kingdom

ACM ISBN 979-8-4007-0434-5/24/03…$15.00
https://doi.org/10.1145/3627508.3638322

et al. [26] studied task complexity for interactive information retrieval systems, and introduced a hierarchy based on a revision of Bloom's taxonomy of learning objectives [15]. Their study showed that tasks with higher cognitive complexity lead to longer search sessions, an increased number of queries submitted, and a higher number of search result clicks.

Based on these previous taxonomies, Bailey et al. [3] propose a task complexity taxonomy tailored for offline information-seeking tasks test collections. The key idea is that information-seeking tasks exhibit a diversity of different characteristics, and task complexity significantly influences search behavior. In their study, Bailey et al. adopted a three-level hierarchy, based on the cognitive complexity hierarchy originally introduced by Wu et al.. The three levels encompass a spectrum of information needs: "*Remember*" tasks involve retrieving facts which answer simple questions (e.g., "How did Eva Peron die?"), "*Understand*" tasks require interpreting, summarizing, and explaining information, while "*Analyze*" tasks require information to be broken down, gaining a deeper understanding of each part, and finally creating a comprehensive overview.

Bailey et al. categorized a set of TREC topics using these complexity types by creating backstories that clearly describe the information need represented in each topic. All four authors annotated the topics independently, leading to an overall inter-annotator agreement of 0.664, using Fleiss' κ. Notably, the agreement varied for different complexity levels, ranging from 0.456 for *Analyze* tasks to 0.907 for *Remember* tasks. This suggests that *Remember* tasks tend to be relatively easy to identify and agree upon. In cases where no majority rating could be determined among the four annotators, a thorough discussion was conducted in order to reach a consensus, resulting in a single confirmed task complexity label for each topic which can be used as a ground truth in future experiments. This comprehensive classification process is the foundation of the experiments and subsequent analyses in this work.

2.2 Human vs. Machine Annotation

There has been significant interest in using LLMs for tasks that traditionally required human labeling [10, 17, 30]. Arguably, ALPACA [25] is the seminal work for instruction-based prompting in LLMs. The key idea is to define a prefix prompt for a chat session, where the initial prompt is a clear set of instructions describing how an LLM should respond. This approach is now widely used to solve a variety of information retrieval tasks [2, 12, 23] and, more widely, ML-based tasks [14]. In fact, one of the results that is now routinely being reported is that LLMs outperform crowdsourced human labeling for a number of tasks [13]. Recently Microsoft described using OpenAI's GPT models for relevance assessment in the Bing search engine [22]. The paper describes a study that compares human experts with LLMs for the relevance assessment task. The authors conclude that LLMs are able to achieve comparable performance to human experts.

These studies challenge our current beliefs about the quality of human relevance labeling and our ongoing dependence on it [4]. They also highlight the potential of using LLMs for annotation tasks, and emphasise their affordability and scalability in annotation.

2.3 Categorizing Search Task Complexity

Previous research on categorizing the complexity of search tasks provides a foundation for our study. The refined taxonomy proposed by Bailey et al. [3] provides a practical set of labels for information-seeking tasks. Their approach to categorizing tasks and measuring inter-annotator agreement provides the foundation for a comprehensive comparison between our approach and human experts. Research using LLMs for tasks traditionally requiring human labeling highlights the value of using LLMs in this domain. We build on this by exploring the use of LLMs for the cognitive complexity classification labeling task in IR.

3 DATA AND EXPERIMENTAL SETUP

We employ two pre-trained LLMs for annotation: GPT-3.5-turbo and GPT-4, accessible through the OpenAI API. These models operated with their default hyper-parameters, and the version used was 0613.

Our dataset consists of 180 annotated topics obtained from Bailey et al. [3]. The topics originate from several TREC collections: the Question Answering Track 2002 (70 topics, 1824-1893), the Robust Track 2003 (60 topics, 303-610), and the Terabyte Track 2004 (50 topics, 701-750). Note that the dataset was enhanced by the authors with a set of new backstories – brief information need statements to accompany each topic. These backstories were designed to clarify the context and motivation of the search requests, making the topic statements easier to understand. The annotation process included all 180 topics, with categorization performed by the four authors, whom we consider human experts as they also created each of the backstories.

Topics are assigned to one of three categories based on the backstory: "*Remember*" (R), "*Understand*" (U), or "*Analyze*" (A). In order to evaluate LLM performance, we compare the categorizations with the original ones provided by human expert annotators. To simplify the initial comparison, we focus on the 107 topics where there was unanimous agreement by the human experts, which we will refer to as "*full agreement*" (FA).[1] The FA set is instrumental in assessing LLM performance using several different models and configurations. Subsequently, to compare LLM performance with human experts, we also consider the entire set of 180 topics, referred to as "*all topics*" (AT). The distribution of categories for each set is shown in Table 1. To measure quality, we compute Krippendorf's α using an ordinal scale [16]. This metric is well-suited to evaluating agreement between multiple annotators and categories, even in cases where data is missing. We adopt an ordinal scale as it aligns with the assumption of an inherent order among the categories, as is often applied in Bloom's Taxonomy, which the labeling scheme is derived from. In this order, "Remember" represents the least complex, followed by "Understand," and "Analyze" denoting the most complex category. Consequently, we assign values of 1, 2, and 3 to the categories, respectively, so that for example a higher level of disagreement is signified between "Remember" and "Analyze", versus a disagreement between "Remember" and "Understand".

The distribution of topic lengths for the FA set is shown in Figure 1, where topic lengths are measured by the number of terms contained in the backstory. To examine the potential separability to

[1] Originally, there were 112 topics that had a full agreement; however, five of these were marked with a question mark and were excluded from the FA set.

Table 1: The distribution of categories for each set. Columns represent the categories assigned by the annotators to a certain topic.

	{R}	{U}	{A}	{A, U}	{R, U}	{A, R, U}	Total
AT set	65	35	12	53	11	4	180
FA set	65	33	9	0	0	0	107

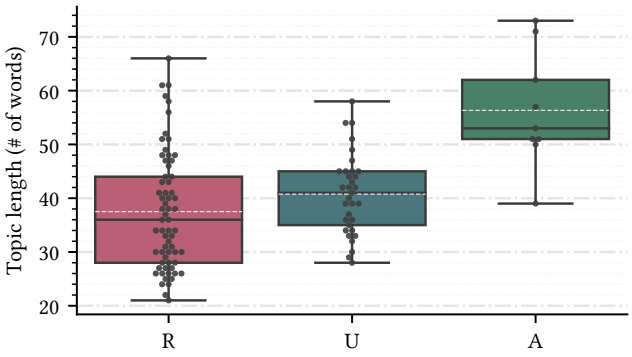
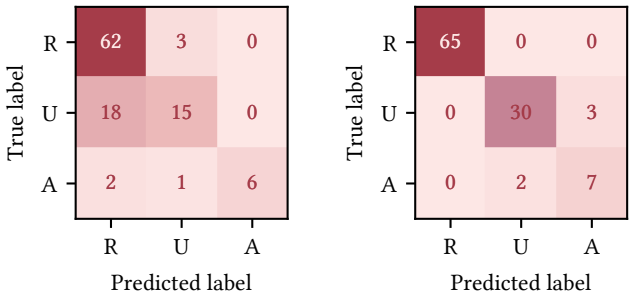


Figure 1: The distribution of the topic lengths for each category in the FA set, which contains 107 topics, where topic length is the number of terms contained in the backstory. The black line represents the median and the white dashed line represents the mean.



(a) Using topic length as the sole feature for classification.

(b) Using GPT-4 with batching for classification.

Figure 2: Confusion matrices for the FA set.

categories by topic length, we employed SVC [19] to fit and classify the entire FA set (both train and evaluate). Topic length, measured in terms, was the sole feature used for classification. The resulting confusion matrix for the FA set is depicted in Figure 2a. While there are noticeable differences in topic lengths depending on the category, length alone does not appear to be a reliable predictor of the category label.

4 RESULTS

Large Language Models (LLMs) are stochastic by design, and so is the model output. To gain a better understanding of the models' behavior, we conduct experiments where the temperature hyperparameter is set to zero, which results in "mostly deterministic" outputs for the same input prompt.[2] We then explore different LLM configurations. More specifically, we examine the following factors:[3]

(1) The Instruction Prompt: We assess the impact of different instruction prompts on LLM categorization.
(2) Batch Topic Order: We investigate how the order of topics in a batch can influence responses from the LLM.
(3) Zero-Shot vs. One-Shot: We compare the performance of LLMs when categorizing topics in a zero-shot manner (without examples) and a one-shot manner (using a single example).

4.1 Batching

Considering that annotation costs depend on the number of input and output tokens when using the OpenAI API, we devise a cost-efficient method to minimize the cost. Our approach involves using the OpenAI "`system`" role to provide an instruction prompt once for each batch of topics.[4] We then maximize the batch size to utilize the tokens available per API call most effectively. It is worth noting that while the OpenAI API supports straightforward batching of topics, other LLM APIs may not offer this feature and may require a specific implementation or fine-tuning.

Tokenization schemes vary for each LLM model. For example, GPT-3.5-turbo allows a maximum of 4,000 tokens per query, and costs 0.0015 USD per 1,000 input tokens and 0.002 USD per 1,000 output tokens.[5] Our calculations factor in an estimated 40 tokens per topic for each classification response. The mean costs and time to annotate 107 topics are shown in Table 2. Implementing this approach led to a 74.6% cost reduction for GPT-3.5-turbo and a 69% reduction for GPT-4 when compared against non-batched annotation. Additionally, annotation time is reduced by 46.3% and 48.1% for GPT-3.5-turbo and GPT-4, respectively. It is also noteworthy that the costs and time associated with each task, even when using the most expensive model, GPT-4, are significantly lower than traditional crowdsourcing methods. To provide context, a rough estimate of using MTurk to annotate the same topics would cost approximately 15 USD for a single worker,[6] assuming an average of roughly one minute per topic, including reading initial instructions.

4.2 Instruction Prompts

Instruction prompts – brief text added to the beginning of a prompt, before the topic text – can have an impact on LLM outputs [22, 28, 29]. Instruction prompts play a pivotal role in guiding an LLM to generate the desired responses, making them a critical component in the annotation process. We examined several properties of instruction prompting:

(i) **Prompt 1:** A comprehensive verbal prompt that describes the role, task, and context of the categories. It instructs the

[2] As per information from the OpenAI Developer Forum.
[3] The code, prompts, and model responses are available at: https://github.com/Zendelo/LLMs-Complexity-Annot.
[4] https://platform.openai.com/docs/guides/gpt.
[5] Source: https://openai.com/pricing, as of October 31, 2023.
[6] Based on the minimum federal wage of 7.25 USD in the US.

Table 2: Annotation costs and time for each LLM model using the FA set.

	Model	Batched	Single	Difference
Cost (USD)	GPT-3.5	0.03	0.13	0.10
	GPT-4	0.44	1.41	0.97
Time (Minutes)	GPT-3.5	1.32	2.46	1.14
	GPT-4	6.93	13.36	6.43

Table 3: Krippendorf's α for each model and prompt. The values for *batched* are the mean values for runs using six randomized orderings for the topics. The best result for each row is underlined.

		Prompt					
Model		1	2	3	4	5	6
GPT-3.5	Batched	0.63	0.86	0.79	<u>0.87</u>	0.72	0.85
	Single	<u>0.87</u>	0.82	0.84	0.86	0.84	0.87
GPT-4	Batched	<u>0.97</u>	0.92	0.93	0.96	0.96	0.95
	Single	0.91	0.87	0.81	0.90	<u>0.92</u>	0.91

model to provide a brief explanation in addition to categorization (226 terms).

(ii) **Prompt 2:** A concise prompt with instructions similar to Prompt 1 (91 Terms).

(iii) **Prompt 3:** Identical to Prompt 2, but with a request for an explanation before categorization (91 terms).

(iv) **Prompt 4:** Similar to Prompt 1, with an additional request for the model to count the number of categorized topics and ensure all are categorized (248 terms).

(v) **Prompt 5:** Similar to Prompt 4, including an input and output example for each category (315 terms).

(vi) **Prompt 6:** Same as Prompt 5 but without the explanation request (244 terms).

The results of the annotation using these different prompts are summarized in Table 3. Several noteworthy findings can be observed: batching appears to enhance the results for GPT-4 but decreases agreement slightly when using GPT-3.5-turbo. Also of interest, if an example for each category is added to the prompt (Prompts 5 & 6), there is no observable improvement in agreement. Asking the model to provide a detailed explanation before the categorization (Prompts 2 vs. 3) also shows minimal impact, with slightly reduced agreement. Removing the explanation altogether and simply requesting labels (Prompts 5 vs. 6) appears to have no effect on agreement, but offers cost and time advantages. While it could be argued that omitting a reason for each label makes the output less *explainable*, it is unclear if the class selected and the reason provided are correlated. This is an interesting question that deserves further exploration. For example, the explanation may simply be information that the model "believes" the user wants to see, rather than the reason the model chose the label it provided.

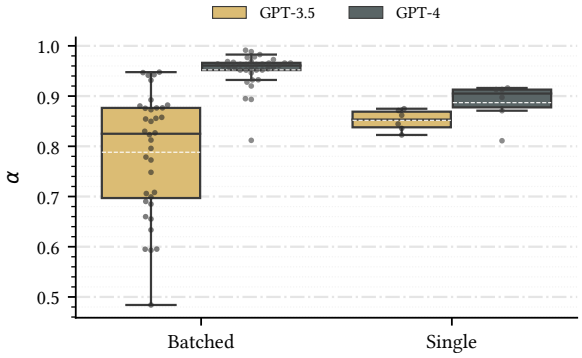


Figure 3: Krippendorff's α for each model when comparing annotation of a single topic vs. a batch of topics. The different values represent six different prompts and six different orderings for the topics. The black line represents the median and the white dashed line represents the mean.

On investigation, the reasons that were provided in several of the model configurations appear to be informative, but the legitimacy of label justification in LLMs is an interesting area that warrants further study.

4.3 Batch Topic Order

The order in which input tokens are presented to an LLM can affect the output that is produced. We therefore investigate how topic order in batching affects LLM responses. Formally, we hypothesize that topic order should not significantly influence LLM responses. To test our hypothesis, we conduct experiments using the FA set, and vary the order of topics within each batch. We tested six different orderings, comprised of four random permutations of topics, as well as sorted ascending and descending based on the topic ID. The results, summarized in Figure 3, show that GPT-3.5-turbo is more sensitive to topic order, indicating that the particular prompt that is used (recall that topic order is part of the prompt) should be constructed carefully. In contrast, GPT-4 is remarkably stable for the different orderings, showing little impact on the final results produced.

Overall, from Table 3 we observe that the agreement with human annotators is remarkably high. Specifically, GPT-4 achieves an agreement in the range of 0.92 to 0.97 for the FA set when using batch processing of topics. In contrast, GPT-3.5-turbo has an agreement in the range of 0.63 to 0.87 for the same task. Although GPT-3.5-turbo exhibits slightly lower agreement and much higher variance than GPT-4, it still demonstrates the potential utility of LLMs in a variety of practical applications. A key concern with GPT-3.5-turbo is its sensitivity to the input prompt. This sensitivity may potentially be mitigated by running the model multiple times using a set of different prompts and then aggregating the results, but this is a prospect we leave for future research.

4.4 GPT-4 as a Human Expert

Next, we compare the performance of GPT-4 with human experts. Building on the insights gained in the previous experiments, we

Table 4: Leave-one-out Krippendorff's α using an ordinal scale for the classes.

	A.1	A.2	A.3	A.4	GPT-4
α	0.84	0.84	0.84	0.83	0.83

Table 5: Krippendorff's α between every pair of annotators using an ordinal scale for the classes. The cell color is based on the α value before rounding.

	A.1	A.2	A.3	A.4	GPT-4
A.1	—	0.82	0.84	0.83	0.83
A.2	0.82	—	0.82	0.84	0.84
A.3	0.84	0.82	—	0.84	0.81
A.4	0.83	0.84	0.84	—	0.89
GPT-4	0.83	0.84	0.81	0.89	—

formulate a hypothesis that GPT-4 performance will be very similar to human experts. To test this hypothesis, we conduct experiments using the AT set, which contains all 180 annotated topics. The agreement between the four human annotators (labeled A.1 to A.4) for this data is $\alpha = 0.83$. We then use the GPT-4 model to create a further set of annotations of the topics in the AT set using Prompt 1, and batch the topics. The model's results on the FA set are illustrated in Figure 2b, with a Krippendorff's α of 0.98 for the FA set.[7]

To evaluate performance, we regard the model as an additional (fifth) annotator and calculate the agreement between all the annotators. Our approach uses leave-one-out cross-validation to compute the agreement between annotators by excluding one annotator at a time, and repeating this process for every annotator. The results are summarized in Table 4.

The results in Table 4 show that using a GPT-4 model result instead of one from a human annotator increases the overall agreement in three out of four of the comparisons. The agreement is only worse when annotator 4 (A.4) is omitted.

To investigate agreement in more detail, we also compute the pairwise agreement between the GPT-4 model and the human annotators, shown in Table 5. The results closely parallel the agreement observed between the human annotators in the original experiment. Notably, the agreement between the GPT-4 model and annotator 4 (A.4) is the highest among all of the annotator pairs. This finding supports the comparability of GPT-4 to human annotation, and suggests it could be valuable to use it as an additional annotator when performing human annotation. In summary, our results indicate that the GPT-4 model can match the quality of human annotators for this classification task, and reinforce our belief that these models are a valuable addition to an annotation pool.

5 DISCUSSION

LLMs are becoming popular across a wide range of fields, including IR, education, medicine, law, finance, and psychology [9]. Their potential to enhance human annotation processes is quite clear based on our findings, and merits further exploration. In this section, we discuss the implications of our findings, highlight potential challenges, and suggest future research directions.

5.1 Cost Comparison

The cost and time efficiency of employing LLMs is a significant advantage, especially when comparing it to traditional crowdsourcing methods [22]. The affordability of existing pre-trained LLMs for cognitive complexity classification in IR underscores their potential in making the dataset construction process more accessible. Our study demonstrates that not only is the cost significantly lower, but the quality of the output is also comparable to expert human annotators, which have been shown to be more reliable than crowdsourced annotations a number of times in previous work [20, 21], but is expensive (consider the cost of using four highly paid IR researchers for 1-2 hours each with the cost of gathering a set of annotations using Mechanical Turk for example). The quality is surprisingly high, and the cost of using the OpenAI API is minimal.

5.2 Caveats and Future Research

While this study demonstrates the value of LLMs for cognitive complexity classification tasks, there are still a number of limitations and potential challenges that must be resolved before an LLM can be reliably used for other important tasks. Future research directions should focus on refining LLM-based annotation processes and addressing these existing issues.

Despite the cost-effectiveness of LLMs, several counterpoints should be considered. The output quality may not always match human-generated results. Humans can better understand context and nuances that LLMs may not currently be capable of, and therefore topic experts should always be involved in the process of evaluating and refining the output of LLMs. Furthermore, it is well-known that the cost of training LLMs can be substantial, both in monetary terms as well as in negative impacts on the environment [5, 31]. LLMs also require regular updates and maintenance in order to capture new information, which adds to their overall costs. Finally, the models are stochastic so there is no guarantee that they will produce the same results every time they are used, which makes the results difficult to reproduce and potentially less reliable. One way to mitigate this issue is to run the models multiple times and aggregate the results, but this increases the costs further, and for the annotation process, it is unclear how much this approach would improve the quality of the final result. We leave this as a topic for future research.

LLMs are often seen as "black boxes" since their decision-making process is not transparent or easily interpretable by humans. Many of the models, such as the OpenAI models used in this study, are proprietary and not open source, and so the data used to train them are not publicly available. This lack of transparency can be problematic as the models may be biased; it is not currently clear how to identify and correct such biases, especially when the underlying data used in the models is not known. This issue is particularly concerning when LLMs are used in applications where the consequences of incorrect decisions could be severe and could potentially harm people (healthcare or insurance approval, for example). Therefore, LLMs should be used cautiously, and their output should be

[7] It is worth noting that the disagreements in the FA set are only between "Understand" and "Analyze".

carefully scrutinized by human experts before being used to make important decisions.

One important consideration is that in our study, information needs are represented by concise backstories that provide a contextual framework for the annotators and the LLMs. While beneficial for cognitive complexity classification and tasks involving user information needs [1, 18, 27], such backstories are not always available. In many real-world IR scenarios, queries are not accompanied by enough information to clearly define the underlying information need. Therefore, the LLMs would need to infer this additional information using only the query, a well-known challenge in the IR community, and a topic that warrants additional study in the future.

The integration of LLMs in the field of IR for tasks such as cognitive complexity classification could potentially be a disruptive technology in the crowdsourcing industry. This disruption could have both positive and negative consequences, representing a significant transformation in the way research and information classification tasks are completed by researchers. The full extent and impact of this transformation warrants further study as it is still unclear what percentage of labeling tasks being performed by human crowdsourcing can be reliably automated using LLMs.

6 CONCLUSION

This study explored the use of LLMs for the task of classifying the cognitive complexity of search scenarios, demonstrating key advantages and disadvantages of automating the task using a LLMs. We presented a cost and time-effective annotation process which integrates LLM annotation with human annotation, and demonstrated the cost and time savings that can be achieved without compromising the quality of the annotations produced. The potential benefits and challenges arising from our study require additional exploration in the evolving landscape of LLM-driven IR.

ACKNOWLEDGMENTS

We thank the reviewers for their helpful comments. This work was supported by the Australian Research Council's *Discovery Projects* Scheme (grant DP190101113).

REFERENCES

[1] Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer, and Paul Thomas. 2023. Can Generative LLMs Create Query Variants for Test Collections? An Exploratory Study. In *Proc. SIGIR*. 1869–1873.
[2] Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-Source Large Language Models Outperform Crowd Workers and Approach ChatGPT in Text-Annotation Tasks. *arXiv preprint arXiv: 2307.02179* (2023).
[3] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User Variability and IR System Evaluation. In *Proc. SIGIR*. 625–634.
[4] Peter Bailey, Paul Thomas, Nick Craswell, Arjen P. De Vries, Ian Soboroff, and Emine Yilmaz. 2008. Relevance Assessment: Are Judges Exchangeable and Does It Matter. In *Proc. SIGIR*. 667–674.
[5] Lotfi Belkhir and Ahmed Elmeligi. 2018. Assessing ICT global emissions footprint: Trends to 2040 & recommendations. *Journal of Cleaner Production* 177 (2018), 448–463.
[6] David J. Bell and Ian Ruthven. 2004. Searcher's Assessments of Task Complexity for Web Searching. In *Proc. ECIR*. 57–71.
[7] Katriina Byström and Kalervo Järvelin. 1995. Task Complexity Affects Information Seeking and Use. *Inf. Proc. & Man.* 2 (1995), 191–213.
[8] Bogeum Choi, Austin Ward, Yuan Li, Jaime Arguello, and Robert Capra. 2019. The Effects of Task Complexity on the Use of Different Types of Information in a Search Assistance Tool. *ACM Trans. Inf. Sys.* 38 (2019).
[9] D. Demszky, D. Yang, D. S. Yeager, C. J. Bryan, M. Clapper, S. Chandhok, C. J. Eichstaedt, C. Hecht, J. Jamieson, M. Johnson, M. Jones, D. Krettek-Cobb, L. Lai, N. JonesMitchell, D. C. Ong, C. S. Dweck, J. J. Gross, and J. W. Pennebaker. 2023. Using Large Language Models in Psychology. *Nature Reviews Psychology* (2023).
[10] Guglielmo Faggioli, Laura Dietz, Charles L A Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proc. ICTIR*. 39–50.
[11] Souvick Ghosh, Manasa Rath, and Chirag Shah. 2018. Searching as Learning: Exploring Search Behavior and Learning Outcomes in Learning-Related Tasks. In *Proc. CHIIR*. 22–31.
[12] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks. *Proc. Nat. Acad. Sci.* 120, 30 (2023), e2305016120.
[13] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks. *Proc. Nat. Acad. Sci.* 120 (2023).
[14] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Proc. NeurIPS*, Vol. 35. 22199–22213.
[15] David R. Krathwohl, Lorin W. Anderson, and Benjamin Samuel Bloom. 2001. *A Taxonomy for Learning, Teaching, and Assessing : A Revision of Bloom's Taxonomy of Educational Objectives* (complete ed.).
[16] Klaus Krippendorff. 2022. *Content Analysis: An Introduction to Its Methodology* (fourth ed.). SAGE Publications, Inc.
[17] Sean MacAvaney and Luca Soldaini. 2023. One-Shot Labeling for Automatic Relevance Estimation. In *Proc. SIGIR*. 2230–2235.
[18] Joel Mackenzie, Rodger Benham, Matthias Petri, Johanne R Trippas, J Shane Culpepper, and Alistair Moffat. 2020. CC-News-En: A Large English News Corpus. In *Proc. CIKM*. 3077–3084.
[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[20] Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Falk Scholer. 2021. On the Effect of Relevance Scales in Crowdsourcing Relevance Assessments for Information Retrieval Evaluation. *Inf. Proc. & Man.* 58 (2021), 102688.
[21] Paul Thomas, Gabriella Kazai, Ryen W White, and Nick Craswell. 2022. The Crowd is Made of People Observations from Large-Scale Crowd Labelling. In *Proc. CHIIR*. 25–35.
[22] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large Language Models Can Accurately Predict Searcher Preferences. *arXiv preprint arXiv:2309.10621* (2023).
[23] Petter Törnberg. 2023. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. *arXiv preprint arXiv: 2304.06588* (2023).
[24] Pertti Vakkari. 1999. Task Complexity, Problem Structure and Information Actions: Integrating Studies on Information Seeking and Retrieval. *Inf. Proc. & Man.* 6 (1999), 819–837.
[25] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions. In *Proc. ACL*. 13484–13508.
[26] Wan-Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. 2012. Grannies, Tanning Beds, Tattoos and NASCAR: Evaluation of Search Tasks with Varying Levels of Cognitive Complexity. In *Proc. IIiX*. 254–257.
[27] Oleg Zendel, Melika P Ebrahim, J Shane Culpepper, Alistair Moffat, and Falk Scholer. 2022. Can Users Predict Relative Query Effectiveness?. In *Proc. SIGIR*. 2545–2549.
[28] Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E. Gonzalez. 2023. TEMPERA: Test-Time Prompt Editing via Reinforcement Learning. In *Proc. ICLR*.
[29] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large Language Models Are Human-Level Prompt Engineers. In *Proc. ICLR*.
[30] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large Language Models for Information Retrieval:A Survey. *arXiv preprint arXiv: 2308.07107* (2023).
[31] Guido Zuccon, Harrisen Scells, and Shengyao Zhuang. 2023. Beyond CO2 Emissions: The Overlooked Impact of Water Consumption of Information Retrieval Models. In *Proc. ICTIR*. 283–289.