Large Language Models-aided Literature Reviews: A Study on Few-Shot Relevance Classification

Large Language Models-aided Literature Reviews: A Study on Few-Shot Relevance Classification

Flavio Giobergia
Politecnico di Torino
Turin, Italy
flavio.giobergia@polito.it

Alkis Koudounas
Politecnico di Torino
Turin, Italy
alkis.koudounas@polito.it

Elena Baralis
Politecnico di Torino
Turin, Italy
elena.baralis@polito.it

***Abstract*—Conducting a comprehensive literature review is a critical step in the research process, often requiring significant time and effort to identify and evaluate relevant academic papers. Traditional methods rely heavily on keyword-based searches to filter potentially relevant articles, which are then manually reviewed for inclusion. This process is not only labor-intensive but also susceptible to subjective biases, potentially affecting the consistency and accuracy of the review. This paper explores the potential of using large language models (LLMs) to automate the relevance assessment phase of literature reviews, thereby addressing some of the limitations of traditional methods. Specifically, we investigate the application of few-shot learning with various LLMs to determine the relevance of papers identified through keyword searches. We evaluate the sensitivity of this approach to the number of shots provided and compare the performance across different open-source LLMs, including Llama-3, Mistral, and Phi-3. Our findings aim to provide insights into the effectiveness of using LLMs for literature review processes, potentially transforming how researchers conduct literature reviews.**

***Index Terms*—Large language models, Few-shot learning, Literature review, Relevance classification**

I. Introduction

The literature review is a foundational component of academic research, providing a critical overview of existing knowledge, identifying gaps, and contextualizing new findings within the broader scholarly discourse. This process traditionally begins with a keyword-based search, often utilizing academic databases such as Scopus, PubMed, and Google Scholar to compile a list of potentially relevant papers. Researchers then manually review these papers to determine their relevance, a task that is both time-consuming and prone to subjective bias [1], [2]. This manual screening often involves extensive reading and critical evaluation, which can introduce inconsistencies and inefficiencies in the research process.

Recent advancements in artificial intelligence, particularly in natural language processing (NLP), have introduced new possibilities for automating and enhancing various stages of the research process. Large language models (LLMs), such as GPT-4 [3] and Llama-3 [4], have demonstrated remarkable capabilities in understanding and generating human-like text. These models have been used in diverse applications ranging from text summarization and translation to generating coherent and contextually relevant writing pieces. Given their proficiency in handling vast amounts of textual data, LLMs present promising tools for automating the literature review process, potentially transforming how researchers interact with existing literature.

This paper explores using LLMs to assist in the relevance assessment phase of literature reviews. Specifically, we investigate the application of few-shot learning, a method where the model is provided with a small number of examples (shots) to guide its understanding and performance on a given task. Few-shot learning leverages the model's ability to generalize from limited data, making it an attractive approach for reducing the manual effort involved in screening papers. By providing just a few examples of relevant and non-relevant papers, the model can learn to discern relevance criteria and apply them consistently across a larger dataset.

We conduct a series of experiments to evaluate the sensitivity of the few-shot learning approach to the number of shots provided. This involves varying the number of examples given to the LLM and observing how its performance changes in terms of accuracy and consistency in relevance assessments. Additionally, we compare the performance of different open-source LLMs, including Llama-3 [4], Mistral [5], and Phi-3 [6], to determine which model offers the best support for this task. Our goal is to identify the optimal conditions under which LLMs can effectively assist in literature reviews, thereby enhancing the efficiency and reliability of this critical research activity.

The potential benefits of integrating LLMs into the literature review process are manifold. Automating relevance assessments can significantly reduce the time researchers spend on initial screenings, allowing them to focus more on in-depth analysis and synthesis of the literature. Moreover, by minimizing subjective bias, LLMs can contribute to more consistent and objective evaluations of research papers. This, in turn, can improve the overall quality and robustness of literature reviews, facilitating more informed and impactful research outcomes.

The remainder of this paper is organized as follows: Section II reviews related work in the application of NLP and LLMs in literature reviews and other academic tasks. Section III describes our methodology, including the selection of LLMs, the few-shot learning approach, and the evaluation criteria. Section IV defines the experimental setup, while Section V presents the results of our experiments with a discussion of the findings. Finally, Section VI concludes the paper and outlines

potential directions for future research.

II. Related Works

The literature review process involves multiple steps to ensure rigor and reproducibility [2], [7]. Among these, the relevance assessment of papers manually done by researchers is particularly time-consuming and prone to subjective bias [1], [2]. Recent advancements have focused on using deep learning algorithms for automation to expedite this process.

The authors of [2] introduced an end-to-end solution for citation screening utilizing deep neural networks, achieving significant workload reductions of at least 10% in various domain data analyses. The BERT [8] model and its variants, known for excelling in numerous NLP tasks, have been employed for document screening. Its application is relatively new [9]–[11], with many studies highlighting the importance of prioritizing eligible citations early in the screening process.

However, pre-trained algorithms for citation screening and relevance classification require a substantial number of labeled papers to avoid overfitting, necessitating large domain-specific datasets. Alternative methods, such as meta-learning and few-shot learning, are explored to mitigate the effort needed for labeling. [7] proposed using a model-agnostic meta-learning algorithm (MAML) [12] to paper classification in the literature review process, aiming to enhance efficiency and accuracy.

Recently, Large Language Models (LLMs) have shown exceptional capabilities in storing factual knowledge and achieving state-of-the-art results in NLP tasks but face challenges like hallucination and outdated knowledge [13]–[15]. Retrieval Augmented Generation (RAG) addresses these issues by integrating external databases, enhancing the accuracy and relevance of LLMs [15]. Recent research also highlights the potential of LLMs in information retrieval, ranking, zero-shot, and few-shot learning, with improvements through instruction tuning and prompt engineering [16]–[18]. These advancements suggest a promising direction for using LLMs in tasks such as citation screening and paper relevance classification [19]–[21].

III. Methodology

Our proposed architecture for automating the relevance assessment phase of literature reviews consists of two primary components: the *Engine Retrieval* block and the *LLM-Based Relevance Classification* block. This section offers a detailed overview of each component and the overall workflow.

A. Engine Retrieval Block

The initial step in our methodology involves retrieving potentially relevant academic papers using a keyword-based search strategy. This is achieved by leveraging existing academic search engines, such as Scopus, through their respective APIs. The process is as follows:

1) **Keyword Selection**: Identify a comprehensive set of keywords related to the research topic. These keywords form the basis for querying the search engine.
2) **Query Execution**: Utilize the search engine's API to execute the keyword search, generating a list of papers that match the search criteria.
3) **Initial Filtering**: Apply basic filtering criteria (e.g., publication date, language, and document type) to refine the list of retrieved papers and ensure relevance.

The output of this block is a curated collection of academic papers that are potentially relevant to the literature review.

B. LLM-Based Relevance Classification Block

The second component of our architecture focuses on determining the relevance of the retrieved papers using a large language model (LLM). This process leverages few-shot learning to classify papers as relevant or irrelevant. The steps involved are:

1) **Few-Shot Prompting**: Select a small number of example papers (shots) that are pre-labeled as relevant or irrelevant. These examples serve as prompts to guide the LLM's understanding and classification task.
2) **Relevance Assessment**: For each paper retrieved in the previous block, input the paper's abstract (or other pertinent sections) along with the few-shot prompt into the LLM. The model then generates a classification, indicating whether the paper is relevant or irrelevant.

The output of the LLM-based block is a refined and annotated list of papers, each labeled as relevant or irrelevant based on the model's assessment. This refined list is expected to be more accurate and consistent, helping researchers efficiently identify the most pertinent literature for their review.

C. Data Annotation

To identify a specific use case, we decided to focus on the growing body of work examining biases in speech models and their varying performance across different subpopulations [22]–[32]. We began by retrieving all relevant papers that matched a set of predefined keywords, which resulted in a total of 795 works published from 2018 onwards.

Next, we removed duplicate entries, reducing the number of documents to 490. These papers were then subjected to an annotation process conducted by two human annotators. The annotators evaluated each paper's relevance based on its content and alignment with the provided keywords, marking them as either relevant or irrelevant. To ensure the reliability of our selection, we retained only those papers where the annotators' scores were in agreement, discarding the rest. This filtering process left us with a curated pool of 300 papers. From this refined collection, we randomly selected a subset to use as input within the prompt for the LLMs.

IV. Experimental Setup

To evaluate the effectiveness of our proposed methodology, we conducted a series of experiments. This section details the performance metrics, sensitivity analysis, and model comparison used to assess the approach.

A. Large Language Models

We conducted a comparative analysis of various open-source LLMs, including Llama-3 [4], Mistral v0.2 [5], and Phi-3 [6], to evaluate their performance in the relevance

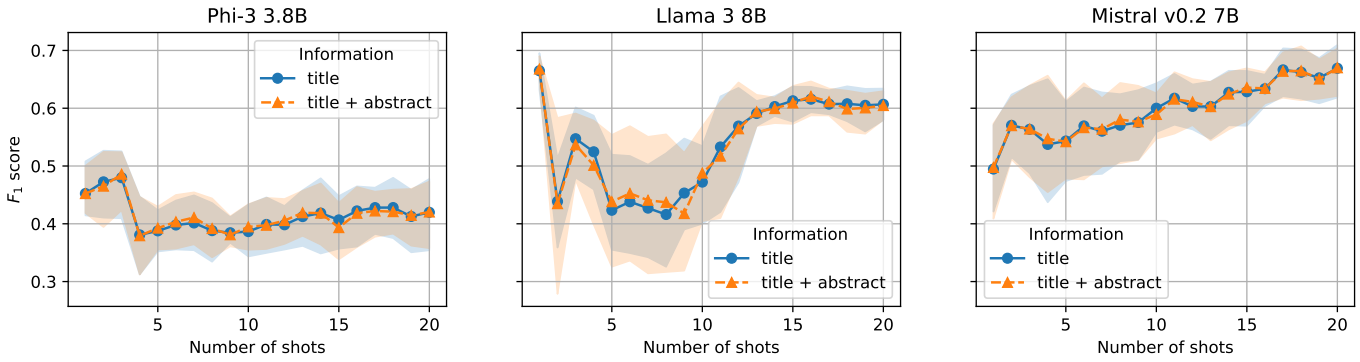

Fig. 1. Performance of various models, as the number of shots changes, when prompted with the title only, or the title and the abstract of each document.

classification task. We aimed to determine which LLM is most effective for automating the literature review process.

Specifically, we focused on the instruction fine-tuned versions of these models [33]. After their initial training, the models underwent a post-training process that included supervised fine-tuning and direct preference optimization. This additional training aimed to enhance their ability to follow instructions accurately and maintain safety standards.

We tested the 7B Mistral-v0.2 model with a 32k context window, the 8B Llama-3 model with an 8k context window, and the 3.8B Phi-3 model with a 128k context window. We sourced the instruction-tuned checkpoints for these models from the Hugging Face hub [34].

B. Prompting setup

To understand whether additional information can be of use in improving the classification performance, we tested two separate prompting approaches:

- *Title-only*. In this kind of prompt, we only include in the prompt the title of the manuscript being analyzed, both for the document to be annotated and for the shots passed. This is meant to include fewer overall tokens in the resulting prompt.
- *Title and abstract*. In this kind of prompt, we provide information about both the title and the abstract of all documents (target document and shots). This is meant to provide further context about the articles at the cost of increasing the number of tokens included in the prompt.

C. Performance Metrics

We utilized various classification metrics to evaluate the performance of the LLM-based classification against the manually-curated ground truth set: (i) accuracy, the ratio of correctly classified papers to the total number of papers, providing an overall measure of the model's performance; (ii) F_1 score, the harmonic mean of precision and recall, offering a single metric that balances precision and recall; (iii) the False Positive Rate (FPR), representing the fraction of negative samples that have been incorrectly labeled as positives; and (iv) the False Negative Rate (FNR), representing the fraction of positive samples that have been incorrectly predicted as negative.

A False Positive case represents a situation where a document is reported as relevant to the topic of interest when it is not. This implies that some time will be wasted manually going through irrelevant documents. Conversely, a False Negative case arises when a relevant document is discarded as it is not considered relevant, meaning a potentially useful resource will not be taken into account.

Different situations may prioritize one type of error over the other. For instance, when conducting a literature review for a survey, it might be acceptable to spend additional time reviewing (and then discarding) false positives, while it may be more detrimental if a relevant paper is not included in the analysis.

By using these metrics, we aimed to thoroughly evaluate and compare the effectiveness of the LLMs in classifying relevant literature.

D. Sensitivity Analysis

To gain a comprehensive understanding of the sensitivity of the few-shot learning approach, we varied the number of shots provided in the prompts and observed the corresponding impact on model performance. This analysis was designed to identify the optimal number of examples necessary to achieve high classification accuracy.

In our study, we considered a range of 1 to 20 shots for each prompt. Since shots need to be manually labeled, we identified this range as a number of documents that can be manually annotated relatively quickly by a human. To preserve the balance between classes, we provide the same number of positive (relevant) and negative (irrelevant) shots. To ensure the robustness and reliability of our findings, we repeated the analysis with 5 different seeds for every model-shot combination. By doing so, we aimed to account for variability in model performance due to different initializations and to determine the most effective number of shots for consistent and accurate relevance classification.

TABLE I
PERFORMANCE FOR THE VARIOUS LLMS, WHEN PROMPTED WITH THE BEST PERFORMING NUMBER OF SHOTS, WITH INFORMATION ON THE TITLE ONLY, OR TITLE AND ABSTRACT.

Model	Information	$\mathbf{F_1}$ **score** ↑	**Accuracy** ↑	**FPR** ↓	**FNR** ↓
Phi-2 (3-shots)	w/ title	0.4797 ± 0.0763	0.4852 ± 0.0777	0.6336 ± 0.1083	0.1417 ± 0.0317
	w/ title + abstract	0.4858 ± 0.0671	0.4913 ± 0.0688	0.6274 ± 0.1005	**0.1361 ± 0.0385**
Llama 3 8B (1-shot)	w/ title	0.6649 ± 0.0411	0.7167 ± 0.0554	0.2661 ± 0.0897	0.3370 ± 0.0561
	w/ title + abstract	0.6678 ± 0.0267	**0.7207 ± 0.0342**	**0.2599 ± 0.0548**	0.3397 ± 0.0427
Mistral 7B v0.2 (20-shots)	w/ title	0.6691 ± 0.0603	0.7167 ± 0.0746	0.2917 ± 0.1069	0.2540 ± 0.0463
	w/ title + abstract	**0.6700 ± 0.0560**	0.7189 ± 0.0697	0.2872 ± 0.0997	0.2603 ± 0.0443

V. RESULTS

Figure 1 compares, for each considered model, whether the inclusion of additional information about each document (the abstract) provides any helpful information or not, as the number of shots varies. It can be seen that the performance, as measured with the F_1 score metric, does not benefit from introducing additional information. This indicates that the information contained in the title is already, in general, enough to provide information as to whether an article should be relevant or not to a topic of interest in a concise way.

Since adding the abstract does not appear to provide a meaningful benefit, we compare different models when prompted with the information about the title only. Figure 2 compares the three models considered in this study. First, we note that increasing the number of shots does not always lead to an improvement in terms of performance. This appears to be the case for Mistral. Instead, Phi-3 works best for a low number of shots. Increasing them produces lower, stable performance. The behavior of Llama-3 is instead rather curious. It works particularly well when provided with a single shot (and achieves the best overall performance observed in this experimental section). However, increasing the number of shots (in the range of 2-10) does not improve the results and, instead, produces a meaningful drop. However, further increasing the number of shots does lead to a steady improvement until saturation (around 15 shots). Any increase in the number of shots does not produce a meaningful change.

Based on these considerations, we can identify – for each model – the best configuration of shots. We do this by choosing the number of shots that produces the highest performance in terms of F_1 score. We report in Table I the performance in terms of F_1 score, accuracy, FPR, and FNR for the best-performing configurations of shots for each model. The main conclusion that can be drawn from these results is that there is no one model that achieves the best overall performance. For instance, while Llama 3 achieves the best performance in terms of False Positive Rate, it is affected by the largest False Negative Rate. It is also interesting to note that, despite being much smaller than its competitors (3.8B vs. 7B vs. 8B parameters), Phi-3 achieves the best performance in terms of FNR. However, it performs the worst on all other metrics. As already discussed, different contexts imply a different interest in terms of metrics to be maximized.

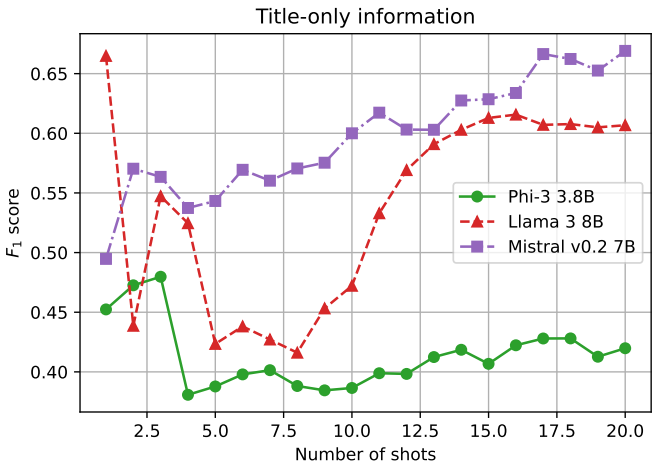

Fig. 2. Models comparison, when prompted with the title-only information, as the number of shots provided in the prompt varies.

VI. CONCLUSION AND FUTURE WORKS

In this work, we explored initial results that were obtained by using Large language models (LLMs) to define the relevance of scientific papers to a specified topic. The initial results are promising and indicate that LLMs can understand the assigned task and produce relevant results. In particular, we show that some models (e.g., Phi-3) are generally good at discarding irrelevant articles (low False Negative Rate), whereas others produce more balanced results (e.g., Llama 3, Mistral v0.2). This work also highlights the fact that abstracts do not appear to provide additional information to the models: using only the title produces comparable results, with the added benefit of producing a more concise prompt. For this work, we have mainly focused on the title of the paper (and optionally the abstract). Of course, the actual relevance of a paper depends on the entire contents of the document itself, as well as the information conveyed by its structure. An obvious next step consists of leveraging the entire structure of the paper as additional information: various techniques have been proposed in the literature to extract graph data from unstructured data (e.g., visually rich documents, images) by segmenting them and modeling the positional relationship occurring therein [35], [36]. Additionally, we plan on expanding the information provided by prompting with a summary of the entire paper

rather than simply the title (and optionally the abstract). In this way, key information that may be missing from the abstract but is included in the paper would be passed to the model for a more accurate classification.

ACKNOWLEDGEMENTS

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

REFERENCES

[1] G. Tsafnat, P. Glasziou, G. Karystianis, and E. Coiera, "Automated screening of research studies for systematic reviews using study characteristics," *Systematic reviews*, vol. 7, pp. 1–9, 2018.

[2] R. van Dinter, C. Catal, and B. Tekinerdogan, "A multi-channel convolutional neural network approach to automate the citation screening process," *Applied Soft Computing*, vol. 112, p. 107765, 2021.

[3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[4] A. Meta, "Introducing meta llama 3: The most capable openly available llm to date," *Meta AI.*, 2024.

[5] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023.

[6] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

[7] M. K. de Melo, A. V. A. Faria, L. Weigang, A. G. Nery, F. A. R. de Oliveira, I. T. Barreiro, and V. R. R. Celestino, "Few-shot approach for systematic literature review classifications." in *WEBIST*, 2022, pp. 33–44.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] A. Ioannidis, "An analysis of a bert deep learning strategy on a technology assisted review task," *arXiv preprint arXiv:2104.08340*, 2021.

[10] X. Qin, J. Liu, Y. Wang, Y. Liu, K. Deng, Y. Ma, K. Zou, L. Li, and X. Sun, "Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews," *Journal of clinical epidemiology*, vol. 133, pp. 121–129, 2021.

[11] W. Kusa, A. Hanbury, and P. Knoth, "Automation of citation screening for systematic literature reviews using neural networks: A replicability study," in *European Conference on Information Retrieval*. Springer, 2022, pp. 584–598.

[12] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[13] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023.

[14] F. Borra, C. Savelli, G. Rosso, A. Koudounas, and F. Giobergia, "Malto at semeval-2024 task 6: Leveraging synthetic data for llm hallucination detection," 2024.

[15] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[16] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," *arXiv preprint arXiv:2211.01910*, 2022.

[17] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, Z. Dou, and J.-R. Wen, "Large language models for information retrieval: A survey," *arXiv preprint arXiv:2308.07107*, 2023.

[18] Y. Hou, J. Zhang, Z. Lin, H. Lu, R. Xie, J. McAuley, and W. X. Zhao, "Large language models are zero-shot rankers for recommender systems," in *European Conference on Information Retrieval*. Springer, 2024, pp. 364–381.

[19] M. Haman and M. Školník, "Using chatgpt to conduct a literature review," *Accountability in research*, pp. 1–3, 2023.

[20] B. J. Jansen, S.-g. Jung, and J. Salminen, "Employing large language models in survey research," *Natural Language Processing Journal*, vol. 4, p. 100020, 2023.

[21] S. Agarwal, I. H. Laradji, L. Charlin, and C. Pal, "Litllm: A toolkit for scientific literature review," *arXiv preprint arXiv:2402.01788*, 2024.

[22] R. Tatman and C. Kasten, "Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions." in *Proc. INTERSPEECH*, 2017.

[23] J. L. Martin and K. Tang, "Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be"," in *Proc. INTERSPEECH*, 2020.

[24] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," *arXiv preprint arXiv:2103.15122*, 2021.

[25] Y. Zhang, Y. Zhang, B. M. Halpern, T. Patel, and O. Scharenborg, "Mitigating bias against non-native accents," in *Proc. INTERSPEECH*, 2022.

[26] C. Liu, M. Picheny, L. Sarı, P. Chitkara, A. Xiao, X. Zhang, M. Chou, A. Alvarado, C. Hazirbas, and Y. Saraf, "Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions," in *ICASSP*, 2022.

[27] P. Dheram, M. Ramakrishnan, A. Raju, I.-F. Chen, B. King, K. Powell, M. Saboowala, K. Shetty, and A. Stolcke, "Toward fairness in speech recognition: Discovery and mitigation of performance disparities," in *Proc. INTERSPEECH*, 2022.

[28] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, L. Cagliero, L. de Alfaro, E. Baralis, and D. Amberti, "Exploring subgroup performance in end-to-end speech models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[29] Y. Zhang, A. Herygers, T. Patel, Z. Yue, and O. Scharenborg, "Exploring data augmentation in bias mitigation against non-native-accented speech," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[30] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, E. Reale, L. Cagliero, S. Cumani, L. de Alfaro, E. Baralis, and D. Amberti, "Towards comprehensive subgroup performance analysis in speech models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1468–1480, 2024.

[31] A. Koudounas, E. Pastor, G. Attanasio, L. de Alfaro, and E. Baralis, "Prioritizing data acquisition for end-to-end speech model improvement," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1–5.

[32] A. Koudounas, F. Giobergia, E. Pastor, and E. Baralis, "A contrastive learning approach to mitigate bias in speech models," in *Proc. INTERSPEECH*, 2024.

[33] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[34] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, and A. M. et al., "Transformers: State-of-the-art natural language processing," in *EMNLP: System Demonstrations*, Oct. 2020.

[35] A. Pasini, F. Giobergia, E. Pastor, and E. Baralis, "Semantic image collection summarization with frequent subgraph mining," *IEEE Access*, vol. 10, pp. 131 747–131 764, 2022.

[36] X. Liu, F. Gao, Q. Zhang, and H. Zhao, "Graph convolution for multimodal information extraction from visually rich documents," *arXiv preprint arXiv:1903.11279*, 2019.