Harnessing Large Language Models for Cost-Effective Relevance Labeling in Job Search Systems

Ishan Shrivastava, Nadav Barkai, Ritvik Kharkar

Ziprecruiter, Inc
604 Arizona Ave, Santa Monica, California, 90401, USA
{ishans, nadavb, ritvikk}@ziprecruiter.com

Abstract

Relevance labeling is a crucial step in optimizing search systems. Traditional methods, including user feedback and crowdsourced annotation, are often expensive and time-consuming. This paper explores the use of large language models (LLMs) for relevance labeling in job search tasks, showing they can achieve comparable quality to crowdsourced labels at a fraction of cost and turnaround time. We demonstrate that systematic modifications to prompts and innovative feature engineering can help smaller open-source LLMs (such as Qwen2.5-32B-Instruct) achieve comparable performance to a powerful closed-source LLM (GPT-4o). We also outline our approach to monitoring potential biases.

1 Introduction

Relevance labeling is a critical component of optimizing job search systems at ZipRecruiter. Traditional methods, such as crowdsourcing, can be costly and time-consuming, especially as scale and data demand grows. Recent successes of LLMs on general data labeling tasks , (Pavlovic and Poesio 2024), (Wang et al. 2024) including relevance labeling tasks (Faggioli et al. 2023), (Thomas et al. 2024), (He et al. 2024) offer an alternative. We therefore turned to these models for producing relevance labels with high agreement to human annotators at significantly reduced cost and turnaround time.

This paper explores the use of LLMs for job search relevance labeling, focusing on prompt engineering to achieve optimal performance. We evaluate closed-source and open-source LLMs, monitor biases and enumerate the cost and turnaround time reductions when using LLMs for data labeling versus when using crowdsourced annotators. Our findings suggest that LLMs can match human annotators in quality while offering superior scalability and efficiency, making them a compelling choice for modern relevance labeling needs. Particularly, we show how open-source LLMs can benefit from additional features providing comparable performance to closed-source LLMs thereby opening doors to better data privacy, faster research cycles, and cost savings.

2 Related Work

Microsoft Bing's work (Thomas et al. 2024) has been a crucial inspiration for this work. The authors highlight how queries alone might not provide sufficient context and therefore utilized additional features representing "what the query means (description) and which documents should be considered responsive (narrative)". This served as a motivation to generate/extract query intention and job detail features that we explore in this work. The authors also observed the LLM to over estimate the relevance on a given query-document for longer documents motivating us to check for this bias as well.

In their work, (He et al. 2024) showed how to utilize LLM explanations and rationales along with few-shot examples for relevance labeling. We were inspired by this to consider LLM explanations for error analysis and used these explanations qualitatively to iterate and improve our prompts. In the future, we would also like to utilize these LLM explanations more systematically as done by the authors.

3 Methodology

We selected Cohen's Kappa (McHugh 2012) as our primary evaluation metric for measuring the level of agreement of an LLM's relevance score predictions with those sourced from crowdsourced annotators. We experimented with different prompts and different LLMs (both open and closed source) as well as explored various input features for enhancing these prompts.

Our workflow included the following steps:

- Given an LLM and specific prompt structure, generate relevance predictions on a set of query and job pairs and compute Cohen's Kappa to understand the level of agreement between crowdsourced labels and LLM predictions.
- Perform bootstrapped significance testing to see whether or not the current prompt is better than the previous baseline. If the current prompt is better, consider this prompt as the new baseline. Perform error analysis and iterate.

We started developing and iterating on the prompts over a prominent closed source LLM: GPT-4o (OpenAI 2023). After designing the best prompt based on GPT-4o, we evaluated it on well performing open source LLMs such as the Qwen2.5 family of LLMs (Team 2024), particularly Qwen2.5-14B-Instruct and Qwen2.5-32B-Instruct models. We hypothesized that to achieve a comparable performance to closed source LLMs such as GPT-4o, the smaller open

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

source LLMs might require additional features, such as explicitly introducing extracted information, to compensate for the lower capacity. We elaborate on this in Sections 4 and 5.

3.1 Defining Relevance

To measure the relevance of a query, job pair, we considered different grading scales depending on the use case. When training a relevance classifier, for example, one might consider a binary relevance scheme while for a ranking task, one might consider a ternary or quinary relevance scheme. We therefore aimed at getting labels at a quinary level so that we could map those down to ternary or binary grades.

The quinary relevance grades (**Q**) along with the corresponding mapping to binary (**B**) and ternary (**T**) grades are shown in Table 1. Crowdsourced annotators or LLMs were provided a prompt containing these relevance definitions and asked to label roughly 5000 query, job pairs accordingly. Consistent with the language in these definitions, we are trying to understand how likely a job seeker would be to apply to a job if they had searched a given query.

Q	Description	T	B
0	This is an **extremely irrelevant** job, makes me wonder why I am being shown this job for my search query.	0	0
1	This is an **irrelevant** job and it is unlikely that I will apply to such a job given my job search query.	0	0
2	This job is **somewhat relevant** to my job search query and there's a chance I might apply to this if I am desperate.	1	1
3	This job is **relevant** to my job search query as it is close enough to what I am looking for. Although it doesn't satisfy all of my requirements from a job, I will still apply to this job.	2	1
4	This is an **extremely relevant** job. I would definitely apply to this as it satisfies all of my requirements.	2	1

Table 1: Job Relevance Grades. Q: Quinary, T: Ternary, B: Binary.

4 Prompt Engineering

Now we describe how we iterated and optimized through various prompt structures, with both closed source (GPT-4o) and open source (Qwen2.5-14B-Instruct and Qwen2.5-32B-Instruct) LLMs. We also describe how we explored different features to encapsulate job seeker's search intention and a given job's details. Due to confidentiality we cannot share the exact prompts but will share their basic structure.

4.1 Existing Features

We started with the obvious input features for our prompts which were also readily available: **query**: The search query executed by the job seeker; **title**: Title of the job; **description**: Description of the job; **company**: Company that this job belongs to.

4.2 Baseline Prompt

Our first prompt was heavily inspired by the labeling guidelines written for crowdsourced annotation efforts. We had the LLM to assume the role of a job seeker, briefly explained the task of job relevance labeling and provided a definition for the quinary relevance grades as described in Table 1.

4.3 With Examples Prompt

Error analysis on the baseline prompt (Section 4.2) showed us how the LLM was being too strict and therefore underestimating the relevance of a given query, job pair. Consider the examples in Table 2, that show how the baseline prompt would label a pair irrelevant if *any* of the intentions specified in the query were not satisfied by the job. We therefore decided to add a few descriptive examples for each of the relevance grades. Adding these examples resulted in statistically significant improvement over the baseline prompt.

Query	Job	True Label	Pred Label
dollar tree	Assistant Store Manager at Dollar General	3	0
Walmart cashier	PT-Head Cashier at Lowe's	3	0
French	Senior Mining Engineer requiring proficiency in French	2	0

Table 2: False Negatives with Baseline Prompt

4.4 Chain of Thought (CoT) Prompt

It is a well-known practice to utilize the Chain-of-Thought prompting technique (Wei et al. 2023) where we explicitly prompt the LLM with a step-by-step reasoning process. We therefore break the entire job search labeling process into multiple steps and add those in our previous best prompt (Section 4.3) enabling the LLM to think step-by-step:

- **Step 1**: Intention analysis of the job seeker based on their search query.
- **Step 2**: Analyze the details (seniority level, skills required etc) present in the job based on its title, description and company.
- **Step 3**: Assess Step 1 and Step 2 to determine the relevance of the job to the query.

Utilizing CoT prompting resulted in statistically significant improvement over the previous best prompt (with examples prompt).

4.5 Semantic Relevance Prompt

Upon seeing the improvements with the introduction of CoT (Section 4.4), we looked at the rationales given by the LLM for their predictions and discovered that the LLM was focusing on "Attractiveness" over pure "Semantic Relevance". A job might be semantically relevant to a given query, but job aspects such as the salary or the geographic location, could make it more attractive to one person compared to some other person. Figure 1 shows an example of such a false negative prediction made by the CoT prompt (Section 4.4) along with its explanation. The LLM identifies that the job matches the "Graveyard" shift aspect of the query. However, the specific role of "Housekeeping Shampooer" may not directly align with *all* the job seekers entering this query.

So far our prompts used a "Would you apply?" kind of language while describing the relevance grades as shown to human annotators (Table 1) as well. Human annotators might have a more general inductive bias and naturally infer relevance despite application phrasing; LLMs on the other job seeker's search query: "graveyard"; job title: "Housekeeping Shampooer"; job description: "... Housekeeping Starting Wage $17.00 (Graveyard) Full Time ...."

LLM Explanation for the label: "While the job does match the 'graveyard' shift aspect of the query, indicating that it is offered during the overnight hours which the job seeker is interested in, the specific role of 'Housekeeping Shampooer' may not directly align with what all job seekers entering this query might be looking for, as it is quite specific to housekeeping and maintenance within a casino environment. Therefore, the relevance is somewhat limited by the specificity of the job's industry and role, making it a less than perfect match for a general 'graveyard' shift query without more context on the job seeker's desired industry or role.

Figure 1: False Negative from the CoT Prompt, where the LLM focuses on Attractiveness over Semantic Relevance.

hand treated it as a binary choice, being overly strict in labeling. We therefore remove this apply-centric language and instead focus on Semantic Relevance by adding phrases like "semantically related" or "semantically relevant/irrelevant". This change led to our best performance prompt, which was a statistically significant improvement over the CoT prompt (Section 4.4).

4.6 Binary Semantic Relevance Prompt

Until now we had the LLM generate quinary grades and mapped those to ternary or binary grades. We experimented with asking the LLM to directly generate binary grades hoping it would lead to a better performance on the binary labels but to our surprise it did not (more on this in Section 5).

4.7 Prompts with Additional Engineered Features

So far we had only used the obvious and available features specified in Section 4.1. These features were able to provide really satisfactory results when using a bigger and better closed source LLMs such as GPT-4o. Such closed source LLMs are great at reasoning and therefore perfect for such data labeling tasks but if we'd like to scale our data labeling these costs can easily stack up. On the other hand, smaller open source LLMs, while being weaker at reasoning, are easier to host and therefore can be comparatively cheaper. But to make up for their weak reasoning capabilities, we hypothesized that they will require additional features that can probably help them get comparable performance to the closed source LLMs. We therefore created three feature engineering prompts to generate/extract three types of features: **Job Seeker's Descriptive Intention Feature**, **Job Seeker's Granular Intention Features** and **Job's Detail Features**. These were the additional input features that we provided the prompt described in Section 4.5 with in order to improve its performance when using open source LLMs.

Job Seeker's Descriptive Intention Feature: We were inspired by Microsoft Bing's work (Thomas et al. 2024) where they showed how using short descriptive text, provided by the search query originator and indicative of their search intention, proved to be an important feature for their relevance labeling task. Bing's search team has invested a lot of resources in collecting these additional features using expert and domain labeling along with asking original users about their search intention. We decided to generate this feature (Figure 2) by prompting Qwen2.5-14B-Instruct to analyze queries and their historically applied job titles.

Job Seeker's Granular Intention Features: We also generated detailed granular features describing the job seeker's intent as shown in Figure 3. Similar to how we generated the query descriptive intention feature, the granular intention features were also generated by prompting an LLM (Qwen2.5-14B-Instruct) given the historically applied job titles for a given job search query.

query	descriptive_intention
Dental Hygienist	The job seekers are primarily looking for positions as dental hygienists, with a focus on roles that offer various scheduling options including full-time, part-time, and temporary positions. Some seekers also have an interest in bilingual and community-focused dental hygienist roles.
CDL A	Job seekers with a commercial driver's license (CDL) are looking for a variety of driving positions ranging from local to regional and over-the-road (OTR) roles, including truck driving, delivery driving, and various specialized driving jobs such as ready mix truck drivers and dump truck drivers.
SAP Consultant	The job seekers are looking for a variety of roles within the SAP ecosystem, ranging from technical to functional consulting positions, with a focus on specific modules such as Basis, MM, SD, and FICO. They also show interest in both remote and on-site opportunities, indicating a preference for flexible work arrangements while seeking expertise in SAP solutions.
walmart cashier	The job seekers are specifically looking for cashier positions at Walmart, focusing on roles that involve handling customer transactions and potentially managing cash registers.

Figure 2: Example of Job Seeker's Descriptive Intention Feature.

query	employment_type_intention	seniority_type_intention	job_title_intention	skills_intention	industry_intention	company_intention
Dental Hygienist	['full time', 'part time', 'temporary']	NaN	Dental Hygienist	NaN	NaN	NaN
CDL A	[]	NaN	NaN	commercial drivers license	NaN	NaN
SAP Consultant	[]	senior, mid, entry level	SAP Consultant	NaN	NaN	NaN
walmart cashier	[]	NaN	Cashier	NaN	Retail	Walmart

Figure 3: Example of Job Seeker's Granular Intention Features.

Job's Detail Features: We also extracted various details from the job descriptions such as the job's industry, seniority of the job, location type of the job, skills and certifications required by the job (Figure 4).

5 Experiments

In this section we share the experiments we ran with both open and closed source LLMs over a dataset of $\sim$5000 query, job pairs. Both use zero temperature. Our primary metric has been Cohen's Kappa (McHugh 2012) particularly over the binary relevance grade (**B**) but we also observe the agreements over the quinary (**Q**) and ternary (**T**) relevance grades. As per (McHugh 2012) Cohen, ≤ 0: no agreement; 0.01–0.20: none to slight; 0.21–0.40: fair; 0.41– 0.60: moderate; 0.61–0.80: substantial; 0.81–1.00: almost perfect agreement with crowdsourced labels. We also report Precision and Recall for the binary relevance grade.

The dataset used in these experiments is a representative sample containing both head queries, representing the most frequent searches and also the torso queries to be representative of our online search traffic. Although we had $\sim$20000 query, job pairs labeled by crowdsourced annotators, we observed 5000 pairs to be representative enough and hence experimented over them to save closed-source experimentation costs.

5.1 Performance of Closed Source LLM (GPT-4o)

Table 3 shows the performance of GPT-4o on prompts described in Sections 4.2 to 4.6. We observed that our baseline prompt (4.2), inspired by the crowdsourced annotator instructions, resulted in substantial agreement for the binary relevance scheme (κ_B) and moderate agreement for the quinary (κ_Q) and ternary (κ_T) relevance schemes showing the powerful reasoning capabilities of GPT-4o. Adding examples and CoT reasoning improved performance incrementally. **Upon introducing the notion of Semantic Relevance (Section 4.5) we got our best performing prompt with substantial agreement over the binary and**

title	location_type	employment_type	seniority_level	is_managerial_job	certifications
Creative Traffic Manager	not_remote	['full_time']	Mid	True	[{'certification_name': 'Credit security clearance', 'issuing_organization': 'N/A'}]
Senior ABAP PI/PO Developer (Hybrid Role)	remote_hybrid_flexible	['full_time']	Senior	False	NaN

Figure 4: Example of some Job's Detail Features.

ternary relevance schemes but not over quinary relevance schemes. This gives us enough confidence in trusting these labels for these two schemes.

Interestingly, directly generating binary labels underperformed compared to quinary-to-binary mapping, as granular grades helped handle edge cases. There were no cases where the quinary semantic relevance prompt (Section 4.5) missed labeling a relevant query, job pair that the binary prompt corrected for. However the quinary semantic relevance prompt was able to make up for the binary semantic relevance prompt's poor recall by correctly catching many relevant pairs that the binary semantic relevance prompt had missed labeling correctly. This finding is also in line with what (Zhuang et al. 2024) found for the Ranking task.

Section	κ_B	κ_T	κ_Q	Precision	Recall
4.2	0.656	0.593	0.433	0.974	0.781
4.3	0.705	0.618	0.442	0.948	0.852
4.4	0.719	0.63^*	0.461	0.932	0.883
4.5	**0.723**	$\mathbf{0.627}^*$	0.438	0.899	0.935
4.6	0.678	—	—	0.967	0.807

Table 3: Experimental results for closed source LLM prompts. All experiments use GPT-4o. Precision and Recall is over the binary relevance scheme. *indicates no statistically significant difference.

5.2 Performance of Open Source LLMs

The performance of Qwen2.5-14B and Qwen2.5-32B (shown in Table 4), was evaluated using the best semantic relevance prompt (Section 4.5). Initially, with only existing features like query and job title and descriptions, both models performed similarly, achieving moderate agreement on binary labels ($\kappa_B \approx 0.60$). The additional engineered features (Section 4.7) led to significant improvements. **The 32B parameter model gave us the best performance with 0.695 Cohen's κ score on the binary relevance labeling scheme when Job Seeker's descriptive intention feature was provided in addition to the existing features**. This improvement suggests that context-rich unstructured features compensate for the model's weaker reasoning capabilities compared to closed-source alternatives. We were unable to achieve substantial agreement over the other two labeling schemes with open-source LLMs reflecting limitations in their ability to handle more granular distinctions. In contrast, the Job Seeker's granular intention features and Job's Detail Features did not prove to be as useful likely because these structured features lacked the nuanced context of unstructured feature needed for relevance predictions.

When the job seeker's descriptive intention feature was added to GPT-4o, along with the existing features, its performance actually decreased (κ_B=0.711 vs. 0.723 without the feature). This decline may be due to noise introduced by the generated descriptive intent, which does not originate directly from the job seeker but is inferred by the LLM. GPT-4o, with its strong reasoning capabilities can likely interpret job seeker intent directly from queries, making the additional feature redundant and potentially confusing. Conversely, the Qwen2.5-32B-Instruct model, which has fewer parameters and weaker reasoning, benefited significantly from the descriptive intent feature. Despite some noise, it captured enough true intent to help the model achieve performance closer to GPT-4o on binary relevance labeling task.

N_P	**D,G,J**	κ_B	κ_T	κ_Q	**Pr.**	**Re.**
14B	_,_,_	0.602^*	0.504	0.365	0.839	0.957
14B	D,_,_	0.65^-	0.533	0.365	0.872	0.924
14B	_,G,_	0.65^-	0.545	0.393	0.87	0.929
32B	_,_,_	0.609^*	0.5	0.337	0.837	0.967
32B	D,_,_	$\mathbf{0.695}^+$	0.57	0.372	0.893	0.924
32B	D,G,_	0.692^+	0.56	0.358	0.885	0.935
32B	_,G,_	0.67	0.54	0.348	0.867	0.953
32B	D,_,J	0.656	0.535	0.345	0.865	0.944
32B	_,G,J	0.612	0.496	0.319	0.840	0.961

Table 4: Experimental results for open source LLM prompts with Qwen2.5 Instruct models. N_P is the number of parameters in the LLM. We vary whether Descriptive Intention Features (**D**), Granular Intention Features (**G**), and Job Detail Features (**J**) were used or not. **Pr.** is Precision and **Re.** is recall over the binary relevance scheme. *, $^-$, $^+$ indicates no statistically significant difference.

6 Cost and Turnaround Time Comparison

In this section we compare crowdsourced annotation and LLM labeling in terms of cost and turnaround time as shown in Table 5. For crowdsourced annotations we consider the cost and turnaround time to get 20000 query, job pairs. For labeling with GPT-4o, we consider the cost based on OpenAI's Batch API. For labeling with the Qwen models we utilized vLLM framework (Kwon et al. 2023) and compute cost based on the hourly rate of using a machine with 4 NVIDIA A10G GPUs ($0.80 per hour) and the number of hours it takes to run the inference on the same 20k Query, Job Pairs. Please note with efficient batching and other optimizations this cost can be optimized further.

	crowdsourced	**GPT-4o**	**Qwen-32-Instruct**
Cost	$9000	$80	$6.5
Time	2 weeks	3 - 24 hours	8 hours

Table 5: Cost and Turnaround time comparison for crowdsourced and LLM labeling of 20k Query, Job Pairs.

7 Understanding and Checking for Biases

We end by checking whether either of our prompts (for either closed source or open source LLMs) display biases in their outputs for a certain characteristic of their input. Based on previous works on data labeling using LLMs (Faggioli et al. 2023), (Thomas et al. 2024), (Pavlovic and Poesio 2024), we know that this step can highlight potential pitfalls.

Effect of Document Length In (Thomas et al. 2024), the authors found that longer prompts correlated with more positive relevance labels. We checked for this effect by measuring the correlation between a job's description and the output signed error (+1 or -1 when LLM over or underestimates relevance and 0 otherwise). Our best performing closed source LLM shows no statistically significant correlation (at the 95% level) between job description length and signed error. However, our best performing open source LLM shows a small but statistically significant positive correlation (95% CI [0.01, 0.07]) between job description length and the tendency for the open source prompt to overestimate the true relevance.

Effect of LLM Confidence Scores on True Relevance We next check to see whether the LLMs are systematically more confident on positive or negative signed errors. Therefore, we measure the correlation between the LLM's confidence score on the output token (obtained by exponentiating the log probability outputted by the LLM) and the signed error. Our best performing closed-source LLM shows a small yet statistically significant positive correlation for both binary (95% CI [0.11, 0.15]) and quinary (95% CI [0.20, 0.23]) labeling scheme. However, open-source LLM prompt only had a small yet statistically significant positive correlation for quinary (95% CI [0.02, 0.09]) labeling scheme. While some of these correlations are statistically significant and positive, we ruled them low enough not to warrant immediate action. Still, it is important to be aware of such biases when solving data labeling problems using LLMs.

Consistency Based on two runs, re-testing inputs with the best-performing prompts showed consistent outputs across both closed and open-source models. While this suggests reliability, additional runs are recommended for thorough validation.

8 Conclusion

This work delves into the integration of LLMs into relevance labeling workflows for job search systems, addressing challenges of cost-effectiveness, scalability, and quality. By carefully designing prompts, we demonstrate that GPT-4o can generate relevance labels comparable to human annotators, significantly reducing costs and turnaround time, achieving substantial agreement with crowdsourced labels on binary and ternary relevance schemes. Qwen2.5-32B-Instruct, on the other hand rely on engineered descriptive intention feature to achieve competitive performance with closed-source models on binary relevance label. However, the poor performance when predicting granular quinary grades and the small yet significant biases in LLM predictions such as overestimations influenced by job description length emphasize the need for ongoing refinement and monitoring.

Our results showcase how promising this approach can be for reducing the cost and turnaround time for large scale relevance labeling task in Job search systems. Future work should focus on systematically utilizing LLM-generated explanations, enhancing bias mitigation strategies, and exploring ways to further close the gap in performance between open-source and closed-source models.

References

Faggioli, G.; Dietz, L.; Clarke, C. L. A.; Demartini, G.; Hagen, M.; Hauff, C.; Kando, N.; Kanoulas, E.; Potthast, M.; Stein, B.; and Wachsmuth, H. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '23, 39–50. ACM.

He, X.; Lin, Z.; Gong, Y.; Jin, A.-L.; Zhang, H.; Lin, C.; Jiao, J.; Yiu, S. M.; Duan, N.; and Chen, W. 2024. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. arXiv:2303.16854.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3): 276–282.

OpenAI. 2023. GPT-4 Omni: OpenAI's Advanced Multimodal Model.

Pavlovic, M.; and Poesio, M. 2024. The Effectiveness of LLMs as Annotators: A Comparative Overview and Empirical Analysis of Direct Representation. In Abercrombie, G.; Basile, V.; Bernadi, D.; Dudy, S.; Frenda, S.; Havens, L.; and Tonelli, S., eds., *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, 100–110. Torino, Italia: ELRA and ICCL.

Team, Q. 2024. Qwen2.5: A Party of Foundation Models.

Thomas, P.; Spielman, S.; Craswell, N.; and Mitra, B. 2024. Large Language Models Can Accurately Predict Searcher Preferences. In *2024 International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. An earlier version of this paper appeared as arXiv preprint arXiv:2309.10621v1 [cs.IR].

Wang, X.; Kim, H.; Rahman, S.; Mitra, K.; and Miao, Z. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Zhuang, H.; Qin, Z.; Hui, K.; Wu, J.; Yan, L.; Wang, X.; and Bendersky, M. 2024. Beyond Yes and No: Improving Zero-Shot LLM Rankers via Scoring Fine-Grained Relevance Labels. arXiv:2310.14122.