International Journal on Natural Language Computing (IJNLC) Vol.13, No.2, April 2024

Identifying Key Terms in Prompts for Relevance Evaluation with GPT Models

Jaekeol Choi

Division of AI Data Convergence, Hankuk University of Foreign Studies, Seoul, South Korea

Abstract. Relevance evaluation of a query and a passage is essential in Information Retrieval (IR). Recently, numerous studies have been conducted on tasks related to relevance judgment using Large Language Models (LLMs) such as GPT-4, demonstrating significant improvements. However, the efficacy of LLMs is considerably influenced by the design of the prompt. The purpose of this paper is to identify which specific terms in prompts positively or negatively impact relevance evaluation with LLMs. We employed two types of prompts: those used in previous research and generated automatically by LLMs. By comparing the performance of these prompts in both few-shot and zero-shot settings, we analyze the influence of specific terms in the prompts. We have observed two main findings from our study. First, we discovered that prompts using the term 'answer' lead to more effective relevance evaluations than those using 'relevant.' This indicates that a more direct approach, focusing on answering the query, tends to enhance performance. Second, we noted the importance of appropriately balancing the scope of 'relevance.' While the term 'relevant' can extend the scope too broadly, resulting in less precise evaluations, an optimal balance in defining relevance is crucial for accurate assessments. The inclusion of few-shot examples helps in more precisely defining this balance. By providing clearer contexts for the term 'relevance,' few-shot examples contribute to refine relevance criteria. In conclusion, our study highlights the significance of carefully selecting terms in prompts for relevance evaluation with LLMs.

Keywords: chatGPT, GPT-3.5, GPT-4, Information Retrieval, Large Language Models (LLMs), relevance evaluation, prompt engineering, passage ranking.

1 Introduction

Ranking models are foundational in the domain of Information Retrieval (IR). Their success relies heavily on relevant sets that are used as standards during both training and testing stages. Traditionally, crowd-sourced human assessors have been used for relevance judgement, as indicated by several studies [1, 2]. However, this method is often time-consuming, expensive, and can yield inconsistent results due to the inherent subjectivity of human judgement [3, 4].

As technology keeps advancing, diverse machine learning techniques have stepped into the realm of relevance judgment [5, 1, 6, 7]. Driven by sophisticated algorithms,

DOI: 10.5121/ijnlc.2024.13201

these methods tried to replicate or even enhance the human ability to discern relevance within vast information collections. Despite their potential, there remains skepticism among researchers about whether these techniques can match human accuracy and reliability in relevance judgment.

The major change came about with the advent of LLMs, notably GPT-3 and GPT-4. With their large architectures and extensive training datasets, these LLMs brought the possibility of automated relevance judgments. The performance of these models across diverse natural language processing tasks has fostered a renewed belief in the ability of machines to evaluate passage relevance accurately. Encouraged by this paradigm shift, a couple of relevance judgment [8, 9] and ranking models [10] rooted in GPT architectures have been proposed. These models have demonstrated exceptional performance, often equaling or surpassing traditional methods.

However, The accuracy and robustness of relevance assessment using LLMs are significantly influenced by the prompts employed during the evaluation [11, 12]. These prompts serve as critical guides, aligning the model's responses with the user's intent. Consequently, prompt formulation becomes a pivotal component, demanding careful design and optimization.

In this paper, we primarily focus on the prompts used for relevance evaluation in GPT models, particularly examining which terms in the prompts are beneficial or detrimental to performance. We investigate how the performance of LLMs varies with the use of different types of prompts: those utilized in previous research and those generated by LLMs. Our aim is to identify which terms in the prompts improve or impair the performance in relevance assessment tasks. To provide a comprehensive understanding, we conduct these experiments in both few-shot and zero-shot settings.

This study concludes that the term 'answer' in prompt design is notably more effective than 'relevant' for relevance evaluation tasks using LLMs. This finding emphasizes the importance of a well-calibrated approach to defining relevance. While 'relevant' broadly encompasses various aspects of the query-passage relationship, 'answer' more directly targets the core of the query, leading to more precise and effective evaluations. Therefore, balancing the scope of 'relevance' in prompt design is crucial for enhancing the efficiency and accuracy of LLMs in relevance assessment.

The rest of this paper is organized as follows: '2 Related Work' delves into the background and previous studies. '3 Methodology' outlines the methods and approaches used in our study, including the details of the LLMs and the dataset. '4 Experimental Results' presents the findings from our experiments, providing a comprehensive analysis of the performance of different prompts. '5 Discussion' explore the implications of our findings. Finally, '6 Conclusions' summarizes the key insights from our study.

2 Related Works

The field of IR has seen a significant evolution with the advent of advanced machine learning models and techniques. This section reviews the relevant literature, focusing on the development of relevance judgment methods in IR and the role of prompt engineering in the effective utilization of LLMs.

2.1 Relevance Judgement in Information Retrieval

The relevance evaluation between a query and a passage has been a fundamental task since the inception of ranking systems. This assessment has historically been conducted in a binary manner, categorizing results as either relevant or non-relevant, but has evolved to include graded relevance scales offering more detailed evaluations.

In the realm of traditional IR, the reliance on human assessors for relevance judgment has been extensively documented [1, 2]. Despite their ability to provide nuanced evaluations, this approach has been criticized for its time and cost inefficiencies, as well as the subjective variability in results it can produce [3, 4].

The advancement of machine learning and its integration into IR has marked a transition towards automated relevance judgment. This area, particularly the use of transformer-based models like BERT, has been the focus of recent research [7]. The challenge, however, lies in achieving a balance between the precision offered by human assessment and the scalability of automated methods.

The introduction of LLMs, especially GPT-3 and GPT-4, has further transformed the landscape of relevance judgment. Initial studies, such as those by [13] and [8], explored the use of GPT-3 in annotation tasks, including relevance judgment. [10]'s research extends this to examining GPT-3's broader capabilities in data annotation. In a distinct approach, [14] investigated the use of LLMs for evaluating unassessed documents, aiming to improve the consistency and trustworthiness of these evaluations. Complementing this, [12] delved into the integration of LLMs for comprehensive relevance tagging, highlighting their comparable precision to human annotators. On the contrary, [9] has presented theoretical concerns regarding the exclusive use of GPT models for independent relevance judgment.

While extensive research has been conducted in this field, the specific influence of terms within a prompt on relevance evaluation remains unexplored. This study seeks to bridge this gap by investigating the impact of individual terms used in prompts.

2.2 Few-shot and Zero-shot Approaches

Recent advancements in LLMs have emphasized their capability for in-context learning, classified as either few-shot or zero-shot based on the presence of in-context

examples. Few-shot learning, where a model is given a limited set of examples, has historically shown superior performance over zero-shot learning, which relies on instructions without examples, as highlighted by [15].

The "pre-train and prompt" paradigm emphasizes the distinction between few-shot prompts (conditioned on task examples) and zero-shot prompts (template-only). While few-shot learning was traditionally favored, recent studies, including those on GPT-4, suggest that zero-shot approaches can sometimes outperform few-shot methods, particularly in specific domains [16, 17].

In our study, to investigate the terms in prompts, we conduct experiments using both few-shot and zero-shot settings and compare their outcomes.

2.3 Advanced in Prompt Engineering

Prompt engineering has emerged as a critical factor in harnessing the full potential of LLMs across various natural language processing applications. The formulation of a prompt is instrumental in guiding an LLM's output, significantly influencing its performance in diverse tasks [18, 15]. The art of crafting effective prompts involves meticulous design and strategic engineering, ensuring that prompts are precise and contextually relevant [19, 20, 21].

The increasing complexity of LLMs has spurred interest in developing sophisticated prompt tuning methods. These methods often utilize gradient-based approaches to optimize prompts over a continuous space, aiming for maximal efficiency and efficacy [22, 23]. However, the practical application of these methods can be limited due to constraints such as restricted access to the models' gradients, particularly when using API-based models. This challenge has led to the exploration of discrete prompt search techniques, including prompt generation [24], scoring [25], and paraphrasing [26].

In the broader context of prompt-learning, or "prompting," the approach is increasingly recognized as a frontier in natural language processing, seamlessly bridging the gap between the pre-training and fine-tuning phases of model development [27, 28]. This technique is particularly valuable in low-data environments, where conventional training methods may be less effective [29, 30, 31].

Within the realm of prompt-learning, two primary strategies are employed: few-shot and zero-shot learning. [32] demonstrated a few-shot technique for generating relevance, while studies like those by [10] and [33] have successfully applied few-shot learning in various scenarios. Conversely, [28] suggested that with an appropriate template, zero-shot prompt-learning could yield results surpassing those of extensive fine-tuning, emphasizing the power and flexibility of well-engineered prompts.

So far, there has been little focus on the terms within a prompt in existing research. This study is important because even small changes in a prompt can lead to different results. Our research, which concentrates on word terms, can be considered a form of micro-level prompt engineering.

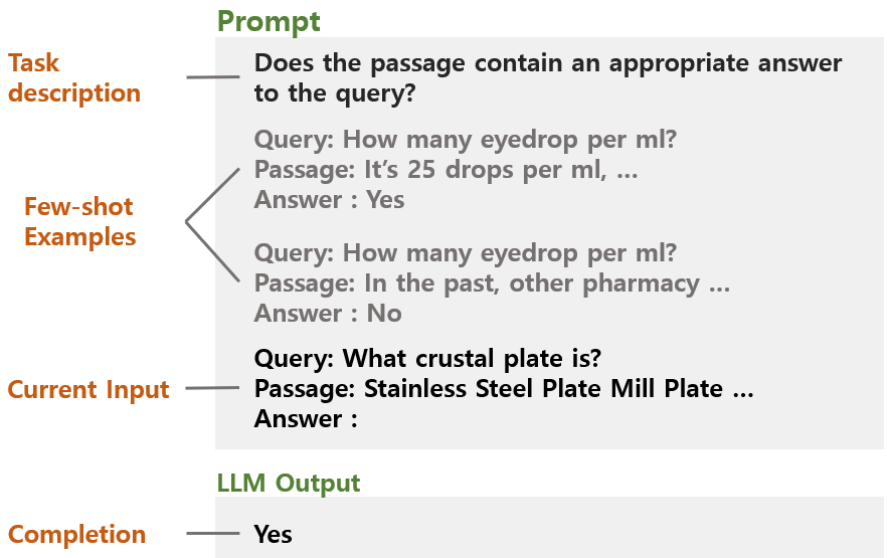

Fig. 1. A prompt example for relevance evaluation. This example utilizes 2-shot examples.

3 Methodology

Prompts for relevance evaluation, as shown in Figure 1, include an instruction to guide the LLM, in-context few-shot examples for clarity, and an input as the target task. Using these elements, LLMs generate the corresponding output. We apply this template in conduting our experiments for finding out which terms in prompts have impact on the performacne.

3.1 Evaluation method

To evaluate the effectiveness of each prompt in the relevance evaluation task, an objective metric is required. For this purpose, we decided to use the similarity between the evaluations conducted by humans and those conducted by the LLM using the prompt. To measure the similarity between the two sets of evaluations, we utilize Cohen's kappa (κ) coefficient, a statistical measure for inter-rater reliability that accounts for chance agreement. This measure compares the agreement between relevance labels generated by the LLM and human judgments, reflecting the quality of the prompt. Higher kappa values indicate a stronger alignment between the LLM and human evaluations. The Cohen's kappa coefficient is calculated using the following formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

Table 1. Templates used for generating and analyzing by LLMs.

Usage	Template for generating prompts
Generation	Instruction: When given a query, a passage, and a few examples, generate a prompt that can make an output from the given input. Example 1 - Input:[query,passage]\nOutput:[Yes/No] Example 2 - Input:[query,passage]\nOutput:[Yes/No] ... Generate prompt:
Analysis	Instruction: Which terms are common in these prompts that have a key role to evaluate relevance? Prompt 1: [Prompt] Prompt 2: [Promtp] ... Find terms :

In this equation, P_o represents the observed agreement between the two sets of evaluations, and P_e is the expected agreement by chance. The kappa value ranges from -1 to 1, where 1 indicates perfect agreement, 0 no agreement other than what would be expected by chance, and -1 indicates total disagreement. A higher kappa value suggests that the LLM's relevance evaluations are more closely aligned with human assessments, indicating a higher quality of the prompt in guiding the LLM to make evaluations similar to those of human judges.

3.2 Prompts and Few-shot Examples

We utilize two types of prompts, as shown in Table 7 of Appendix B. The first type consists of prompts named with an 'M', sourced from previous research [32, 10, 9]. The second type includes prompts generated using the template in Table 1, which are named with a 'G'. After assessing the performance of both prompt types, we aim to determine which prompts perform better. Following the experiments, we will analyze whether there are any terms common to the more effective prompts. If common terms are identified, it would suggest that these terms play a crucial role in the effectiveness of the prompt.

We conduct the experiments under both zero-shot and few-shot settings. Few-shot examples, derived from [9], are illustrated in Table 6 of Appendix A. These few-shot examples consist of four instances: two are positive examples, and the other two are negative ones. To ensure a fair comparison, we apply the same set of few-shot examples across all prompts.

Table 2. Overview of the TREC DL Passage datasets utilized in the study. The datasets from 2019 to 2021 are used for evaluating the performance of prompts. The table details the year of the dataset, the number of queries, the total number of query relevance judgments (qrels), and the number of sampled qrels used in the study.

Usage	TREC DL year	Number of queries	Number of qrels	Number of sampeld qrels
Evaluation	2019	43	9,260	200
	2020	54	11,386	200
	2021	53	10,828	200

3.3 Analysis

We analyze which terms are beneficial for relevance evaluation. Initially, we compare the performance of the prompts illustrated in Table 7. We then categorize the prompts into those with high performance and those with lower performance and look for distinguishing characteristics in each group. To identify the specific terms that play a role, we utilize the analysis prompts provided in Table 1. Furthermore, we compare how the results of each group vary depending on the presence or absence of few-shot examples.

We advance our analysis by constructing confusion matrices for the prompts, allowing for a more in-depth evaluation of their impact. Through the examination of precision and recall values derived from these matrices, we gain insights into the roles played by different terms within the context of relevance evaluation.

4 Experimental Result

We presents the results of our experimental investigation into the effectiveness of various prompts in relevance evaluation tasks using LLMs. We detail the experimental setup, including the models and datasets used, and then delve into the outcomes of our experiments. These results provide crucial insights into how different prompt designs and key terms influence the performance of LLMs in relevance judgment tasks.

4.1 Experimental Setup

Large Language Models For our experiments, we utilize GPT-3.5-turbo and GPT-4, both accessed via OpenAI's APIs. GPT-3.5-turbo, with its 178 billion parameters, enhances user interaction by providing clearer and more precise answers. As the most advanced in the series, GPT-4 has 1.76 trillion parameters and outperforms its predecessors in processing and contextual understanding.

Table 3. Comparative Results of Relevance Evaluation in Zero-shot and Few-shot Settings: This table presents the performance of various prompts under zero-shot and few-shot scenarios. The top five performing prompts are highlighted in bold, while the bottom five are underlined. We provide the respective average performances for these groups in both GPT-3.5-turbo and GPT-4 models. A '*' symbol denotes a significant difference at the 95% confidence level.

Type	Name	Zero-shot		Few-shot	
		GPT-3.5-turbo	GPT-4	GPT-3.5-turbo	GPT-4
Manual	**M1**	**0.389** (±0.115)	**0.450** (±0.090)	**0.339** (±0.059)	**0.471** (±0.041)
	M2	0.326 (±0.032)	0.426 (±0.061)	0.274 (±0.064)	0.437 (±0.046)
	M3	0.319 (±0.033)	0.396 (±0.086)	0.330 (±0.025)	0.460 (±0.046)
	M4	0.204 (±0.019)	0.344 (±0.073)	0.310 (±0.041)	0.433 (±0.028)
Generated	G1	0.301 (±0.046)	0.209 (±0.116)	0.309 (±0.052)	0.408 (±0.029)
	G2	0.356 (±0.064)	0.384 (±0.099)	0.315 (±0.033)	0.425 (±0.050)
	G3	0.279 (±0.044)	0.424 (±0.060)	0.303 (±0.026)	0.427 (±0.067)
	G4	0.268 (±0.053)	0.426 (±0.082)	0.312 (±0.017)	0.432 (±0.054)
	G5	0.342 (±0.007)	0.429 (±0.101)	0.257 (±0.031)	0.461 (±0.071)
	G6	0.363 (±0.085)	**0.462** (±0.073)	**0.333** (±0.073)	**0.472** (±0.046)
	G7	**0.393** (±0.074)	**0.450** (±0.066)	**0.379** (±0.042)	**0.464** (±0.051)
	G8	**0.382** (±0.075)	**0.455** (±0.084)	**0.349** (±0.066)	**0.463** (±0.039)
	G9	**0.398** (±0.089)	**0.443** (±0.074)	**0.351** (±0.078)	**0.468** (±0.046)
	G10	**0.366** (±0.086)	0.442 (±0.074)	0.327 (±0.050)	0.445 (±0.055)
Top-5 average		0.386 (±0.013)*	0.452 (±0.007)*	0.352 (±0.018)*	0.468 (±0.004)*
Bottom-5 average		0.274 (±0.044)	0.351 (±0.084)	0.291 (±0.024)	0.425 (±0.010)

Dataset For our experiments, we utilize the test sets from the MS MARCO TREC DL Passage datasets spanning three years[1]. As depicted in Table 2, We randomly sampled 200 data points from each year's test dataset, ensuring every query in the full set is included. These sampled datasets are then used to evaluate the prompts.

Relevance in these dataset is rated on a 4-point scale: "Perfectly relevant," "Highly relevant," "Related," and "Irrelevant."

For binary classification tasks, we simplify this 4-point relevance scale to a binary "Yes" or "No" judgment. Specifically, the categories of "Perfectly relevant" and "Highly relevant" are consolidated into a "Yes" category to indicate relevance, while "Related" and "Irrelevant" is classified as "No."

4.2 Relevance Evaluation Result of Prompts

The evaluation of prompt efficacy in relevance assessments, as outlined in Table 3, reveals notable trends. A key observation is the significant performance variation among semantically similar prompts, highlighting the impact of subtle differences

[1] https://microsoft.github.io/msmarco/TREC-Deep-Learning-2019
https://microsoft.github.io/msmarco/TREC-Deep-Learning-2020
https://microsoft.github.io/msmarco/TREC-Deep-Learning-2021

in prompt design on evaluation outcomes. For example, although M3 and G3 are similar prompts asking if the query and passage are 'relevant,' they yield different results. Moreover, despite all prompts addressing the relevance between the query and passage, their outcomes vary substantially.

When comparing results between GPT-3.5 and GPT-4 across both few-shot and zero-shot settings, Prompts M1, G7, G8, and G9 consistently rank in the top five across both GPT-3.5-turbo and GPT-4, indicating their inherent effectiveness. Conversely, certain prompts consistently underperform in both models. Specifically, prompts M4, G1, and G3 are found in the bottom five, underscoring elements that may detract from the efficacy of relevance evaluations.

Examining the performance of individual models reveals distinct characteristics in response to the prompts. Each model demonstrates unique preferences in prompt efficacy, illustrating that LLMs may respond differently to the same prompt structures. Certain prompts show high efficacy in GPT-3.5-turbo, while others perform better in GPT-4. Notably, GPT-4 generally exhibits superior performance compared to GPT-3.5-turbo across a range of prompts. A particular case of interest is prompt G1 in the zero-shot setting, where GPT-4's performance is the only instance of falling behind GPT-3.5-turbo. Aside from this case, GPT-4's performance is generally superior to that of GPT-3.5-turbo.

Further statistical analysis, involving a paired t-test on the averages of the top five and bottom five prompts, reinforces these findings. Specifically, the top five prompts in GPT-3.5-turbo had an average performance of 0.386, while in GPT-4, this average was higher at 0.452. Conversely, the bottom five prompts averaged 0.274 in GPT-3.5-turbo and 0.351 in GPT-4. These results indicate a statistically significant difference in performance at a 95% confidence level, emphasizing the pivotal role of prompt design in influencing the effectiveness of relevance evaluations in LLMs.

4.3 Analysis of Terms in prompts

In our analysis, we utilized the template from Table 1 to identify key terms in prompts that play a significant role in relevance evaluation using LLMs. The findings are summarized in Table 4.

We observed that prompts demonstrating top performance commonly used the term 'answer' or its variations. For instance, in M1, the prompt asks if the passage 'answers' the query. Similarly, G7 and G9 emphasize whether the passage contains or directly 'answers' the query. This pattern is also evident in G10, where the prompt focuses on whether the passage 'correctly answers' the query.

On the other hand, prompts associated with lower performance frequently included the term 'relevant' or related terms. For example, M3's prompt requires indicating if the passage is 'relevant' for the query, while G1 asks if the query and

Table 4. Key terms that have an crucial role. In prompts demonstrating good performance, the term 'answer' is commonly used, whereas in prompts indicating low performance, the term 'relevant' is commonly used.

Efficacy	Key Term	Prompt
High	**Answer**	G9: ... if the passage provides a direct **answer** to ... G7: ... the passage contains the **answer** to the query ... M1: Does the passage **answer** the query? ... G10: Determine if the passage correctly **answers** to ...
Low	**Relevant**	G1: Do the query and passage **relate** to the same topic.. M4: 2 = highly **relevant**, very helpful for ... M3: Indicate if the passage is **relevant** for the query? ... G3: In the context of the query, is the passage **relevant**?

passage 'relate' to the same topic. This trend continues in M4 and G3, where the term 'relevant' is central to the prompt's structure.

These findings indicate that the choice of key terms in prompts significantly impacts the performance of LLMs in relevance evaluation tasks. Terms like 'answer' seem to guide the LLM towards more effective evaluation, while the use of 'relevant' appears to be less conducive for this purpose.

4.4 Analysis of Zero-shot and Few-shot Results

The differences in performance between zero-shot and few-shot models for GPT-3.5-turbo and GPT-4 are illustrated in Figure 2, which presents the average results for each approach. From this analysis, we can discern two interesting observations.

Firstly, there is a notable variation in performance across the top and bottom five performers between the two model versions. In the case of GPT-3.5-turbo, while there is an improvement in the performance of the bottom five prompts (from an average of 0.274 in zero-shot to 0.291 in few-shot), the top five prompts exhibit a decrease in performance (from 0.386 in zero-shot to 0.352 in few-shot). This indicates that while few-shot examples enhance GPT's ability to handle previously lower-performing prompts, they might detrimentally affect the performance of the highest-performing prompts.

In contrast, GPT-4 shows a consistent improvement in both the top and bottom performers with few-shot examples. The top five prompts improve from an average of 0.452 in zero-shot to 0.468 in few-shot, and the bottom five improve from 0.351 to 0.425. This shows that few-shot examples enhance the overall performance in evaluation tasks with GPT-4.

Secondly, both models demonstrate a reduction in the performance gap between the top and bottom five prompts with few-shot learning. This convergence is more pronounced in GPT-4, which sees a more significant increase in performance for the bottom five prompts. It suggests that few-shot examples is particularly effective in

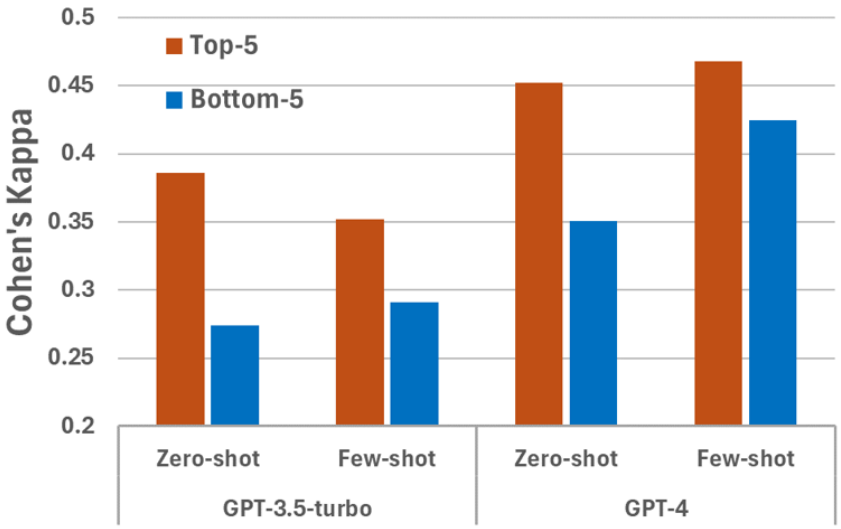

Fig. 2. Average Cohen's kappa values for top-5 and bottom-5 prompts in GPT-3.5-turbo and GPT-4 across few-shot and zero-shot settings.

refining the model's responses to less optimal prompts, leading to a more consistent performance across different types of prompts.

Given the role of few-shot examples in providing clearer instructions and context, these results suggest that GPT-4 is more adept at adapting to varied prompt structures and content than GPT-3.5-turbo.

5 Discussion

This section offers an analysis of our experimental results, focusing on the impact of specific prompt terms on the performance of LLMs in relevance evaluation. We also discuss the potential and challenges of using LLMs as direct rankers in IR, compared to their current role in generating relevance judgments.

5.1 Why 'Answer' Is Better Than 'Relevant'

The analysis of confusion matrices in Table 5 provides key insights into the effectiveness of different prompt types in relevance evaluation. This analysis highlights G6, which had the highest performance, G1 with the lowest performance, and G10, known for its use of the term 'correctly.'

G6, achieving the highest performance, questions if the passage provides 'an answer' to the query. This prompt led to a significant agreement between LLM predictions and human assessors, as evident by a high Cohen's kappa value of

Table 5. Confusion Matrices for three prompts using the TREC DL 2021 test set in a zero-shot setting. This table includes Cohen's kappa values, along with calculated precision and recall. The analysis focuses on G6 with the highest performance, G1 with the lowest, and G10, which has the narrowest definition by using the term of 'correctly'.

Prompt	Prediction	Human assessors Relevant	Human assessors Irrelevant	Cohen's κ	Precision	Recall
G6	Relevant	**43**	24	**0.528**	0.641	0.716
	Irrelevant	17	**116**			
G1	Relevant	**59**	84	0.275	0.413	**0.983**
	Irrelevant	1	**56**			
G10	Relevant	**38**	20	0.495	**0.655**	0.633
	Irrelevant	22	**120**			

G6 : Given a query and a passage, determine if the passage provides **an answer** to the query. ...
G1 : Do the query and passage **relate** to the same topic? ...
G10 : Determine if the passage **correctly answers** a given query. ...

0.528, along with strong precision and recall. The high number of true positives (43) and true negatives (116) in G6's matrix suggests that focusing on 'answering' is highly effective in evaluating the relevance of the passage to the query.

Conversely, G1, which demonstrated the lowest performance, focuses on whether the query and passage 'relate' to the same topic. Despite its high recall, this prompt yielded a lower Cohen's kappa value of 0.275. The comparatively fewer true negatives (56) against G6 indicate that a broader 'relevance' focus may lead to less precise evaluations.

G10, with its emphasis on whether the passage 'correctly answers' the query, shows a distinct performance, marked by a Cohen's kappa value of 0.495. Its precision is notably high, but the recall is somewhat limited, suggesting that while it is effective in identifying specific relevant answers, it may overlook some broader aspects of relevance.

This comparison underlines the varying effectiveness of prompts based on their focus in the context of information retrieval. Prompts like G6, with an 'answering' focus, tend to lead to more accurate and precise evaluations, while 'relevance'-focused prompts like G1 might not capture the entire scope of the query-passage relationship. G10's specific focus on 'correctly answering' demonstrates a particular effectiveness in identifying precise answers but at the potential expense of broader relevance. Therefore, the choice of key terms and their emphasis is crucial in designing prompts for efficient retrieval and ranking in LLMs.

5.2 Balancing the Definition of 'Relevance'

As discussed in the previous section, defining 'relevance' in the context of LLM prompts varies significantly in its scope. G10's approach, using the term 'correctly answers', tends to give a slightly narrow definition in relevance evaluation. It focuses on whether the passage precisely addresses the query, potentially overlooking broader aspects of relevance.

On the other hand, we explored a more balanced approach with G6's prompt. This prompt, focusing on whether the passage provides 'an answer' to the query, strikes a middle ground. It covers not just the direct answer but also the broader context, leading to a more comprehensive consideration of relevance.

Conversely, G1's prompt offers the broadest definition of relevance by asking if the query and passage 'relate' to the same topic. This wide approach, while inclusive, risks being too expansive. As reflected in the confusion matrix for G1 in Table 5, this broad definition results in high recall but at the cost of lower precision, as it captures a wide net of potentially relevant information, including false positives.

This analysis highlights the need for a balanced definition of relevance in prompt design. While G1's broad approach increases recall, its precision suffers. G10's narrow focus may miss broader relevance aspects. In contrast, G6's approach appears to offer a more optimal balance. It captures a wide array of relevant information without being overly narrow or inclusive, leading to more accurate and balanced performance in relevance evaluations. These findings are pivotal for crafting prompts that precisely measure the relevance of information in LLM-based retrieval and ranking systems.

5.3 Influence of Few-shot Examples

As can be seen in Figure 2, in GPT-3.5-turbo, the performance of zero-shot is slightly higher than that of few-shot. In contrast, in GPT-4, the performance of few-shot exceeds that of zero-shot. This variation indicates that a conclusive determination of the relative impacts of few-shot and zero-shot approaches is complex and model-dependent.

However, there is a characteristic that appears consistently in both models: the use of few-shot examples reduces the performance gap between the top-5 and bottom-5 groups. In GPT-3.5-turbo, the gap decreased from 0.112 to 0.061, and in GPT-4, it nearly halved from 0.101 to 0.043. These results suggest that few-shot examples help in defining unclear aspects in the bottom-5 instructions. For instance, consider the case of the G1 prompt. In the zero-shot setting, GPT-4 shows a low performance of 0.209, but when few-shot examples are used, the performance dramatically increases to 0.409. This could indicate that while the term 'relate' in G1 has a broad meaning, the use of few-shot examples helps in clarifying its interpretation.

5.4 Direct Ranking vs. Relevance Judgment Using LLMs

An emerging area of interest is the potential for using LLMs directly as rankers in IR, rather than just for generating relevance judgments. However, the practical application of LLMs as direct rankers faces significant challenges, primarily due to efficiency concerns. Directly ranking with LLMs, especially when reliant on API calls, can be slow and costly, as it requires repeated, resource-intensive interactions with the model for each ranking task. This approach, therefore, becomes impractical for large-scale or real-time ranking applications.

Given these constraints, future research in this domain should consider the development and utilization of downloadable, standalone LLMs. Such models, once sufficiently advanced, could potentially be integrated directly into ranking systems, offering a more efficient and cost-effective solution compared to API-dependent models. This shift would allow for the direct application of LLMs in ranking tasks, potentially overcoming the limitations currently posed by API reliance. However, this path also necessitates further advancements in LLM technology to ensure these models can operate effectively and reliably in a standalone capacity.

6 Conclusions

In this paper, we have examined the nuances of prompt design in relevance evaluation tasks using Large Language Models such as GPT-3.5-turbo and GPT-4. Our research reveals the profound impact that specific terms within prompts have on the effectiveness of these models. Contrary to initial expectations, our findings indicate that prompts focusing on 'answering' the query are more effective than those emphasizing broader concepts of 'relevance.' This highlights the importance of precision in relevance assessments, where a direct answer often more closely aligns with the intended query-passage relationship.

Furthermore, our investigations into few-shot and zero-shot scenarios revealed contrasting impacts on model performance. We found that few-shot examples tend to enhance the performance of LLMs, particularly in GPT-4, by bridging performance gaps between differently functioning prompts.

Our study also underscores the need for a well-balanced definition of 'relevance' in prompt design. We observed that overly broad definitions, while helpful in increasing recall, can compromise precision. Conversely, narrowly defined prompts, though precise, risk missing broader relevance aspects, failing to capture a comprehensive relevance assessment. Therefore, striking the right balance in prompt design is crucial for enhancing the efficiency and accuracy of LLMs in relevance evaluation tasks.

In summary, this paper contributes to the field by providing new insights into optimizing LLMs for relevance evaluation tasks. These insights offer crucial guidelines for creating effective prompts, ensuring that LLM outputs align more accurately

Table 6. Four few-shot exsmples

#	Few-shot examples
1	Query: how many eye drops per ml Passage: Its 25 drops per ml, you guys are all wrong. If it is water, the standard was changed 15 - 20 years ago to make 20 drops = 1mL. The viscosity of most things is temperature dependent, so this would be at room temperature. Hope this helps. Answer: Yes
2	Query: how many eye drops per ml Passage: RE: How many eyedrops are there in a 10 ml bottle of Cosopt? My Kaiser pharmacy insists that 2 bottles should last me 100 days but I run out way before that time when I am using 4 drops per day.In the past other pharmacies have given me 3 10-ml bottles for 100 days.E: How many eyedrops are there in a 10 ml bottle of Cosopt? My Kaiser pharmacy insists that 2 bottles should last me 100 days but I run out way before that time when I am using 4 drops per day. Answer: No
3	Query: can you open a wells fargo account online Passage: You can transfer money to your checking account from other Wells Fargo. accounts through Wells Fargo Mobile Banking with the mobile app, online, at any. Wells Fargo ATM, or at a Wells Fargo branch. 1 Money in — deposits. Answer: No
4	Query: can you open a wells fargo account online Passage: You can open a Wells Fargo banking account from your home or even online. It is really easy to do, provided you have all of the appropriate documentation. Wells Fargo has so many bank account options that you will be sure to find one that works for you. They offer free checking accounts with free online banking. Answer: Yes

with nuanced, human-like relevance judgments. As LLM technology continues to evolve, understanding the subtleties of prompt design becomes increasingly important in natural language processing and information retrieval applications.

Acknowledgment

This work was supported by Hankuk University of Foreign Studies Research Fund of 2024.

Appendix

A Few-shot Exmaples

We utilize four few-shot exmaples for our experiments.

Table 7. List of 14 prompts used in the experiments, detailing their names and instructions.

	Name	Prompt instruction
Manual	M1	Does the passage answer the query? Respond with 'Yes' or 'No'.
	M2	Given a passage and a query, predict whether the passage includes an answer to the query by producing either "Yes" or "No".
	M3	Indicate if the passage is relevant for the query. Respond with "Yes" or "No".
	M4	You are a search quality rater evaluating the relevance of passages. Given a query and a passages, you must provide a score on an integer scale of 0 to 2 with the following meanings: 2 = highly relevant, very helpful for this query 1 = relevant, may be partly helpful but might contain other irrelevant content 0 = not relevant, should never be shown for this query
Generated	G1	Do the query and passage relate to the same topic? Respond with 'Yes' or 'No'.
	G2	Is the passage pertinent to the query? Indicate with 'Yes' or 'No'.
	G3	In the context of the query, is the passage relevant? Reply with 'Yes' or 'No'.
	G4	Would a user find the passage relevant to their query? Respond with 'Yes' or 'No'.
	G5	Does the passage contain information relevant to the query? Answer with 'Yes' or 'No'.
	G6	Given a query and a passage, determine if the passage provides an answer to the query. If the passage answers the query, respond with "Yes". If the passage does not answer the query, respond with "No".
	G7	Your task is to determine whether the passage contains the answer to the query or not. If the passage contains the answer to the query, your response should be 'Yes'. If the passage does not contain the answer, your response should be 'No'
	G8	Given a query and a passage, determine if the passage provides a satisfactory answer to the query. Respond with 'Yes' or 'No'.
	G9	Given a query and a passage, determine if the passage provides a direct answer to the query. Answer with 'Yes' or 'No'
	G10	Determine if the passage correctly answers a given query. Respond with 'Yes' or 'No'

B Prompts

We utilize 14 prompts for our experiments.

Bibliography

[1] Omar Alonso, Stefano Mizzaro, et al. Can we get rid of trec assessors? using mechanical turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, volume 15, page 16, 2009.

[2] Roi Blanco, Harry Halpin, Daniel M Herzig, Peter Mika, Jeffrey Pound, Henry S Thompson, and Thanh Tran Duc. Repeatable and reliable search system evaluation using crowdsourcing. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 923–932, 2011.

[3] Eddy Maddalena, Marco Basaldella, Dario De Nart, Dante Degl'Innocenti, Stefano Mizzaro, and Gianluca Demartini. Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 4, pages 129–138, 2016.

[4] Zahra Nouri, Henning Wachsmuth, and Gregor Engels. Mining crowdsourcing problems from discussion forums of workers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6264–6276, 2020.

[5] Ian Soboroff, Charles Nicholas, and Patrick Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 66–73, 2001.

[6] Ben Carterette, James Allan, and Ramesh Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 268–275, 2006.

[7] Laura Dietz, Shubham Chatterjee, Connor Lennox, Sumanta Kashyapi, Pooja Oza, and Ben Gamari. Wikimarks: Harvesting relevance benchmarks from wikipedia. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3003–3012, 2022.

[8] Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*, 2022.

[9] Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Perspectives on large language models for relevance judgment, 2023.

[10] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. Is chatgpt good at search? investigating large language models as re-ranking agent, 2023.

[11] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

[12] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences, 2023.

[13] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*, 2021.

[14] Sean MacAvaney and Luca Soldaini. One-shot labeling for automatic relevance estimation. *arXiv preprint arXiv:2302.11266*, 2023.

[15] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[16] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[17] OpenAI. Gpt-4 technical report, 2023.

[18] Timo Schick and Hinrich Schütze. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402, 2021.

[19] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[20] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

[21] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

[22] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 2023.

[23] Guanghui Qin and Jason Eisner. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*, 2021.

[24] Eyal Ben-David, Nadav Oved, and Roi Reichart. Pada: A prompt-based autoregressive approach for adaptation to unseen domains. *arXiv preprint arXiv:2102.12206*, 3, 2021.

[25] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.

[26] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

[27] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021.

[28] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998*, 2021.

[29] Teven Le Scao and Alexander M Rush. How many data points is a prompt worth? *arXiv preprint arXiv:2103.08493*, 2021.

[30] Chengxi Li, Feiyu Gao, Jiajun Bu, Lu Xu, Xiang Chen, Yu Gu, Zirui Shao, Qi Zheng, Ningyu Zhang, Yongpan Wang, et al. Sentiprompt: Sentiment knowledge enhanced prompt-tuning for aspect-based sentiment analysis. *arXiv preprint arXiv:2109.08306*, 2021.

[31] Chengwei Qin and Shafiq Joty. Lfpt5: A unified framework for lifelong few-shot language learning based on prompt tuning of t5. *arXiv preprint arXiv:2110.07298*, 2021.

[32] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[33] Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples, 2022.