Advances in Artificial Intelligence and Machine Learning; Research 4 (3) 2687-2702 Received 13-07-2024; Accepted 21-09-2024; Published 28-09-2024

Binary or Graded, Few-Shot or Zero-Shot: Prompt Design for GPTs in Relevance Evaluation

Jaekeol Choi jaekeol.choi@hufs.ac.kr
Division of AI Data Convergence
Hankuk University of Foreign Studies
Seoul, South Korea

Corresponding Author: Jaekeol Choi

Abstract

Evaluating the relevance between a query and a passage is a pivotal task in Information Retrieval (IR). Utilizing such relevance evaluations can assist in ranking as well as in the creation of datasets for training and testing. The recent advancements in Large Language Models (LLMs) like GPT-4 have contributed to performance enhancements across many natural language processing tasks. Specifically, in the IR domain, many studies are being conducted on tasks related to relevance judgment, showing notable improvements. However, the efficacy of LLMs is considerably influenced by the design of the prompt. Despite this significance, there is a lack of research on prompts specifically tailored for relevance evaluation. The proposed prompts for this evaluation can be categorized based on how they distinguish relevance (binary or graded) and their reliance on in-context examples (few-shot or zero-shot). In this study, we experimentally investigate these two dimensions to determine which configurations are most advantageous for relevance evaluation. Our findings, based on the GPT-4 model, demonstrate that graded prompts in a zero-shot format are more effective.

Keywords: GPT-4, LLM, Passage ranking, Prompt engineering, Relevance evaluation.

1. INTRODUCTION

Large Language Models (LLMs) hold the potential to advance the field of Information Retrieval (IR). There is a growing trend in IR to leverage LLMs in addressing various challenges [1, 2]. A pivotal challenge among these is relevance evaluation [3–5]. The process of determining the relevance between a user's query and a corresponding passage is at the heart of various IR tasks, integral to both IR ranking and evaluation systems.

The accuracy and robustness of relevance assessment using LLMs are significantly influenced by the prompts employed during the evaluation [5, 6]. They serve as guiding beacons for models, ensuring alignment between the model's response and the user's intent. Consequently, prompt formulation emerges as a critical task, necessitating rigorous design and optimization. Historically, approaches to prompt design have faced significant challenges. They often required intensive labor and yielded inconsistent results [7, 8]. Traditional methodologies predominantly relied on

Citation: Jaekeol Choi. Binary or Graded, Few-Shot or Zero-Shot: Prompt Design for GPTs in Relevance Evaluation. Advances in Artificial Intelligence and Machine Learning. 2024;4(3):156.

manual prompt engineering. While this allowed for custom tailoring of prompts to specific needs, it introduced an inherent degree of subjectivity. This subjectivity, when combined with the intricate nuances of language, frequently resulted in significant variability in model performance [8, 9]. Such inconsistencies, as highlighted by various studies, emphasize the need for a more understanding to prompt design.

The prompt for relevance evaluation comprises three key components as illustrated in TABLE 1: the Instruction phrase, relevant granularity, and in-context few-shot examples. The Instruction phrase represents the task definition, guiding the evaluation of relevance between the provided query and passage. Relevance between the query and passage can be represented in a binary format, such as 'yes' or 'no', or in a graded manner that ranges from 0 to 4 points. In-context few-shot examples are samples presented to the model with the aim of achieving superior results. For instance, in TABLE 1, Sun et al. [2], utilized a prompt with a binary output complemented by four few-shot examples. Conversely, Thomas et al. [5], employed a three-tier graded output with a zero-shot approach. The definition of these three components can influence the results, leading to diverse outcomes even for semantically similar prompts.

Table 1: Prompt instruction examples of the variations in relevance granularity and in-context few-shot examples.

Prompt instruction	Relevance granularity	In-context examples
Faggioli et al. [3] (Binary)	Binary	Zero-shot
Sun et al. [2]	Binary	Few-shot
Thomas et al. [5]	Graded	Zero-shot
Faggioli et al. [3] (Graded)	Graded	Few-shot

In this study, we aim to empirically determine the most effective way to utilize these two prompt elements for relevance evaluation. We first experiment to ascertain whether binary or graded granularity yields superior performance. Following that, we investigate which setting, either few-shot or zero-shot, is more advantageous. To conduct this experiment, it is imperative to first establish a benchmark for assessing the superiority of relevance evaluations. We determined that evaluations aligning closely with human judgments are more reliable. Thus, the standard we employ measures the similarity between the outcomes generated by each evaluated prompt and existing human evaluations.

Our experimental results showed that for GPT-4, the best performance was observed with a graded granularity in a zero-shot setting. In contrast, for GPT-3.5, a combination of binary granularity and zero-shot yielded superior performance. The performance relative to granularity varied depending on the model, but there was a consistent trend favoring zero-shot settings.

2. RELATED WORKS

2.1 Relevance Judgement in Information Retrieval

Relevance judgement, or determining the relevance of retrieved passages to a specific user query, has been an integral part of IR systems. Traditionally, crowd-sourced human assessors have been used for relevance judgement, as indicated by several studies [10, 11]. However, this method is often time-consuming, expensive, and can yield inconsistent results due to the inherent subjectivity of human judgement [12, 13]. The rise of advanced machine learning models has marked a shift towards automatic relevance judgement, a rapidly growing area of research [14]. Transformer-based models like BERT have been the focus of many studies for this task [15]. Even so, finding the right balance between the precision of human judgement and the scalability of automatic methods remains a challenge.

Recent work has begun to investigate the role of LLMs like ChatGPT and GPT-4 as relevance judgements evaluators. Ding et al. [16], Wang et al. [17], embarked on experimental endeavors to establish GPT-3's capabilities as an annotator. MacAvaney and Soldaini [4], probed into the application of LLMs in gauging the relevance of yet-to-be-assessed documents utilizing a one-shot strategy, aiming to enhance the consistency and trustworthiness of such evaluations. Additionally, Thomas et al. [5], looked into the integration of LLMs for extensive relevance tagging, underscoring their efficacy and matching human labelers in precision. On the flip side, Faggioli et al. [3], articulated profound theoretical concerns related to the sole reliance on GPT for generating relevance judgments autonomously.

In these research efforts, the prompt itself has not been extensively studied. Thus, our study concentrates on the prompt itself. We aim to empirically determine, through experimentation, which is more advantageous: binary or graded granularity, and which approach, few-shot or zero-shot, leads to superior results.

2.2 Binary and Graded Relevance

Relevance between a query and a document or passage can typically be classified as binary or graded. Historically, the prevalent evaluation metrics like precision, recall and NDCG have leaned towards binary relevance, making it a standard approach for gauging the efficacy of ranking models [18]. However, as the field has evolved, several studies have introduced metrics, such as Expected Reciprocal Rank (ERR) and Rank-Biased Precision (RBP), that harness graded evaluations for more nuanced insight [19–21].

The TREC DL test sets[1], widely employed for assessing ranking models, incorporate a 4-graded relevance categorization. This provides two potential pathways for evaluating relevance: one can either directly use a binary categorization or first employ a graded relevance system, subsequently converting the results into binary categories for model evaluation.

[1] https://microsoft.github.io/msmarco/TREC-Deep-Learning.html

These distinct approaches raise a pressing question: which is more effective when utilizing LLMs such as ChatGPT for relevance assessment? Our research is driven by this question, and we strive to empirically establish the superior method via comprehensive experimentation.

2.3 Few-Shot and Zero-Shot Approaches

Recent LLMs exhibit an intriguing ability to quickly adapt to a variety of tasks through in-context learning. Depending on how many in-context examples or demonstrations are provided within the prompt, we can categorize the approach as either few-shot or zero-shot.

Historically, the few-shot learning paradigm, where a limited set of examples are provided, has demonstrated superior performance over zero-shot learning, where the model is only given an instruction without any examples. This is aligned with the findings of Brown [22], who demonstrated that GPT-3, a 178 billion parameter model, performs better under few-shot conditions as compared to one-shot or zero-shot settings across various natural language processing tasks.

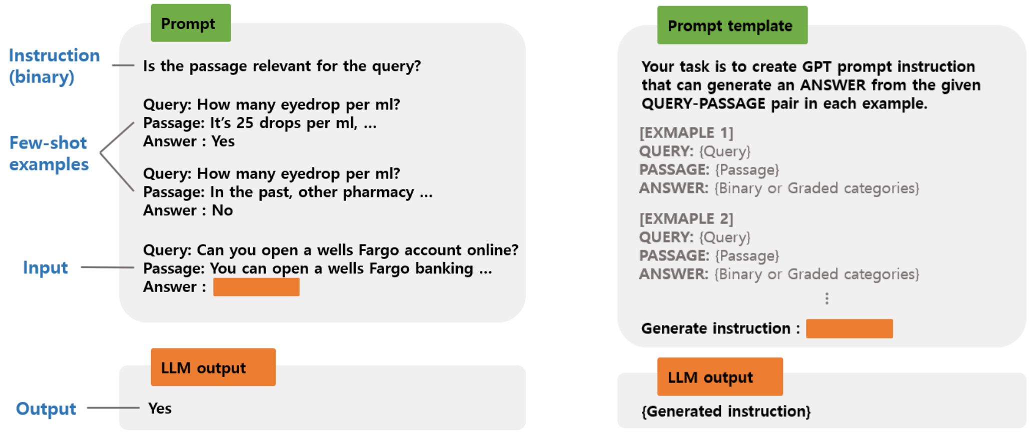
The “pre-train and prompt” paradigm, which has arisen with the proliferation of LLMs, has underscored the effectiveness of using prompts to guide model generation [23]. Here, the distinction between few-shot prompts (with explicit conditioning on several task examples) and zero-shot prompts (template-only without examples) becomes crucial.

However, as the field advances, there is a growing body of evidence that challenges this historically accepted advantage of few-shot learning. For instance, Kojima et al. [24], found scenarios where zero-shot learning can outperform its few-shot counterpart. Moreover, recent studies on GPT-4 have revealed that in some domains, it tends to excel even with zero-shot approaches [25].

In the context of relevance evaluation, the question remains: Which of these approaches, few-shot or zero-shot, yields superior results? Our research aims to shed light on this debate by experimentally comparing the efficacy of both methods.

3. METHODOLOGY

The prompts for relevance evaluation consist of three components, as illustrated in FIGURE 1 (a): instruction, few-shot examples, and input. The instruction acts as a task definition that guides LLMs in generating an output for a given input. Few-shot examples serve to clarify the instruction, aiding LLMs in grasping the context presented within the instruction. The input is the specific target that LLMs are tasked to address, utilizing the guidance from the instruction and few-shot examples. Employing prompts composed of these three elements, LLMs produce the output corresponding to the provided input. We apply this template in conducting two principal experiments: comparing binary versus graded relevance and exploring few-shot versus zero-shot approaches.



Is the passage relevant for the query?

Query: How many eyedrop per ml?
Passage: It's 25 drops per ml, ...
Answer : Yes

Query: How many eyedrop per ml?
Passage: In the past, other pharmacy ...
Answer : No

Query: Can you open a wells Fargo account online?
Passage: You can open a wells Fargo banking ...
Answer :

Yes

Your task is to create GPT prompt instruction that can generate an ANSWER from the given QUERY-PASSAGE pair in each example.

[EXMAPLE 1]
QUERY: {Query}
PASSAGE: {Passage}
ANSWER: {Binary or Graded categories}

[EXMAPLE 2]
QUERY: {Query}
PASSAGE: {Passage}
ANSWER: {Binary or Graded categories}
⋮

Generate instruction :

{Generated instruction}

(a) An example of a prompt for relevance evaluation (b) Prompt template for generating instructions

Figure 1: (a) This example employs binary relevance categories and includes 2-shot examples. Granularity is described in the instruction and further elucidated through the few-shot examples (b) Prompt template for generating instructions automatically by LLMs with sampled examples from dataset.)

3.1 Performance Evaluation Metric

To assess the efficacy of each prompt in the relevance evaluation task, we employ Cohen's kappa coefficient (κ), which is a statistical measure of inter-rater reliability [3, 5]. Cohen's κ evaluates the agreement between the judgments made by the human raters and the LLM, which is indicative of the prompt's quality. When we define the set evaluated by humans as $\mathsf{rel}_{\mathsf{truth}}$ and the relevance predicted by the LLM given a prompt as $\mathsf{rel}_{\mathsf{pred}}$, our goal is for $\mathsf{rel}_{\mathsf{pred}}$ to closely align with the ground truth $\mathsf{rel}_{\mathsf{truth}}$. To calculate the prompt's effectiveness using Cohen's κ, we define the score as follows:

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

where P_o is the observed agreement between the LLM's predictions $\mathsf{rel}_{\mathsf{pred}}$ and the human judgments $\mathsf{rel}_{\mathsf{truth}}$, and P_e is the hypothetical probability of chance agreement. To calculate P_o, the observed agreement, we measure the proportion of instances where both the human raters and the LLM predictions agree. In the binary case, this is the sum of the true positives and true negatives, divided by the total number of instances:

$$P_o = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\mathsf{rel}_{\mathsf{truth}\,i} = \mathsf{rel}_{\mathsf{pred}\,i}) \tag{2}$$

where $\mathbf{1}(\cdot)$ is the indicator function that returns 1 if the condition is true and 0 otherwise, and N is the total number of instances. To calculate P_e, the expected agreement by chance, we use the marginal probabilities of the human judgments and LLM predictions. For the binary case, we calculate P_e as:

$$P_e = \sum_{j\in\{0,1\}} \left(\frac{\sum_{i=1}^{N} \mathbf{1}(\mathsf{rel}_{\mathsf{truth}\,i} = j)}{N} \times \frac{\sum_{i=1}^{N} \mathbf{1}(\mathsf{rel}_{\mathsf{pred}\,i} = j)}{N} \right) \tag{3}$$

Using this measure, we assess the performance of the given prompt, with a higher κ value indicating a stronger agreement between the LLM and human evaluations.

For prompts resulting in graded relevance, we convert the outcomes to binary by considering 'Perfectly Relevant', 'Highly Relevant' as relevant, and 'Related' and 'Irrelevant' as not relevant, before calculating Cohen's κ.

3.2 Binary and Graded Instructions

We first use four prompts detailed in TABLE 1, that were employed in prior studies [2, 3, 5]. Two prompts of these pertain to binary categories, while the other two address graded categories. Recognizing that four prompts may not yield sufficiently robust research results, we secondly generate four additional prompts automatically using LLMs. We adopt the instruction generation method proposed by [1, 26], which leverages LLMs to craft instructions from sampled examples within a dataset. FIGURE 1(b) presents the template we employ for deriving prompt instructions. With this template, we generate two binary instructions and two graded prompts, resulting in four generated instructions.

In total, we have eight prompts: four derived from previous research and four generated by LLMs. The complete set of prompts used in this study is summarized in TABLE 2, for comparison.

3.3 Few-Shot and Zero-Shot Approaches

Eight instructions listed in TABLE 2, are utilized for both zero-shot and few-shot approaches. For a comparative analysis between few-shot and zero-shot, we employ the same instructions; however, the key difference lies in the inclusion or exclusion of few-shot examples.

To ensure a fair comparison, we consider that the number of few-shot examples may influence the performance of each prompt. To calibrate this effect, we experiment with varying quantities of few-shot examples: 2, 4, 8, 16, and 32, all of which are randomly selected from the dataset. The experiment is conducted twice, each time with a different set of few-shot examples, and the results are averaged to mitigate any variability stemming from the example selection.

4. EXPERIMENTAL SETUP

4.1 Large Language Models

For our experiments, we utilize five models depicted in Table 3, from OpenAI's GPT series, accessed via OpenAI's APIs. Each model offers unique capabilities due to its varying size and specialized training. GPT-4 model among them has the largest parameter size, 1.76 trillion.

Table 2: Eight instructions for the comparison experiments: four instructions with names starting with ‘M’ are from previous research, and four with names starting with ‘G’ are generated by GPT-4.

Name	Prompt instruction	Granularity
M1	Given a passage and a query, predict whether the passage includes an answer to the query by producing either “Yes” or “No.” Does the passgae answer the query?	Binary
M2	Indicate if the passage is relevant fort the question. Respond with ‘Yes’ or ‘No’.	Binary
G1	You are asked to determine if a given passage answers a specific query. You will be provided with a query and a passage. Your task is to analyze the passage and decide whether it provides a satisfactory answer to the query. If the passage answers the query, respond with ‘Yes’. If the passage does not answer the query, respond with ‘No’.	Binary
G2	Given a query and a passage, determine if the passage provides an answer to the query. If the passage contains the answer to the query, respond with ‘Yes’. If the passage does not contain the answer to the query, respond with ‘No’.	Binary
M3	Indicate if the passage answers the question. Respond with ‘Perfectly relevant’, ‘Highly relevant’, ‘Related’, or ‘Irrelevant’.	Graded
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 Assume that you are writing a report on the subject of the topic. If you would use any of the information contained in the passage in such a report, mark it 1. If the passage is primarily about the topic, or contains vital information about the topic, mark it 2. Otherwise, mark it 0.	Graded
G3	Given a ‘Question’ and a ‘Passage’, determine the relevance of the ‘Passage’ in relation to the ‘Question’. Use the following criteria to categorize the relevance of the information: ... Respond with ‘Highly relevant’, ‘Perfectly relevant’, ‘Related’, or ‘Irrelevant’ based on the criteria.	Graded
G4	You are an AI model tasked with determining the relevance of a passage to a given query. The input will be a query and a passage. Your task is ... provides some information related to the query, and ‘Irrelevant’ should be used when the passage does not provide any useful information in relation to the query.	Graded

Table 3: Five LLMs that are utilized for this experiment

LLM	Model Parameter Size
Text-babbage-001	1.2 billion
Text-curie-001	6.7 billion
Text-davinci-003	178 billion
GPT-3.5-turbo	178 billion
GPT-4	1.76 trillion

4.2 Dataset

4.2.1 Evaluation data set

For our experiments, we utilize the test sets from the MS MARCO TREC DL Passage datasets spanning three years[2]. As depicted in TABLE 4, we randomly sampled 500 data points from each year's test dataset due to the resource-intensive nature of evaluating the entire dataset, while ensuring that every query in the full set is included. These sampled datasets are then used to evaluate GPT prompts. To confirm that the results from our sampling are consistent with those obtained using the entire dataset, we conducted tests and presented the results in Appendix A. The paired t-test results confirmed that there is no significant difference between them.

Table 4: 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, and the 2022 dataset is employed for sampling few-shot examples. Due to the resource-intensive nature of evaluating the entire dataset, 500 data points from each year are sampled for evaluation purposes. For few-shot examples, 2 to 32 data points are randomly sampled as part of the experimental design. The table details the year of the dataset, the number of queries, the total number of data, and the number of sampled data used in this study.

Usage	TREC DL Year	Number of queries	Number of data	Number of sampled data
Evaluation	2019	43	9,260	500
	2020	54	11,386	500
	2021	53	10,828	500
Few-shot examples	2022	500	369,638	2,4,8,16, and 32

The TREC DL 2019 dataset consists of 9,260 query-passage pairs, with the relevance between each pair assessed by human evaluators. Similarly, the TREC DL 2020 and 2021 test sets follow the same format, comprising 11,386 and 10,828 query-passage pairs with relevance judgments, respectively. Relevance in these pairs is rated on a 4-point scale: "Perfectly relevant," "Highly relevant," "Related," and "Irrelevant." For binary classification tasks, we simplify this 4-point

[2] 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

relevance scale to a binary ”Yes” or ”No” judgment. Specifically, the categories of “Perfectly relevant,” and “Highly relevant,” are consolidated into a “Yes” category, while “Related” and “Irrelevant” is classified as “No.”

4.2.2 Few-shot data set

For our few-shot dataset, we utilize the test set from the MS MARCO TREC DL Passage datasets of 2022[3]. As illustrated Table 4, we sampled between 2 and 32 data points (specifically 2, 4, 8, 16, and 32) as part of the experimental design to serve as few-shot examples. In the binary setting, we convert the original 4-point relevance scale to a binary format. For the graded setting, we ensure that we sample an equal number of examples from each relevance category. For instance, if we are using 4 few-shot examples, we select one example from each of the four relevance categories. TABLE 5, shows few-shot examples.

Table 5: Examples of Few-shot examples: This table illustrates two types of few-shot examples used in our study. The first example demonstrates a binary type, where the answer is a straightforward ’Yes’ or ’No’, while the second example is of the graded type, where relevance is categorized on a scale, in this case as ’Irrelevant’.

Type	Example
Binary	**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 ... **Answer: Yes**
Graded	**Query:** can you open a wells fargo account online? **Passage:** You can transfer money to your checking account from other Wells Fargo. accounts ... **Answer: Irrelevant**

5. EXPERIMENTAL RESULT

In this section, we present three sets of results: a comparison between graded and binary relevance, a comparison of few-shot versus zero-shot approaches. Based on these findings, we propose the best policy for prompt formulation.

5.1 Binary vs. Graded

TABLE 6 provides a detailed comparison of binary and graded prompts performance on various GPT models. For GPT-3 models, kappa values are below 0.1 for both prompt types, indicating limited reliability. GPT-3.5 models, with 178 billion parameters, show improved performance and GPT-4 model, with 1.76 trillion parameters, achieves kappa values over 0.6 for graded prompts, averaging

[3] https://microsoft.github.io/msmarco/TREC-Deep-Learning.html

Table 6: Comparative Performance of GPT Models Using Binary and Graded Prompts: This table summarizes the kappa values for various GPT models across both binary and graded prompts. ‘∗’ rpresent t-test result, which means it’s significantly different with 95% confidence level.

Type	**Prompt**	**GPT-3** (< 10 billion)		**GPT-3.5** (178 billion)		**GPT-4** (1.76 trillion)
		Text-Babbage-001	**Text-Curie-001**	**Text-Davinci-003**	**GPT-3.5-turbo**	
Binary	**M1**	0.049 (±0.031)	0.019 (±0.018)	0.348 (±0.078)	0.458 (±0.069)	0.531 (±0.060)
	M2	0.019 (±0.007)	-0.008 (±0.014)	0.292 (±0.115)	**0.532** (±0.079)	**0.598** (±0.072)
	G1 (Ours)	0.018 (±0.086)	0.022 (±0.020)	0.322 (±0.071)	0.469 (±0.082)	0.446 (±0.083)
	G2 (Ours)	**0.052** (±0.028)	**0.036** (±0.019)	**0.356** (±0.086)	0.473 (±0.064)	0.486 (±0.063)
Graded	M3	**0.060** (±0.043)	**0.070** (±0.058)	0.302 (±0.142)	0.337 (±0.073)	0.604 (±0.042)
	M4	0.007 (±0.052)	-0.014 (±0.020)	0.343 (±0.069)	0.399 (±0.056)	0.628 (±0.078)
	G3 (Ours)	0.036 (±0.045)	0.018 (±0.064)	0.345 (±0.058)	0.356 (±0.036)	0.605 (±0.042)
	G4 (Ours)	0.021 (±0.075)	0.011 (±0.052)	**0.411** (±0.078)	**0.455** (±0.054)	**0.638** (±0.051)
Binary prompts avgerage		**0.038** (±0.022)	0.017 (±0.018)	0.326 (±0.029)	**0.483*** (±0.033)	0.515 (±0.065)
Graded prompts avgerage		0.031 (±0.023)	**0.021** (±0.035)	**0.350** (±0.045)	0.387 (±0.052)	**0.619*** (±0.017)

0.619. As the results for GPT-3.5 and GPT-4 models are significant, the focus will henceforth be on these two models.

Examining the GPT-3.5 models first, the Davinci-003 model exhibits higher kappa values for binary prompts compared to graded ones. However, according to t-test results, the difference is not statistically significant. In contrast, the GPT-3.5-turbo model demonstrates a notable preference for binary prompts, with significant differences in kappa values: 0.483 for binary and 0.387 for graded prompts at a 95% confidence level. The smallest kappa value in binary prompts for this model is 0.458, which even surpasses the highest value in graded prompts, 0.455. Therefore, we can conclude that binary prompts are more effective for the GPT-3.5-turbo model.

On the other hand, the GPT-4 model presents a different result. Graded prompts achieve higher kappa values than binary ones. The average value for binary prompts is 0.515, while for graded prompts it is 0.619, indicating a statistically significant difference. Furthermore, the highest value in binary prompts, 0.598, is smaller than the smallest value in graded prompts, 0.604. From these results, it is evident that graded prompts perform better in the case of GPT-4.

In summary, our experimental results reveal that the efficacy of binary and graded prompts varies depending on the model. While the GPT-3.5-turbo model shows significantly better performance with binary prompts, GPT-4 excels with graded prompts. The GPT-3 models, including the Davinci-003, do not demonstrate a significant preference for either type. Therefore, the choice between binary and graded prompts may depend on the specific LLM model in use.

5.2 Few-Shot vs. Zero-Shot

To compare the few-shot and zero-shot approaches, we selected the M2 and G4 prompts for our experiments. These prompts are chosen because they demonstrated the best performance for GPT-3.5-turbo and GPT-4 in the binary and graded prompt categories, respectively. We tested M2 and G4 with 2, 4, 8, 16, and 32 few-shot examples, and the results are detailed in TABLE 7.

Table 7: Performance Comparison of Few-Shot and Zero-Shot Approaches: This table displays the kappa values for GPT-3.5-turbo and GPT-4 models using Binary (M2) and Graded (G4) prompts across different few-shot scenarios (2, 4, 8, 16, and 32 examples). It includes the average kappa values for few-shot prompts and the results for zero-shot prompts.

	Binary (M2)		Graded (G4)	
# of examples	**GPT-3.5-turbo**	**GPT-4**	**GPT-3.5-turbo**	**GPT-4**
2-shots	0.442 (±0.051)	0.584 (±0.032)	-	-
4-shots	0.446 (±0.047)	0.577 (±0.080)	**0.429** (±0.036)	0.585 (±0.076)
8-shots	**0.464** (±0.079)	0.571 (±0.053)	0.400 (±0.040)	0.584 (±0.088)
16-shots	0.425 (±0.062)	**0.613** (±0.077)	0.429 (±0.045)	0.587 (±0.041)
32-shots	0.415 (±0.059)	0.564 (±0.045)	0.380 (±0.028)	**0.593** (±0.024)
Few-shots avg.	0.438 (±0.019)	0.582 (±0.019)	0.409 (±0.024)	0.587 (±0.004)
Zero-shot	**0.532*** (±0.026)	**0.598** (±0.076)	**0.455*** (±0.054)	**0.638*** (±0.051)

A crucial observation emerging from our experimental data is the relative efficacy of zero-shot approaches compared to few-shot methods. Across both M2 (binary) and G4 (graded) prompts, the zero-shot approach consistently outperforms the few-shot approach, regardless of the number of examples used. For instance, when examining the performance of the GPT-4 model, zero-shot prompts yield a significantly higher kappa value of 0.598 for binary prompts and 0.638 for graded prompts. The disparity is evident despite varying the number of few-shot examples from 2 to 32. In every instance, the zero-shot setup maintains an edge over its few-shot counterpart. This trend is also noticeable with the GPT-3.5-turbo model. The zero-shot approach achieves a kappa value of 0.532 for binary prompts and 0.455 for graded prompts. These values surpass the few-shot averages of 0.438 for binary and 0.409 for graded prompts.

These results highlight a critical aspect of the performance of GPT-4 and GPT-3.5-turbo: the zero-shot approach, which relies solely on the model's innate understanding and processing of the prompt without the aid of example data, is more effective than the few-shot method. This finding suggests that the advanced capabilities of these models are better leveraged without the potential biases or constraints introduced by few-shot examples.

5.3 The Optimal Prompt Configuration

In light of our experimental results and subsequent analyses, as consolidated in TABLE 8, we have identified what appears to be the most effective prompt configuration for each GPT model in relevance evaluation tasks.

Table 8: Summary of the best prompt settings for GPT-3.5-turbo and GPT-4 models, highlighting preferred relevance granularity and in-context example usage based on our experimental findings.

GPT model	Relevance granuality	In-context few-shots
GPT-3.5-turbo	Binary	Zero-shot
GPT-4	Graded	Zero-shot

For the GPT-3.5-turbo model, our findings suggest that binary prompts used in a zero-shot setting yield the best performance. This indicates that, for this particular model, a straightforward binary approach to relevance, devoid of additional in-context examples, aligns best with its processing and evaluative capabilities. Conversely, with the GPT-4 model, graded prompts in a zero-shot setting emerge as the most effective. It seems to benefit from the nuanced and layered structure of graded prompts. The absence of few-shot examples in this context also appears to play a significant role, potentially allowing the model to leverage its extensive pre-training and inherent language understanding more effectively.

These findings underscore a critical aspect of prompt engineering for relevance tasks: the optimal prompt configuration can vary significantly between different GPT models. This variability may be attributed to the differences in model architectures, training data, and inherent capabilities. Therefore, while our study provides valuable insights into prompt design for GPT models, it also highlights the importance of context-specific experimentation. Different models may respond differently to the same prompt structure, and as such, prompt design should be tailored to the specific model in use.

In summary, based on our experimental data, we propose that the best prompt configuration for the GPT-3.5-turbo model is a binary, zero-shot prompt, while for GPT-4, it is a graded, zero-shot prompt. However, these recommendations are grounded in the specific parameters and scope of our study, and further research is necessary to explore their applicability in broader contexts or with other models.

6. DISCUSSION

6.1 Practical Implication

The significance of this study lies in its potential to streamline the creation of training sets for IR tasks. Traditionally, generating relevance judgments for IR has been a labor-intensive and costly process, as it requires human annotators to carefully label large datasets. By LLMs like GPT to assist in the creation of training data, we can potentially reduce the reliance on human annotators, making the process more cost-effective and scalable. Furthermore, the ability of LLMs to generate relevance judgments provides an opportunity to automate the creation of high-quality training sets, which are essential for the development and refinement of IR systems. This shift from humandependent data generation to machine-assisted annotation could significantly accelerate progress in the field of IR.

However, there remains an ongoing discussion about the reliability and trustworthiness of LLM-generated evaluations [3]. While these models have demonstrated impressive capabilities, questions persist regarding whether their relevance judgments can be fully trusted. Despite these concerns, research efforts to leverage LLM-evaluated datasets continue to gain traction, highlighting the potential of such approaches to transform IR and related fields. Future work should focus on validating the consistency of LLM judgments against human benchmarks to further solidify their role in training set generation.

6.2 Limitations

This study has several limitations that should be acknowledged. First, the evaluation was conducted solely on GPT-based models, meaning that the results may differ when applying the same methodology to other LLMs such as Claude[4] or LLaMA[5]. The performance and behavior of different models could lead to variations in outcome, which limits the generalizability of our findings. Second, due to resource constraints, we were unable to evaluate the entire dataset and instead sampled 500 data points for our experiments. While statistical tests confirmed that this sample was representative, the results might still differ if a larger portion or the entirety of the dataset were evaluated. Third, there is a potential for data leakage, as GPT may have been trained on some or all of the datasets used in this study. This raises the possibility that the model could have prior exposure to the test data, which could influence the results.

Despite these limitations, our study demonstrates the potential of using LLMs to aid in constructing datasets for IR tasks, highlighting the promising role of LLMs in this field.

7. CONCLUSION

We presented a comprehensive analysis of the efficacy of different prompt types in variant GPT models such as GPT-3.5 and GPT-4 for relevance evaluation tasks. Through our experiments, we have studied on the impact of binary versus graded prompts and the comparative effectiveness of few-shot and zero-shot approaches. Our findings reveal that the choice between binary and graded prompts significantly influences the performance of LLMs in relevance evaluation tasks. Specifically, while the GPT-3.5-turbo model shows a preference for binary prompts, GPT-4 demonstrates superior performance with graded prompts. This distinction underscores the necessity of prompt customization based on the specific LLM in use.

Furthermore, our experiments highlight the superior efficacy of zero-shot approaches over few-shot ones. In scenarios where precision and alignment with human judgment are paramount, zero-shot prompts offer a more balanced and effective method, particularly with GPT-4.

[4] https://www.anthropic.com/index/claude

[5] https://github.com/meta-llama/llama

In conclusion, our research contributes to a deeper understanding of prompt engineering in IR domain, offering valuable insights into the optimization of LLMs for relevance evaluation. Future work should explore the adaptability of these findings across various domains and the development of more nuanced prompt designs to further enhance the performance and reliability of LLM-based systems.

8. ACKNOWLEDGMENT

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

References

[1] Honovich O, Shaham U, Bowman SR, Levy O. Instruction Induction: From Few Examples to Natural Language Task Descriptions. 2022. ArXiv preprint: https://arxiv.org/pdf/2205.10782

[2] Sun W, Yan L, Ma X, Wang S, Ren P, et al. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. 2023. ArXiv preprint: https://arxiv.org/pdf/2304.09542

[3] Faggioli G, Dietz L, Clarke CL, Demartini G, Hagen M, et al. Perspectives on Large Language Models for Relevance Judgment. In: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval. 2023:39-50.

[4] MacAvaney S, Soldaini L. One-Shot Labeling for Automatic Relevance Estimation. 2023. ArXiv preprint: https://arxiv.org/pdf/2302.11266

[5] Thomas P, Spielman S, Craswell N, Mitra B. Large Language Models Can Accurately Predict Searcher Preferences. 2024. ArXiv preprint: https://arxiv.org/pdf/2309.10621

[6] Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. 2021. ArXiv preprint: https://arxiv.org/pdf/2104.08786

[7] Gao T, Fisch A, Chen D. Making Pre-trained Language Models Better Few-Shot Learners. 2020. ArXiv preprint: https://arxiv.org/pdf/2012.15723

[8] Reynolds L, McDonell K. 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. 2021:1-7.

[9] Webson A, Pavlick E. Do Prompt-Based Models Really Understand the Meaning of Their Prompts? 2021. ArXiv preprint: https://arxiv.org/pdf/2109.01247

[10] Alonso O, Mizzaro S. 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. 2009;15.

[11] Blanco R, Halpin H, Herzig DM, Mika P, Pound J, et al. Repeatable and Reliable Search System Evaluation Using Crowdsourcing. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2011:923-932.

[12] Maddalena E, Basaldella M, De Nart D, Degl'Innocenti D, Mizzaro S, et al . Crowdsourcing Relevance Assessments: The Unexpected Benefits of Limiting the Time to Judge. InProceedings of the AAAI conference on human computation and crowdsourcing. 2016;4:129-138.

[13] Nouri Z, Wachsmuth H, Engels G. Mining Crowdsourcing Problems From Discussion Forums of Workers. In: Proceedings of the 28th International Conference on Computational Linguistics. 2020:6264-6276.

[14] Keikha M, Park JH, Croft WB. Evaluating Answer Passages Using Summarization Measures. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. 2014:963-966.

[15] Dietz L, Chatterjee S, Lennox C, Kashyapi S, Oza P, et al. Wikimarks: Harvesting Relevance Benchmarks From Wikipedia. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022:3003-3012.

[16] Ding B, Qin C, Liu L, Bing L, Joty S, et al. Is GPT-3 a Good Data Annotator? 2022. ArXiv preprint: https://arxiv.org/pdf/2212.10450

[17] Wang S, Liu Y, Xu Y, Zhu C, Zeng M. Want to Reduce Labeling Cost? GPT-3 Can Help. 2021. ArXiv preprint: https://arxiv.org/pdf/2108.13487

[18] Kekäläinen J. Binary and Graded Relevance in IR Evaluations—Comparison of the Effects on Ranking of IR Systems. Inf Process Manag. 2005;41:1019-1033.

[19] Chapelle O, Metlzer D, Zhang Y, Grinspan P. Expected Reciprocal Rank for Graded Relevance. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. 2009:621-630.

[20] Robertson SE, Kanoulas E, Yilmaz E. Extending Average Precision to Graded Relevance Judgments. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2010:603-610.

[21] Sakai T. On the Reliability of Information Retrieval Metrics Based on Graded Relevance. Inf Process Manag. 2007;43:531-548.

[22] Brown TB. Language Models Are Few-Shot Learners. 2020. ArXiv preprint: https://arxiv.org/pdf/2005.14165

[23] Liu J, Shen D, Zhang Y, Dolan B, Carin L, et al. What Makes Good In-Context Examples for GPT-3? 2021. ArXiv preprint: https://arxiv.org/pdf/2101.06804

[24] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models Are Zero-Shot Reasoners. Adv Neural Inf Process Syst. 2022;35:22199-22213.

[25] https://cdn.openai.com/papers/gpt-4.pdf

[26] Bavarian M, Jun H, Tezak N, Schulman J, McLeavey C, et al. Efficient Training of Language Models to Fill in the Middle. 2022. ArXiv preprint: https://arxiv.org/pdf/2207.14255

Appendix A. Comparison Between Sampled Data and Entire Data

We conduct experiments to verify that the Cohen's kappa values for our sampled data are not significantly different from those obtained with the entire dataset. Utilizing the TREC DL 2019 passage test set, we compare the Cohen's kappa values for the prompts M2, G4, and G4 with 32 few-shot examples, between the sampled and the full data sets. As indicated by the results in TABLE 9, although there are minor differences between them, these are not substantial. The paired t-test yield a p-value of 0.47, indicating that the differences are not statistically significant.

Table 9: A comparison of Cohen's kappa values for prompts M2 and G4 (zero-shot and 32-shots) between sampled and the entire TREC DL 2019 passage test set, with a paired t-test yielding a p-value of 0.47.

Prompt	Few/zero-shot	Sampled data	Entire data
M2	Zero-shot	0.549	0.543
G4	Zero-shot	0.574	0.572
G4	32-shots	0.565	0.567