LLM-Specific Utility: A New Perspective for Retrieval-Augmented Generation

Hengran Zhang
State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
zhanghengran22z@ict.ac.cn

Keping Bi
State Key Laboratory of AI Safety, ICT, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
bikeping@ict.ac.cn

Jiafeng Guo
State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
guojiafeng@ict.ac.cn

Jiaming Zhang, Shuaiqiang Wang
Baidu Inc
Beijing, China
zhangjiaming04@baidu.com
wangshuaiqiang@baidu.com

Dawei Yin
Baidu Inc
Beijing, China
yindawei@acm.org

Xueqi Cheng
State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
cxq@ict.ac.cn

Abstract

Retrieval-augmented generation (RAG) enhances large language models (LLMs) by incorporating external knowledge. While traditional retrieval focuses on relevance, RAG's effectiveness depends on the utility of retrieved passages, i.e., the usefulness in facilitating the generation of an accurate and comprehensive answer. Existing studies often treat utility as a generic attribute, ignoring the fact that different LLMs may benefit differently from the same passage due to variations in internal knowledge and comprehension ability. In this work, we introduce and systematically investigate the notion of LLM-specific utility. Through large-scale experiments across multiple datasets and LLMs, we demonstrate that human-annotated passages are not optimal for LLMs and that ground-truth utilitarian passages are not transferable across different LLMs. These findings highlight the necessity of adopting the LLM-specific utility in RAG research. Our findings indicate that some human-annotated passages are not ground-truth utilitarian passages for specific LLMs, partially due to the varying readability of queries and passages for LLMs, a tendency for which perplexity is a key metric. Based on these findings, we propose a benchmarking procedure for LLM-specific utility judgments. We evaluate existing utility judgment methods on six datasets and find that while verbalized methods using pseudo-answers perform robustly, LLMs struggle to assess utility effectively—failing to reject all passages for known queries and to select truly useful ones for unknown queries. Our work highlights the necessity of LLM-specific utility and provides a foundation for developing more effective RAG systems. Our code and datasets can be found at https://anonymous.4open.science/r/LLM_specific_utility-4260/README.md.

CCS Concepts

• **Information systems** → **Language models**; **Learning to rank**; **Novelty in information retrieval**.

Keywords

RAG, LLM-Specific Utility, Utility Judgments

ACM Reference Format:
Hengran Zhang, Keping Bi, Jiafeng Guo, Jiaming Zhang, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. 2018. LLM-Specific Utility: A New Perspective for Retrieval-Augmented Generation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 13 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The retrieval-augmented generation (RAG) framework enhances large language models (LLMs) by incorporating external knowledge. Traditional retrieval aims to find documents relevant to a query. However, RAG's effectiveness hinges on the utility of retrieved passages. Relevance typically focuses on the topical matching between a query and retrieved passages [22, 23]. Utility, in contrast, emphasizes the usefulness of a passage in facilitating the generation of an accurate and comprehensive answer to the question [20, 38, 40]. Recognizing this distinction, the research community has shifted from relying solely on relevance annotations towards evaluating passage quality based on downstream LLM performance

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

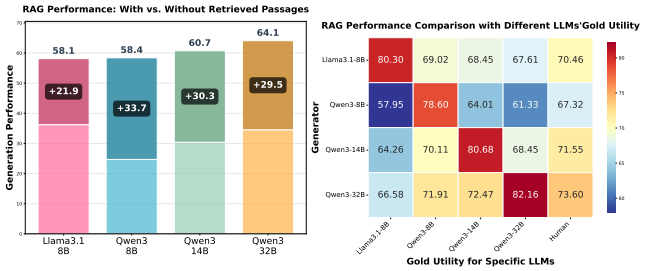

Figure 1: Left: Answer generation performance (*has_answer*, %) with the same top-20 retrieval results upon different LLMs. Right: RAG performance (*has_answer*, %) of LLMs with gold utilitarian passages from different LLMs.

metrics [6, 25]. These metrics include the likelihood of generating ground-truth answers or exact match scores. Another line of work [36, 37, 40] involves prompting LLMs to identify useful passages from a set of relevance-oriented retrieval results for use in RAG.

Current research on passage utility in RAG typically treats utility as a generic attribute, operating under the assumption that a passage considered useful will provide equivalent benefits to any LLM. This view, however, fails to account for the notion of LLM-specific utility. Drawing an analogy from pre-LLM web search, the usefulness of identical search results often varied across users depending on their individual background knowledge. In the contemporary RAG framework, the LLM itself acts as the "consumer" of retrieved passages, synthesizing information to address a given query. Crucially, LLMs differ in two key aspects that shape the actual utility a passage provides in helping them generate a better answer: (1) Different internal knowledge. Each LLM is pre-trained on distinct corpora, resulting in unique knowledge bases encoded within its parameters. Consequently, the same passage is novel and critical for one LLM might be redundant for another that already possesses it. (2) Different ability to understand a potentially useful passage. Variations in scale and training cause LLMs to differ in their capacity to comprehend and draw inferences from the same text. Therefore, a passage rich in information for one LLM might be underutilized or misinterpreted by another with lower comprehension skills. These differences mean that the utility of a passage is not an intrinsic property but is co-determined by the specific LLM using it.

Therefore, it is essential to investigate utility from the perspective of individual LLMs. As shown in Figure 1 (left), the performance improvement from the same top-20 retrieval set varies considerably across LLMs. Figure 1 (right) further illustrates that the gold utilitarian passages constructed for one LLM are not optimal for others; each LLM achieves its best performance when provided with passages specifically constructed for its own needs. These findings highlight the necessity of adopting an LLM-specific notion of utility in RAG research.

Research Goal. The preceding findings concerning LLM-specific utility motivate a deeper and more systematic investigation. We aim to uncover the fundamental relationship between human-annotated general utility and LLM-specific utilitarian passages, along with the underlying causes of their divergence. To this end, we formulate the following research questions: (**RQ1**) *Are human-annotated passages optimal for LLMs in retrieval-augmented generation (RAG)?* (**RQ2**) *What underlies the divergence between human-annotated and LLM-specific utility?* (**RQ3**) *To what extent can LLMs accurately assess LLM-specific utility?*

Benchmarking Procedure. To address these questions, we construct an evaluation benchmark centered on a novel task: LLM-specific utility judgment. This task requires the LLM, when provided with a query and a set of candidate passages, to identify utilitarian passages from the candidate passages. We evaluate this capability through two complementary approaches, both leveraging gold utilitarian passages: set-based evaluation (selecting a utilitarian subset) and ranking-based evaluation (generating a ranked list by utility). The construction of the gold utilitarian passage is LLM-dependent. A passage is considered utilitarian for a given LLM only if it provides a performance gain over the LLM's inherent ability to answer the query without external information. This ensures that the utility is measured as a tangible improvement over the LLM's pre-existing knowledge. We conduct large-scale experiments across multiple knowledge-intensive datasets: NQ [13], TriviaQA [9], MS MARCO-FQA (derived from MS MARCO QA [17]), FEVER [27], HotpotQA [33], and 2WikiQA [4]. We evaluate four LLMs of varying scales and architectures: Qwen2-8B, Qwen2-14B, Qwen2-32B [26], and Llama3.1-8B [2]. We further analyze how well existing utility judgment methods align with ground-truth LLM-specific utility. Current approaches fall into two categories: (1) *verbalized utility judgments* [36, 37, 40], which employ prompting to select utilitarian passages or ranking passages based on utility, and (2) *probabilistic utility estimation* [6, 20, 21, 25, 35, 39], including attention-based methods (using the average attention weight from generated answers over input passages as a utility proxy) and likelihood-based methods (using the likelihood of a pseudo-answer given a passage).

Our empirical investigation yields the following key findings:

- **RQ1**: Human-annotated passages are *not* optimal for LLMs in RAG. LLM-specific gold utilitarian passages consistently yield better downstream performance. Moreover, these utility sets are not transferable across LLMs, underscoring the need for LLM-personalized utility judgments.
- **RQ2**: The incomplete overlap between human-annotated and gold utility for LLMs may arise from the readability for LLMs on some queries and passages. This is evidenced by a marked contrast in perplexity between the human-annotated passages included in and those excluded from the gold utility for LLMs. Additionally, on known queries that the LLM can answer correctly without retrieval, even human-annotated passages lead to slight performance degradation. This implies that very relevant passages would hurt LLM generation performance even when they already possess the knowledge, likely because the LLMs may prioritize provided passages over their own internal knowledge base.
- **RQ3**: Under both set-based and ranking-based evaluations: (1) Verbalized methods using pseudo-answers achieve strong performance across all LLMs and evaluation settings. Attention-based utility estimation performs the worst in ranking tasks, indicating that internal attention is not a reliable proxy for a passage's contribution to the final answer. (2) On known queries, RAG performance degrades more with selected passages than with human-annotated ones, further indicating that LLMs over-rely

on the provided passages even when they already possess sufficient knowledge. These results highlight a core direction for LLM-specific utility judgments: effective utility judgments enable LLMs to reject all passages for known queries, while accurately selecting useful ones for unknown queries.

In summary, our contributions are as follows: (1) We highlight the new perspective of utility for RAG, i.e., LLM-specific utility. (2) We introduce the LLM-specific utility judgment task, propose a benchmarking procedure, and provide a comprehensive empirical analysis of various LLMs and methods. (3) We identify the key direction in achieving more effective LLM-specific utility judgment: known queries should reject all passages, while unknown ones must identify useful ones, which need to be analyzed further.

2 Related Work

2.1 Retrieval-Augmented Generation (RAG)

The current challenges in retrieval-augmented generation (RAG) primarily center on three aspects: 1) how to formulate the information needs of LLMs [1, 8, 11, 12, 14, 19, 24, 29, 34]—that is, during the RAG process, enabling such LLMs to formalize their information requirements based on available context and express them as natural language queries to retrieve relevant documents. For example, FLARE [19] proposed two methods for information need verbalization of LLMs: generate a query for low-confidence spans in the already generated sentence via LLM and masked token with low-confidence as implicit queries. LLatrieval [14] prompted LLMs to generate missed information for answering the question. 2) how to fulfill these information needs [36, 37, 39–41], mainly by improving the retrieval of more relevant documents through enhancements to the retriever itself or by selecting useful documents from retrieval results in the post-retrieval stage to better accomplish the task. For example, UtiSel [38] distilled the utility judgments from LLMs to a dense retriever, reducing dependence on costly human annotations and improving retrieval performance with curriculum learning. Further, UtilityQwen [37] distilled the utility judgments from LLMs to a utility-based selector. and 3) enhancing like performance, faithfulness, or helpfulness of the generation process of large language models [3, 31, 31]. Our work mainly focuses on the second domain.

2.2 Utility-focused RAG

RAG comprises the retriever and the generator. Typically, retrieval models are trained and evaluated using the human-annotated query document relevance labels [17, 39]. In RAG, however, the emphasis shifts from optimizing traditional retrieval metrics to maximizing downstream question answering (QA) performance through retrieval that provides genuinely useful evidence [37]. To improve the utility of retrieved information, three main approaches have emerged: (1) Verbalized utility judgments [36, 37, 40]: Prompt the LLM to explicitly assess the usefulness of candidate documents, often aided by pseudo-answer generation and iterative refinement. (2) Downstream-performance-based utility estimation [6, 21, 25, 35, 39]: Score documents by their impact on end-task performance, such as QA accuracy or the likelihood of producing the ground-truth answer. (3) Attention-based utility estimation [5]: Estimate document utility from the attention mass the generator allocates to tokens from each input document during answer generation, aggregated to the document level. However, these works evaluated by fixed general utility for all LLMs and do not consider the utility for specific LLMs.

2.3 3H Principles in LLM Alignment

After being fine-tuned, large language models (LLMs) acquire the ability to follow instructions and engage in dialogue, enabling preliminary interactions with users. Further advancing this line of research, scholars have sought to align LLMs with human values, proposing that LLM outputs should adhere to the 3H principles [15, 32]: Helpfulness (providing accurate and task-aligned responses), Honesty (avoiding hallucinations and misinformation), and Harmlessness (preventing toxic or unethical outputs). These principles, grounded in human-centric criteria, are enforced through Reinforcement Learning from Human Feedback (RLHF) [3, 28, 30] to align LLMs' behavior with human values. Unlike helpfulness in 3H principles, this work investigates the utility of retrieval results for specific LLMs, adopting a LLM-centric perspective on utility.

3 Problem Statement

3.1 Task Description

In RAG, a retriever R and a generator G collaborate to answer a query q using a document corpus $\mathcal{D}$. The process begins with R retrieving a set of top-k relevant passages $C = \{d_1, d_2, \ldots, d_k\}$ from $\mathcal{D}$ as the candidate passages. Then LLMs identify utilitarian passages from C in two forms, which are motivated by the distinct forms of LLM outputs: (1) Subset Selection. The LLM selects a subset $\mathcal{U}_s \subseteq C$ of utilitarian passages. (2) Ranked List. The LLM produces a ranking $\mathcal{U}_r$ of C ordered by the utility. These output types correspond to different evaluation paradigms. Set evaluation assesses how well $\mathcal{U}_s$ aligns with the gold utilitarian set $\mathcal{G}$. Ranking evaluation measures the quality of the ordering in $\mathcal{U}_r$ against an ideal ranking.

3.2 Benchmark Construction

3.2.1 Gold Utilitarian Passages $\mathcal{G}_q$ for Specific LLM Construction. A document possesses intrinsic value, which is independent of any LLM. However, due to differences in their internal knowledge and capabilities, the same document may hold varying utilitarian value for different LLMs. If an LLM can answer a question directly based on its internal knowledge, it indicates that the LLM already possesses the necessary information to address the query. In such cases, external task-related knowledge provides no additional utility to the LLM's answer generation. Therefore, when constructing corresponding gold utilitarian passages for each LLM, we take into account the performance gain achieved by the LLM when using the passages compared to answering directly without them. This gain in performance serves as the criterion for defining gold utilitarian passages. Specifically, the gold utilitarian passages $\mathcal{G}_q$ for a specific LLM $\mathcal{L}$ on the query q is constructed through a pointwise method. The construction process is formalized as follows: For each passage $d_i \in C$, we define a binary utility indicator $u_i \in \{0, 1\}$ based on comparative performance evaluation:

$$u_i = \mathbb{I}[has\_answer(\mathcal{L}(q, d_i)) > has\_answer(\mathcal{L}(q, \emptyset))], \quad (1)$$

where, $\mathcal{L}(q, \emptyset)$ denotes the response generated by $\mathcal{L}$ without any passage context; $\mathcal{L}(q, d_i)$ denotes the response generated by $\mathcal{L}$

when provided with passage d_i; $has\_answer(A)$ is a binary evaluation function that returns 1 if the ground truth answer is contained in the generated answer A, and 0 otherwise. The gold utilitarian passage is then defined as:

$$\mathcal{G}_q = \{d_i \in C \mid u_i = 1\}. \tag{2}$$

This construction ensures that $\mathcal{G}_q$ contains precisely those passages that provide measurable utility to $\mathcal{L}$ in answering query q, as determined by the improvement in answer quality when the passage is utilized. By default, all the answer generation performance in our experiment uses the *has_answer* score.

3.2.2 Datasets. In this work, we employ development sets from several publicly available benchmarks to analyze. Specifically, we utilize four datasets from the KILT benchmark [18]: Natural Questions (NQ), TriviaQA, HotpotQA, and FEVER. Additionally, we incorporate 2WikiQA and the MS MARCO QA development set.

Natural Questions (NQ) [13]. This dataset comprises real user queries from Google Search. To enhance provenance coverage, KILT conducted an Amazon Mechanical Turk campaign for the development and test sets, increasing the average number of provenance pages per question from 1 to 1.57.

TriviaQA [9]. A collection of question-answer-evidence triples where evidence documents are automatically collected from Wikipedia. KILT uses only the Wikipedia-based portion of this dataset.

HotpotQA [33]. This dataset provides question-answer pairs along with human-annotated supporting sentences. We adopt the full wiki setting, which requires systems to retrieve and reason over the entire Wikipedia.

FEVER [27]. A large-scale dataset designed for claim verification, where systems must retrieve sentence-level evidence to determine whether a claim is supported or refuted.

2WikiQA [4]. A multi-hop question answering dataset that incorporates both structured and unstructured data.

MS MARCO-FQA [17]. The development set comprises queries sampled from Bing's search logs, each accompanied by human-annotated passages. Queries are categorized by type: LOCATION, NUMERIC, PERSON, DESCRIPTION, ENTITY. We focused on the factual questions within the MS MARCO dataset (approximately half of all queries), as non-factual DESCRIPTION queries are difficult to evaluate under our framework. A key difference from the NQ dataset is that MS MARCO's ground-truth answers are sentences rather than specific entities. To address this, we employed the Qwen3-32B model to extract precise entity answers from the filtered MS MARCO data. This newly processed collection is termed the MS MARCO-FQA dataset. More dataset statistics are shown in Table 6 in Appendix C.

3.2.3 Retrieval. For the datasets from the KILT benchmark as well as 2WikiQA, the corpus is the Wikipedia dumps and then split into DPR [10] 100-word format as passages, resulting in about 34M passages, which is provided by DPR[1]. For the MS MARCO-FQA, we directly use the MS MARCO v1 passage corpus [17], which contains about 8.8M passages. We utilize the well-performing BGE-M3 [16] as our retriever and retrieve top-20 results for all six datasets.

[1]https://dl.fbaipublicfiles.com/ur/wikipedia_split/psgs_w100.tsv.gz

3.2.4 Analyzed LLMs. We analyze four LLMs from two different families: Llama-3.1-8B-Instruct (denoted as Llama3.1-8B) [2], Qwen3-8B, Qwen3-14B, and Qwen3-32B [26]. To ensure reproducibility, the temperature for all LLMs in this study was set to 0. By default, Qwen3's think function is disabled.

3.3 LLM-Specific Utility Judgment Methods

This section formalizes current methods for enabling LLMs to assess the utility of retrieved passages. The approaches are categorized according to their output format, namely *utility-based selection* [36, 37, 40] and *utility-based ranking* [6, 20, 21, 25, 35].

3.3.1 Utility-Based Selection. This method produces a binary subset $\mathcal{U}_s \subseteq C$ consisting of passages deemed useful by the LLM $\mathcal{L}$. The primary technique employed is verbalization, which can be further divided into two types based on input format: pointwise verbalized judgments and listwise verbalized judgments. The utility definition in the prompt, which accounts for both the intrinsic value of a passage and the model's pre-existing knowledge, is shown in Appendix A.

Pointwise Verbalized Judgments. In this approach, each passage $d_i \in C$ is evaluated independently via a prompting strategy that elicits a binary usefulness decision:

$$\mathcal{U}s = \{d_i \in C \mid \mathcal{L}(q, d_i, P_{\text{binary}}) = \text{“yes”}\}, \tag{3}$$

where P_{binary} is a prompt template designed specifically for binary utility classification based on the definition.

Listwise Verbalized Judgment. Here, the entire candidate set C is presented simultaneously, and the LLM directly outputs the subset of useful passages:

$$\mathcal{U}s = \mathcal{L}(q, C, P_{\text{subset}}), \tag{4}$$

where P_{subset} is a prompt that instructs the model to identify and return only the passages with utility referring to the self-utility definition.

Furthermore, empirical studies indicate that utility judgments yield higher performance when judging utility referring to the pseudo-answers generated from retrieved documents [36, 40]. Accordingly, LLMs can pre-generate pseudo-answer a using retrieval results to inform their utility judgments, i.e., $\mathcal{U}s = \{d_i \in C \mid \mathcal{L}(q, d_i, a, P_{\text{binary}}) = \text{“yes”}\}$ or $\mathcal{L}(q, C, a, P_{\text{subset}})$. Therefore, we considered four verbalized methods for utility-based selection, i.e., verbalized pointwise (Directly judging or *w/* pseudo-answer judge) and verbalized Listwise (Directly selecting passages or *w/* pseudo-answer) to select passages.

3.3.2 Utility-Based Ranking Methods. An alternative approach produces a utility-ordered ranking $\mathcal{U}_r$ of the passages in C according to a predefined notion of utility. Several works [6, 25, 40] for estimating utility scores within LLM-based frameworks, which can be broadly classified into verbalized, attention-based, and likelihood-based paradigms.

Listwise Verbalized Ranking. LLMs directly generate an explicit ranking list with or without a pseudo-answer a via structured output:

$$\mathcal{U}r = \mathcal{L}(q, C, P\text{rank}) \quad \text{or} \quad \mathcal{L}(q, C, a, P_{\text{rank}}), \tag{5}$$

where P_{rank} is a prompt that instructs the model to output a list of passage identifiers in descending order of utility. Further details regarding all prompt templates are provided in Appendix A.

Attention-based Utility Scoring. The utility of a passage is inferred from the LLM's internal attention distributions observed during answer generation:

$$\text{Utility}(d_i) = \frac{1}{T}\sum_{t=1}^{T}\alpha_t(d_i), \tag{6}$$

where $\alpha_t(d_i)$ denotes the aggregate attention weight assigned to passage d_i at decoding step t, and T is the total length of the generated output. This raw utility score is then normalized to produce the final utility score.

Likelihood-based Utility Scoring. This approach quantifies utility as the conditional probability of generating the pseudo-answer given the specific passage:

$$\text{Utility}(d_i) = P(a \mid q, d_i), \tag{7}$$

where a represents the pseudo-answer generated from C via the LLM $\mathcal{L}$. The final ranking $\mathcal{U}_r$ is obtained by sorting all passages in C by their corresponding utility scores.

3.4 Evaluation Setting

LLM-Specific Utility Judgments Evaluation. To comprehensively evaluate the LLM-specific utility judgment capabilities, we adopted two evaluation approaches: (1) set-based evaluation, assessed using Precision, Recall, and F1 score on the non-empty gold utilitarian passages and Accuracy on the empty gold utilitarian passages; and (b) ranking-based evaluation, measured by Normalized Discounted Cumulative Gain (NDCG) [7] and Recall. To ensure fair comparison, the pseudo-answers used in all methods are the same, which are generated based on the top-20 retrieval results.

RAG Evaluation. For RAG performance in our work refers to answer generation performance, using *has_answer* during the whole experiment. To further analyze RAG performance on different queries, we divided all queries into known queries and unknown queries. Known queries refer to those that LLMs can answer correctly without any retrieved passages, while unknown queries are those that cannot be answered without external knowledge.

4 Are the Human-Annotated Passages Optimal for LLMs in RAG?

A user's satisfaction with search results depends heavily on their personal knowledge, meaning the same retrieved passage can be perceived differently by different people. Similarly, in RAG, where the LLM consumes the information, a passage's utility for answering a query can also vary from one LLM to another. However, today's RAG and retrieval systems are almost universally aligned to human-annotated, static datasets. This practice prompts a key question: are the human-annotated passages optimal for LLMs in RAG? To address this, we examine the relationship between human-annotated relevant passages and ground-truth utilitarian passages for specific LLMs in this section. To enable a direct comparison with human-annotated passages, we construct a candidate set U for each query by combining the top-20 retrieved passages with the human-annotated passages, as the original annotation pool for humans is unavailable. On the other hand, in real-world RAG systems, human-annotated passages are generally unavailable. Therefore, we also consider the top-20 retrieved passages alone as another candidate set, i.e., C, to reflect a purely retrieval-based scenario. We focus on three single-hop query datasets—NQ, TriviaQA, and MS MARCO-FQA in this section, to enable a rigorous analysis of the relationship between human-annotated relevant passages and LLM-specific utility.

Table 1: Answer generation performance (*has_answer*, %) when provided with different passages list: no passages (None), the top-20 retrieval results, i.e., C, human-annotated passages (Human), a union of the top retrieval and human passages (Union (U)), and the gold utilitarian passages for the specific LLM, i.e., $\mathcal{G}_q$. Bold indicates the best performance for the specific LLM. The number in parentheses indicates the average number of human-annotated relevant passages, or gold utilitarian passages for specific LLMs.

Generator	NQ					
	None	C	U	Human	$\mathcal{G}_q$ (C)	$\mathcal{G}_q$ (U)
Llama3.1-8B	36.24	58.12	61.16	70.46 (3.0)	76.17 (1.7)	**80.30** (1.9)
Qwen3-8B	24.71	58.41	59.89	67.32 (3.0)	74.59 (2.0)	**78.60** (2.3)
Qwen3-14B	30.38	60.73	63.62	71.55 (3.0)	76.84 (1.9)	**80.68** (2.1)
Qwen3-32B	34.54	64.08	67.36	73.60 (3.0)	78.39 (1.8)	**82.16** (2.1)
Generator	TriviaQA					
	None	C	U	Human	$\mathcal{G}_q$ (C)	$\mathcal{G}_q$ (U)
Llama3.1-8B	75.52	88.30	89.74	90.67 (2.4)	95.84 (1.4)	**97.01** (1.5)
Qwen3-8B	59.19	87.61	89.12	88.04 (2.4)	94.72 (2.3)	**95.84** (2.6)
Qwen3-14B	68.18	89.83	91.23	91.25 (2.4)	95.80 (1.8)	**97.20** (2.0)
Qwen3-32B	71.26	90.41	91.64	91.42 (2.4)	96.21 (1.7)	**97.18** (2.0)
Generator	MS MARCO-FQA					
	None	C	U	Human	$\mathcal{G}_q$ (C)	$\mathcal{G}_q$ (U)
Llama3.1-8B	19.13	40.45	41.36	60.33 (1.1)	62.96 (1.8)	**67.61** (1.9)
Qwen3-8B	17.47	42.76	43.83	64.86 (1.1)	66.71 (2.0)	**71.21** (2.1)
Qwen3-14B	20.69	45.14	46.51	70.37 (1.1)	70.58 (1.9)	**76.24** (2.0)
Qwen3-32B	20.82	47.64	49.27	70.65 (1.1)	70.02 (2.1)	**75.46** (2.2)

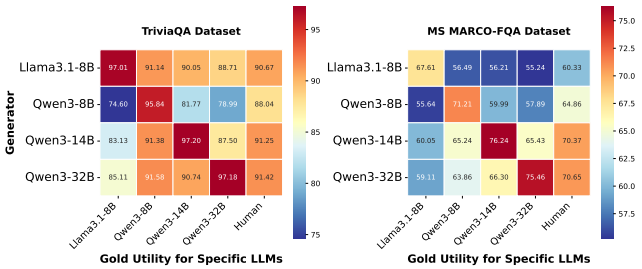

Figure 2: RAG performance (%) of LLMs with different gold utilitarian passages (U candidate) from different LLMs.

4.1 Human-Annotated vs. LLM-Specific Utility

Table 1 compares the performance of Retrieval-Augmented Generation (RAG) using these different sets. We can observe that: (1) **Superiority of LLM-Specific Utility**: The highest RAG performance across all evaluated LLMs and datasets is achieved using gold passages of utility constructed from a union of candidate passages. This result underscores the critical importance of tailoring utility assessment to the specific LLM used for answer generation. (2) **The Upper Bound of Utility Judgments**: Using these

ground-truth utilitarian passages—derived from top-20 retrieval results—yields a significantly greater improvement in answer performance compared to using all top-20 results, as well as union candidate. Furthermore, this LLM-specific utility generally surpasses the performance of human-annotated general utility. This pattern suggests that the quality of retrieval results, as validated by the ground-truth utilitarian passages, is the dominant factor for a given dataset. Collectively, these findings establish the performance upper bound for the utility judgments on the candidate passages. (3) **Impact of Noisy Passages**: While using a union of all retrieved passages and human-annotated general utility leads to better answer performance than the standard top-20 results, it remains significantly worse than using human-annotated passages. This performance gap is particularly pronounced on the NQ and MS MARCO-FQA datasets, indicating that the presence of irrelevant or "noisy" passages can substantially impair LLM performance during answer generation. (4) **Performance Hierarchy of References**: A consistent hierarchy of reference effectiveness is observed across LLMs and datasets: Gold utilitarian passages (Union) > Gold utilitarian passages (Retrieval) > Human-annotated passages >= Union results > Retrieval > None. This confirms that while human annotations are valuable, customizing the reference for the specific LLM yields the greatest benefit.

4.2 Transfer of Gold Utilitarian Passages

Human-annotated general utility assumes that the annotated passages are useful for all LLMs. Then, it is natural to ask whether the gold utilitarian passages for specific LLMs are shared with other LLMs. Therefore, we analyze the gold utilitarian passages transfer experiment. Specifically, the utilitarian passages constructed for a specific LLM are used by other LLMs to answer questions. The interesting results are shown in Figure 2 (Experiments on NQ dataset are shown in Figure 1 (Right)). We can observe that: (1) **Gold Utilitarian Passages for Specific LLM Are Not Transferable**: The gold utilitarian passages constructed for a specific LLM is not shareable with other LLMs, i.e., LLM A has the best answer generation performance with the gold utilitarian passages constructed for LLM A. The reason is likely that different LLMs possess distinct internal knowledge bases and different understanding abilities. This finding underscores the necessity of LLM-specific calibration and challenges the practice of employing a static set of annotated passages to evaluate retrieval quality across diverse LLMs. (2) **Family Similarity Matters**: LLMs from the same family (e.g., the Qwen3 series) have more aligned information needs. This is shown by the fact that Qwen3 models perform worse when using gold passages constructed for Llama3.1-8B than when using passages constructed for other Qwen models. (3) **Human-Annotated Relevant Passages Are Transferable**: Figure 2 shows a consistent performance hierarchy: an LLM achieves its best results with its own utilitarian passages, followed by general human-annotated passages, and performs worst with the utilitarian passages optimized for a different LLM. This pattern confirms that while human-annotated relevant passages may not be optimal for any single LLM, they capture a general utility that is robust, widely applicable, and effective. We also conducted the transfer experiment on the $\mathcal{G}_q$ (C candidate), as shown in Figure 8 in Appendix D, which is consistent with the findings reported in Figure 2.

5 What Underlies the Divergence Between Human-annotated and LLM-Specific Utility?

To investigate the discrepancy between human-annotated passages and the LLM-specific ground-truth utilitarian passages (union candidate set), we compute the intersection between these two sets, as illustrated in Figure 3. Our observations are as follows: (1) **Family Similarity Matters:** LLMs from the same family exhibit a higher degree of overlap in their utilitarian passages compared to those from different families. For instance, on the NQ dataset, the gold utilitarian passages of Qwen3-8B show an average overlap of 0.94 with Llama3.1-8B, while the overlap values with Qwen3-14B and Qwen3-32B are 1.19 and 1.11, respectively. (2) **Not All Human-Annotated Passages Qualify As Gold Utilitarian Passages**. On average, only about half of the human-annotated passages are included in the LLM-specific gold utilitarian passages for both the NQ and MS MARCO-FQA datasets. Regarding why the gold utilitarian passages for LLMs do not fully align with human-annotated passages, our case analysis indicates that approximately 90% of the passages in an LLM's gold utilitarian set possess general utility for the corresponding query. These passages may not have been retrieved during human annotation; otherwise, they would likely have been labeled as positive by human annotators.

As for why some human-annotated passages are absent from an LLM's gold utilitarian set, we hypothesize that the discrepancy stems from the understanding capabilities of LLMs. Figure 4 (up) displays the perplexity of various LLMs on different kinds of human-annotated passages. Figure 4 (bottom) reports the RAG performance of LLMs when using human-annotated passages to show the capabilities of LLMs in the dataset. Our findings can be summarized as follows: (1) **The Perplexity Gap:** The reasoning in human-annotated passages may be absent from LLM-specific utility sets due to the models' readability and confusion when processing the query and passages. LLMs assign lower perplexity to passages within their ground-truth utilitarian passages, a trend that is most evident in the NQ dataset. The MSMARCO dataset shows a less distinct pattern, likely because the majority of its human-annotated passages fall into the ground-truth utilitarian category, leaving fewer samples for comparison. Furthermore, the joint perplexity of a query and its ground-truth utilitarian passage is consistently lower than that of a query with a non-utility passage, a pattern that is consistent across all datasets. (2) **The Over-Reliance on Passages in RAG:** The results of Figure 4 (bottom) indicate that even when provided with highly relevant human-annotated passages, the RAG performance of LLMs degrades compared to the no-passage condition. The reasoning may be that LLMs may prioritize provided passages over their own internal knowledge base. (3) **Better LLMs Lower Perplexity:** Within a given model family, more capable LLMs—those achieving better RAG performance—consistently exhibit lower perplexity on both the human-annotated passages and the query-passage pairs.

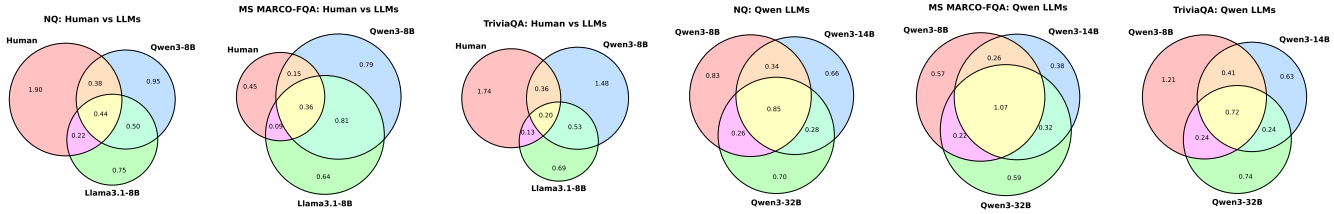

Figure 3: Average number of overlapping passages between the gold utility (U candidate) of LLM and human-annotated passages.

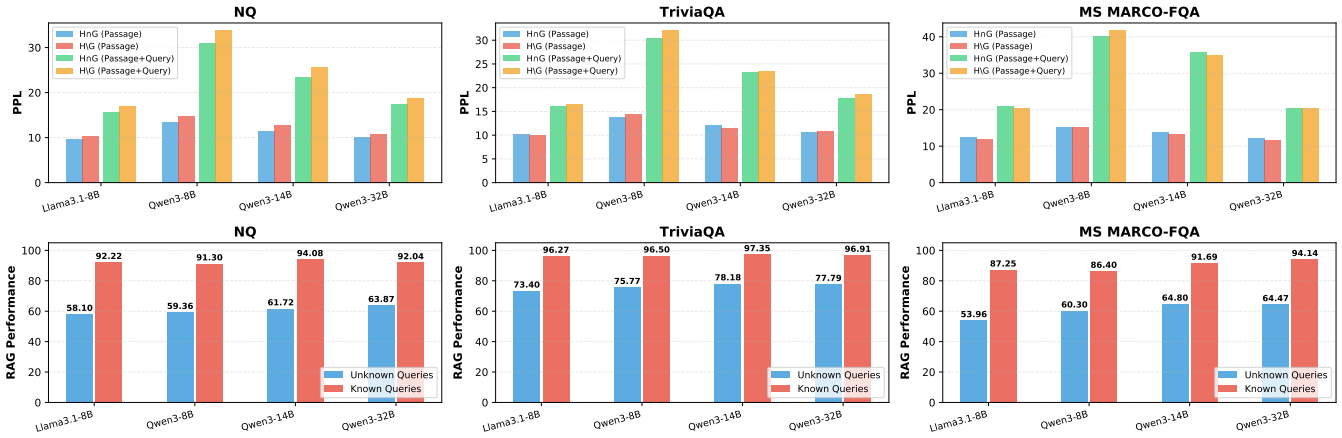

Figure 4: Up: The perplexity (PPL) of LLMs on human-annotated passages of the queries that the gold utilitarian passages (U candidate) for LLMs are not empty. "H" and "G" mean the human-annotated passages and gold utilitarian passages for a specific LLM, respectively. Bottom: RAG performance (*has_answer*, %) of LLMs with human-annotated passages on different queries. "Unk" means "Unknown". The definitions of "Known" and "Unknown" are shown in Section 3.4.

Table 2: LLM-specific utility judgments performance (%) of different utility-based selection methods. "MQA" means the MS MARCO-FQA dataset. "P" and "L" means "Pointwise" and "Listwise", respectively. "$w/$ A" means with pseudo-answer during utilitarian passages selection. Bold indicates the best performance for the specific LLM.

LLM	Type	Method	Gold Utilitarian Passages are Not Empty: F1						Gold Utilitarian Passages are Empty: Accuracy					
			NQ	HotpotQA	2WikiQA	TriviaQA	MQA	FEVER	NQ	HotpotQA	2WikiQA	TriviaQA	MQA	FEVER
Llama3.1-8B	P	Direct	40.28	40.28	36.93	58.07	37.43	50.25	0.18	3.85	5.93	0.74	0.06	15.68
		$w/$ A	48.61	48.61	39.70	**63.05**	44.29	42.45	1.65	**7.00**	**8.31**	1.02	0.77	24.39
	L	Direct	47.86	47.86	38.59	58.58	44.12	52.13	**3.37**	2.75	1.41	**1.10**	**3.25**	7.41
		$w/$ A	**49.88**	**49.88**	**39.91**	60.28	**47.13**	**52.72**	1.65	2.50	1.16	0.38	1.36	**7.45**
Qwen3-8B	P	Direct	48.90	48.90	25.88	63.95	45.29	45.05	3.93	10.90	16.44	2.74	1.87	41.11
		$w/$ A	56.66	56.66	26.51	69.40	49.14	45.83	**8.95**	**17.93**	**30.53**	**5.04**	**8.11**	**38.32**
	L	Direct	56.05	56.05	37.78	67.83	49.50	**59.24**	0.15	0.32	0.13	0.15	0.06	3.72
		$w/$ A	**58.37**	**58.37**	**37.95**	**69.66**	**52.24**	59.22	0.73	0.59	0.65	0.18	0.45	0.95
Qwen3-14B	P	Direct	49.66	49.66	17.10	63.08	47.02	56.83	9.39	22.97	40.76	**4.77**	5.42	**33.34**
		$w/$ A	**57.86**	**57.86**	23.56	71.59	48.91	62.24	**9.87**	**23.30**	**45.67**	4.30	**8.75**	5.74
	L	Direct	55.14	55.14	39.91	67.68	47.50	63.05	1.58	2.90	2.21	0.99	0.72	2.85
		$w/$ A	57.27	57.27	**40.67**	**69.70**	**50.15**	**63.64**	1.58	1.79	2.90	0.70	0.78	0.73
Qwen3-32B	P	Direct	50.17	50.17	17.62	62.29	47.92	40.33	6.87	17.93	35.13	**3.37**	3.50	**8.79**
		$w/$ A	54.79	54.79	24.28	**67.64**	48.65	**49.34**	**9.98**	**20.76**	**48.98**	3.27	**9.33**	3.21
	L	Direct	56.61	56.61	**39.89**	64.24	51.29	44.20	0.20	0.14	0.01	0.05	0.19	1.43
		$w/$ A	**55.87**	**55.87**	39.88	64.62	**53.36**	44.15	0.79	1.61	2.70	0.43	0.58	0.52

6 Can LLMs Assess the LLM-Specific Utility?

6.1 Utility-Based Selection Results

LLM-Specific Utility Judgment Results. Table 2 presents the utility-based selection performance (measured in F1 score) across various LLMs and datasets using different verbalized strategies. Key observations are as follows: (1) **Listwise vs. Pointwise**: The listwise approach consistently outperforms the pointwise method across all LLMs, indicating its advantage in capturing contextual dependencies among passages for utility-based selection. Considering the accuracy when gold utilitarian passages are empty, all methods have an over-selection problem, especially listwise methods.

Table 3: RAG performance (*has_answer*, %) with different references from different verbalized self-utility selection with pseudo-answer methods. All the prompts are shown in Appendix A. Bold indicates the best performance for the specific LLM.

LLM	Source	NQ		HotpotQA		TriviaQA		MS MARCO-FQA		FEVER		2WikiQA		ALL
		Unk	Known	Unk	Known	Unk	Known	Unk	Known	Unk	Known	Unk	Known	AVG
Llama3.1-8B	BGE-M3	42.62	85.41	25.92	80.82	70.05	94.22	31.19	79.58	69.26	92.61	24.97	**62.85**	63.29
	Pointwise Selection	43.45	85.89	27.09	**82.79**	**72.26**	**95.21**	31.23	80.23	56.72	90.75	23.96	60.63	62.52
	Listwise Selection	**44.06**	**85.70**	**28.62**	**82.79**	71.11	94.86	**31.43**	**81.37**	**71.77**	**93.29**	**25.19**	62.11	**64.36**
Qwen3-8B	BGE-M3	49.30	**86.16**	28.65	85.30	77.64	94.48	**34.73**	**80.68**	77.32	**95.99**	23.73	70.57	67.05
	Pointwise Section	46.35	85.73	26.81	**87.93**	75.95	**95.33**	31.36	77.10	60.65	95.67	17.42	**78.32**	64.88
	Listwise Selection	**49.81**	85.02	**30.24**	86.18	**78.83**	94.14	34.36	78.71	**78.03**	95.95	**28.44**	72.79	**67.71**
Qwen3-14B	BGE-M3	48.35	**89.10**	29.76	86.16	77.01	95.81	35.40	82.48	74.69	95.02	27.39	78.31	68.29
	Pointwise Selection	46.78	**89.10**	28.00	**89.86**	76.6	**96.80**	33.11	81.57	75.45	**96.50**	14.72	**89.98**	68.21
	Verbalized Selection	**49.37**	88.52	**31.67**	87.29	**78.06**	95.92	**35.83**	**82.63**	**77.37**	96.29	**30.80**	79.21	**69.41**
Qwen3-32B	BGE-M3	**50.78**	**89.29**	**31.75**	87.31	76.49	96.02	**37.27**	**87.09**	70.54	96.43	30.96	76.54	69.21
	Pointwise Section	47.01	87.76	28.39	**90.39**	75.91	**96.67**	33.24	82.13	70.28	**97.28**	17.28	**86.29**	67.72
	Listwise Selection	50.40	88.67	**31.75**	87.96	**77.66**	95.78	36.44	85.59	**73.42**	96.87	**33.73**	77.42	**69.64**

Table 4: Ranking performance (%) for different LLM-specific utility-based ranking methods. "N", "R" means "NDCG" and "Recall", respectively. Bold means the best performance among the same LLM.

Method	NQ		HotpotQA		2wikiQA		TriviaQA		MS MARCO-FQA		FEVER		ALL AVG	
	N@5	R@5	N@5	R@5	N@5	R@5	N@5	R@5	N@5	R@5	N@5	R@5	N@5	R@5
Llama3.1-8B														
BGE-M3	52.72	51.14	51.11	48.95	43.64	34.59	56.68	42.45	53.35	53.96	66.39	31.55	53.98	43.77
Attention	29.03	32.45	27.95	28.31	38.79	28.09	33.69	26.19	28.23	31.59	60.12	26.38	36.30	28.83
Likelihood	**62.66**	56.71	**64.67**	58.51	51.73	39.73	**78.25**	**59.33**	54.01	49.65	**71.75**	**35.85**	**63.84**	49.96
Verbarlized (Direct)	57.76	56.56	61.09	59.96	51.67	41.59	68.32	52.54	53.72	54.82	69.87	33.81	60.40	49.88
Verbarlized (w/ Answer)	60.37	**58.36**	63.67	**61.63**	**53.54**	**43.05**	69.73	53.68	**55.06**	**55.42**	69.59	33.59	61.99	**50.96**
Qwen3-8B														
BGE-M3	54.47	53.62	49.46	48.12	38.06	35.39	59.89	45.09	55.32	55.74	71.05	42.48	54.71	46.74
Attention	45.69	44.53	43.22	44.27	34.07	30.51	48.98	37.77	42.37	42.42	62.78	35.60	46.18	39.18
Likelihood	64.77	58.41	66.42	60.61	48.94	42.29	**81.27**	**62.37**	52.44	48.06	**82.95**	**51.65**	66.13	53.90
Verbarlized (Direct)	68.25	66.37	66.7	65.67	48.31	45.62	79.52	61.96	59.61	**59.72**	77.82	48.51	66.70	57.98
Verbarlized (w/ Answer)	**69.83**	**67.59**	**68.08**	**66.57**	**49.53**	**46.60**	79.53	61.96	**60.02**	59.60	77.78	48.46	**67.46**	**58.46**
Qwen3-14B														
BGE-M3	54.04	53.20	49.19	48.90	37.03	36.71	59.37	44.66	53.89	55.88	70.23	47.45	53.96	47.80
Likelihood	64.41	58.57	67.51	62.83	**55.91**	50.33	**84.26**	**64.85**	51.98	47.71	74.18	48.86	66.38	55.53
Verbarlized (Direct)	68.47	66.16	67.83	67.56	50.92	51.34	80.92	63.49	**60.35**	**61.24**	**79.32**	55.10	67.97	60.82
Verbarlized (w/ Answer)	**69.51**	**66.60**	**68.59**	**68.02**	52.22	52.00	80.95	63.47	60.34	61.23	79.57	**55.24**	**68.53**	**61.09**
Qwen3-32B														
BGE-M3	53.93	52.90	48.50	47.52	40.06	35.70	59.89	42.72	54.15	54.35	69.96	32.63	54.41	44.30
Likelihood	61.01	55.03	66.58	61.93	**55.89**	46.75	**81.02**	59.10	50.84	46.28	**73.40**	**36.14**	64.79	50.87
Verbarlized (Direct)	66.87	64.97	67.40	65.64	52.81	48.49	79.49	**59.56**	**59.29**	58.60	72.92	35.75	66.46	55.50
Verbarlized (w/ Answer)	**67.72**	**65.10**	**68.95**	**66.53**	53.96	**48.56**	79.50	**59.56**	59.15	**58.80**	73.30	35.73	**67.10**	**55.71**

(2) **Model Comparison**: Qwen-series models (8B, 14B, and 32B) generally achieve higher F1 performance compared to Llama3.1-8B, suggesting stronger inherent capability in utility estimation. (3) **Effect of Pseudo-Answer**: Incorporating pseudo-answers during utility judgments generally leads to improved F1 performance. This finding is consistent with the observations reported by Zhang et al. [40].

RAG Performance. To evaluate the answer generation performance of different utility-based passage selection methods, we employ the two top-performing approaches identified in Table 2—verbalized pointwise and verbalized listwise selection, both incorporating pseudo-answers. The results are summarized in Table 3, from which we derive the following observations: (1) **LLM-specific Utility Judgments vs. BGE-M3**: RAG using self-utility judgments generally outperforms the baseline of directly using the top-20 passages retrieved by BGE-M3. This trend holds across most datasets and LLMs, indicating the benefit of incorporating model-aware utility estimation in passage selection. (2) **The Over-Reliance on Passages in RAG:** For known queries, the introduction of utility-selected passages often leads to performance degradation across all datasets, with Llama3.1-8B being the most noticeably affected. This suggests that LLMs over-rely on the provided passages even when they already possess sufficient knowledge.

6.2 Utility-Based Ranking Results

Table 4 shows the utility-based ranking performance on different datasets. We can observe that: (1) **Attention-Based Ranking Performs Poorly**: The attention weights from the LLM during answer

generation show the poorest alignment with utility, performing worse even than the standalone BGE-M3 retriever. This indicates that the internal attention mechanism is not a reliable indicator of a passage's contribution to the final answer. (2) **LLM-Specific Utility Methods vs. BGE-M3**: Both the likelihood and verbalized methods significantly outperform the BGE-M3 baseline across most datasets and models. This demonstrates that LLMs can effectively rank passages based on their utility for answer generation. (3) **Likelihood is Sensitive to the Pseudo-answer**: The verbalized method with pseudo-answers achieves the best overall performance on most datasets. However, the likelihood method is superior on the TriviaQA dataset for all LLMs. As shown in Table 2, LLMs can generate high-quality answers on TriviaQA, indicating that the queries in TriviaQA are simpler than other datasets. The superior performance of the likelihood method on this dataset suggests it is more sensitive to the quality of the pseudo-answer, benefiting more from accurate generations than the verbalized method.

7 Conclusion

In this paper, we introduced a novel and critical perspective for RAG: LLM-specific utility. We challenged the conventional paradigm of a static, human-annotated general utility by empirically establishing that the passages most useful for answer generation are intrinsically dependent on the specific large language model (LLM) in use. Our demonstration that gold utilitarian passages for a specific LLM outperform human-annotated passages. Moreover, the gold utilitarian passages for specific LLMs are non-transferable across LLMs, fundamentally redefining the objective of LLM-specificity in RAG systems. To systematically investigate this new perspective, we formulated the LLM-specific utility judgment task and constructed a corresponding benchmark to evaluate the ability of LLMs to identify passages utilitarian to themselves. We comprehensively benchmarked three families of utility judgment methods: verbalized, likelihood-based, and attention-based, across multiple LLMs and datasets. Verbalized methods generally have better performance. Attention has the worst performance. Furthermore, the observed performance degradation on known queries, even given highly relevant passages like human-annotated passages. This indicates that the aim of effective utility judgments requires LLMs to not only select useful passages for unknown queries but also to reject all passages when their internal knowledge is sufficient. For future work, critical directions include: 1) Designing more sophisticated LLM-specific utility judgment methods that can truly discern model-specific needs and accurately handle the known/unknown query dichotomy; 2) Developing efficient and lightweight mechanisms to personalize retrieval for target LLMs.

References

[1] Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019. Multi-step retriever-reader interaction for scalable open-domain question answering. *arXiv preprint arXiv:1905.05733* (2019).

[2] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[4] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 6609–6625. https://www.aclweb.org/anthology/2020.coling-main.580

[5] Gautier Izacard and Edouard Grave. 2020. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584* (2020).

[6] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research* 24, 251 (2023), 1–43.

[7] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[8] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 7969–7992.

[9] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551* (2017).

[10] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering.. In *EMNLP (1)*. 6769–6781.

[11] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024* (2022).

[12] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406* (2022).

[13] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

[14] Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2023. Llatrieval: Llm-verified retrieval for verifiable generation. *arXiv preprint arXiv:2311.07838* (2023).

[15] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374* (2023).

[16] Multi-Linguality Multi-Functionality Multi-Granularity. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation.

[17] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset. (2016).

[18] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2020. KILT: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252* (2020).

[19] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350* (2022).

[20] Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2395–2400.

[21] Alireza Salemi and Hamed Zamani. 2025. Learning to rank for multiple retrieval-augmented models through iterative utility maximization. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*. 183–193.

[22] Tefko Saracevic, Paul Kantor, Alice Y Chamis, and Donna Trivison. 1988. A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information science* 39, 3 (1988), 161–176. https://www.researchgate.net/publication/245088184_A_Study_in_Information_Seeking_and_Retrieving_I_Background_and_Methodology

[23] Linda Schamber and Michael Eisenberg. 1988. Relevance: The Search for a Definition. (1988).

[24] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294* (2023).

[25] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652* (2023).

[26] Qwen Team. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

[27] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. *arXiv preprint arXiv:1803.05355* (2018).

[28] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[29] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509* (2022).

[30] Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571* (2024).

[31] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).

[32] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966* (2023).

[33] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600* (2018).

[34] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

[35] Hamed Zamani and Michael Bendersky. 2024. Stochastic rag: End-to-end retrieval-augmented generation through expected utility maximization. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2641–2646.

[36] Hengran Zhang, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. Iterative Utility Judgment Framework via LLMs Inspired by Relevance in Philosophy. *arXiv preprint arXiv:2406.11290* (2024).

[37] Hengran Zhang, Keping Bi, Jiafeng Guo, Jiaming Zhang, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. 2025. Distilling a Small Utility-Based Passage Selector to Enhance Retrieval-Augmented Generation. *arXiv preprint arXiv:2507.19102* (2025).

[38] Hengran Zhang, Minghao Tang, Keping Bi, Jiafeng Guo, Shihao Liu, Daiting Shi, Dawei Yin, and Xueqi Cheng. 2025. Utility-Focused LLM Annotation for Retrieval and Retrieval-Augmented Generation. *arXiv preprint arXiv:2504.05220* (2025).

[39] Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023. From relevance to utility: Evidence retrieval with feedback for fact verification. *arXiv preprint arXiv:2310.11675* (2023).

[40] Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Are Large Language Models Good at Utility Judgments?. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1941–1951.

[41] Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, and Jie Tang. 2024. Longrag: A dual-perspective retrieval-augmented generation paradigm for long-context question answering. *arXiv preprint arXiv:2410.18050* (2024).

Pointwise-Set: with Pseudo-Answer

Question: {question}
Passage: {passage}
The reference answer: {answer} Determine if the passage has utility based on two strict criteria:
1. Usefulness: The passage should not only be relevant to the question but also be useful in generating a correct, reasonable, and perfect answer to the question.
2. Novelty: Is the useful information new to you? This means it must NOT be part of your pre-existing knowledge.
Directly output your response. The format of the output is: 'Utility judgment: Yes/No.'

Listwise-Set: with Pseudo-Answer

User: I will provide you with {num} passages, each indicated by number identifier []. I will also give you a reference answer. Select the passages that have utility for yourself in answering the question: {question}.
Assistant: Okay, please provide the passages and the reference answer.
User: [{rank}] {content}
Assistant: Received passage [{rank}].
…
User: Question: {question}.
Reference answer: {answer}
Determine if the passage has utility based on two strict criteria:
1. Usefulness: The passage should not only be relevant to the question but also be useful in generating a correct, reasonable, and perfect answer to the question.
2. Novelty: Is the useful information new to you? This means it must NOT be part of your pre-existing knowledge.
Directly output the passages you selected that have utility for yourself in answering the question. The format of the output is: 'My selection:[[i],[j],...].'. Only response the selection results, do not say any word or explain.

Figure 5: The prompt of pointwise and listwise self-utility judgment.

User: I will provide you with {num} passages, each indicated by number identifier []. I will also give you a reference answer. Rank the passages based on the passages' utility for yourself in answering the question: {question}.
Assistant: Okay, please provide the passages and the reference answer.
User: [{rank}] {content}
Assistant: Received passage [{rank}].
…
User: Question: {question}.
Reference answer: {answer}
Determine if the passage has utility based on two strict criteria:
1. Usefulness: The passage should not only be relevant to the question but also be useful in generating a correct, reasonable, and perfect answer to the question.
2. Novelty: Is the useful information new to you? This means it must NOT be part of your pre-existing knowledge.
Directly output the ranked the passages in descending order of utility for yourself in answering the question. The format of the output is: '[i]>[j]>...'. Only response the ranked results, do not say any word or explain.

Figure 6: The prompt of verbalized listwise ranking.

You are an expert information extraction system. Your task is to analyze the provided Query and Long Answer to identify and extract every possible concise short answer with absolute precision.
The answer must be a concise phrase or named entity (e.g., a person, place, date, or value), not a complete sentence. Output Format: Respond only with the extracted short answer list, nothing else.
Example:
Query: What is Paula Deen's brother?
Long Answer: Earl W. Bubba Hiers is Paula Deen's brother.
Short Answer: ['Earl W. Bubba Hiers', 'Bubba Hiers']
Extract the short answer from the long answer that directly and precisely responds to the query. The answer should be a concise phrase or named entity (e.g., a person's name, a place, a date, a value), not a full sentence. Query:{query}\n Long Answers: {answer}\n Short Answer:[

Figure 7: The prompt for Qwen3-32B to extract entity answer on MS MARCO-FQA dataset.

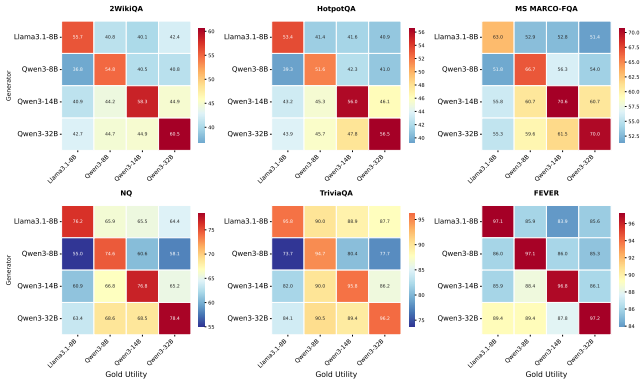

Figure 8: RAG performance (%) of LLMs with golden utility set (top-20 retrieval candidate) from different LLMs.

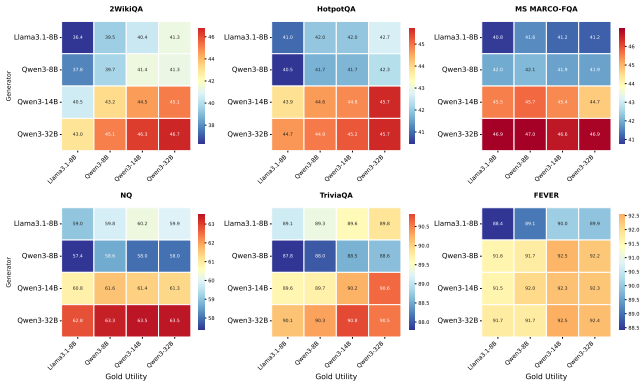

Figure 9: RAG performance (%) of LLMs with utility judgments results (verbalized listwise $w/$ pseudo-answer) from different LLMs.

Table 5: Utility alignment performance (F1, %) of different verbalized methods for LLMs. "(D)" or "(T)", means the pseudo-answer is generated directly from top-20 results or generated from top-20 results with thinking.

LLM	Type	Method	NQ
Qwen3-14B	Pointwise	Direct	49.66
		$w/$ Thinking	49.94
		$w/$ Pseudo-answer (D)	57.86
		$w/$ Pseudo-answer (D) and Thinking	53.23
		$w/$ Pseudo-answer (T) and Thinking	**58.35**
	Listwise	Direct	55.14
		$w/$ Thinking	49.53
		$w/$ Pseudo-answer (D)	57.27
		$w/$ Pseudo-answer (D) and Thinking	51.84
		$w/$ Pseudo-answer (T) and Thinking	**57.85**

Table 6: Dataset statistics. "M-FQA" means the MS MARCO-FQ dataset.

	NQ	HotpotQA	2WikiQA	TriviaQA	M-FQA	FEVER
#Queries	2837	5600	12576	5359	3199	10245
#Qrels/q	3	2	-	2.4	1.1	1.3
#Corpus	32M	32M	32M	32M	8.8M	32M

A Prompts

This section will show the detailed prompts used in this work. Figure 5 and 6 show the prompts designed for self-utility judgments for LLMs. The pseudo-answer is injected for self-utility judgments. Figure 7 shows the prompt designed for Qwen3-32B to extract entity answers on the MS MARCO-FQA dataset. About the answer generation prompt, if no passages are given, the instruction is "`Answer the following question based on your internal knowledge`", otherwise, "`Information: {Passages} Answer the following question based on the given information or your internal knowledge`".

B Thinking in Utility Judgments

To analyze the impact of reasoning on utility alignment, we enabled the "thinking" for the LLMs during self-utility judgments on the NQ dataset using Qwen3-14B. As shown in Table 5, we found that this reasoning process did not yield greater alignment performance gains compared to pseudo-answer injection. Given that the reasoning capability of LLMs incurs significant computational cost, we disabled it for the Qwen3 family models in our subsequent experiments.

C Dataset statistics

Table 6 shows the detailed Dataset statistics for all datasets.

D RAG Performance

D.1 Transfer Experimental on Different Golden Utility Passages from Different LLMs

Figure 8 shows transfer experiments for different golden utility sets of LLMs. We can find that the golden utility set cannot be shared with other LLMs, which is similar to the union candidate setting.

D.2 Transfer of Utility Judgments for Specific LLM

We analyzed the transfer of passages from SOTA utility judgments methods (verbalized selection with pseudo-answer) between LLMs, we observed a markedly different outcome compared to golden utility sets of LLMs. As illustrated in Figure 9, these self-judgments are highly shareable. Notably, Llama3.1-8B's performance improves when it uses the utility judgments from Qwen3-14B or Qwen3-32B, surpassing its performance with its own judgments. This phenomenon opposes the initial hypothesis that retrieval must be fine-tuned to the specific utility signals of each LLM.

D.3 RAG Performance of Utility-based Ranking

Table 7 shows the RAG performance with different top-5 results ranked by self-utility-based ranking methods. We can observe that: (1) RAG performance with selection results in Table 3 generally has better performance ranking results, which is consistent with previous work [37]. (2) On known queries, the RAG performance degrades compared to directly answering for all LLMs, indicating that LLMs cannot reject other passages when the query can be answered without any passages.

Table 7: RAG performance (%) with different top-5 re-ranked results from different self-utility-based ranking methods. "Unk" and "Unknown" are defined in Table 3.

LLM	Method	NQ		HotpotQA		2wikiQA		TriviaQA		MS MARCO-FQA		FEVER		ALL
		Unk	Known	Unk	Known	Unk	Known	Unk	Known	Unk	Known	Unk	Known	AVG
Llama3.1-8B	Likelihood	42.40	83.17	26.73	**80.43**	24.78	61.29	**70.58**	**94.63**	**30.73**	76.96	69.65	92.37	62.81
	Verbalized (w/ Answer)	**44.11**	**84.05**	**27.99**	79.48	**24.91**	**62.27**	69.82	94.29	30.65	**78.59**	**72.72**	**93.71**	**63.55**
Qwen3-8B	Likelihood	46.63	83.88	27.10	84.08	22.01	70.63	75.67	93.53	32.35	75.67	76.28	**96.10**	65.33
	Verbalized (w/ Answer)	**49.44**	**84.45**	**29.32**	**86.70**	**27.42**	**72.10**	**76.77**	**94.32**	**33.71**	**76.21**	**78.37**	95.87	**67.06**
Qwen3-14B	Likelihood	47.59	87.82	28.85	85.33	24.97	76.83	77.14	**95.65**	34.02	79.31	75.49	95.70	67.39
	Verbalized (w/ Answer)	**49.32**	**88.28**	**31.35**	**86.61**	**29.53**	**78.37**	**78.08**	95.56	**36.26**	**81.12**	**77.61**	**96.33**	**69.03**
Qwen3-32B	Likelihood	48.03	86.22	30.34	86.81	29.43	73.14	75.89	95.47	34.66	82.13	70.43	96.11	67.39
	Verbalized (w/ Answer)	**50.62**	**86.73**	**32.27**	**87.38**	**32.94**	**75.98**	**76.93**	**95.86**	**35.49**	**84.68**	**74.19**	**96.80**	**69.16**

Table 8: Utility judgments performance (%) and average number of selected passages of listwise verbalized with pseudo-answer methods. "/" means (w/o / $w/$ Known-Rejection instruct). The average number of selected passages was calculated separately for known and unknown queries. "Precision", "Recall", and "F1" are computed on the queries for which the gold utility is not empty.

Dataset	LLM	Precision	Recall	F1	AVG Count (No Empty)	AVG Count (Empty Gold Utility)
NQ	Llama3.1-8B	39.48 / 40.66	68.40 / 75.44	50.06 / 52.84	7.45 / 6.94	6.84 / 8.01
	Qwen3-8B	50.27 / 50.29	69.58 / 68.61	58.37 / 58.04	6.34 / 6.30	5.30 / 5.22
	Qwen3-14B	49.42 / 49.78	68.07 / 67.33	57.27 / 57.24	5.78 / 5.75	4.95 / 4.86
	Qwen3-32B	53.68 / 55.00	58.25 / 57.52	55.87 / 56.23	4.73 / 4.58	3.94 / 3.82
HotpotQA	Llama3.1-8B	46.29 / 46.69	60.33 / 58.54	52.38 / 51.94	4.92 / 4.57	4.53 / 4.28
	Qwen3-8B	54.44 / 54.47	59.32 / 58.01	56.78 / 56.18	3.63 / 3.47	3.54 / 3.44
	Qwen3-14B	51.27 / 51.70	62.68 / 62.15	56.40 / 56.44	3.94 / 3.83	3.75 / 3.68
	Qwen3-32B	56.69 / 57.49	55.35 / 53.31	56.01 / 55.32	3.07 / 2.92	2.92 / 2.77
2WikiQA	Llama3.1-8B	42.87 / 42.14	37.56 / 34.88	40.04 / 38.17	3.65 / 3.26	3.72 / 3.35
	Qwen3-8B	42.13 / 42.33	34.53 / 33.83	37.95 / 37.61	2.40 / 2.32	2.65 / 2.57
	Qwen3-14B	40.36 / 40.33	40.98 / 39.87	40.67 / 40.10	2.78 / 2.61	2.93 / 2.84
	Qwen3-32B	46.04 / 46.68	35.18 / 34.29	39.88 / 39.54	2.20 / 2.09	2.46 / 2.36
TriviaQA	Llama3.1-8B	61.34 / 61.86	59.53 / 58.13	60.42 / 59.94	8.81 / 8.06	5.90 / 5.71
	Qwen3-8B	72.05 / 72.14	67.42 / 66.53	69.66 / 69.22	7.70 / 7.57	5.54 / 5.43
	Qwen3-14B	73.90 / 73.98	65.96 / 65.82	69.70 / 69.66	6.92 / 6.92	4.93 / 4.91
	Qwen3-32B	77.56 / 77.89	55.38 / 54.10	64.62 / 63.85	6.09 / 6.00	4.11 / 4.00
MS MARCO-FQA	Llama3.1-8B	33.30 / 33.31	80.70 / 79.07	47.15 / 46.88	11.54 / 10.96	11.31 / 9.80
	Qwen3-8B	38.88 / 38.92	79.56 / 80.02	52.24 / 52.37	10.65 / 10.77	8.90 / 9.02
	Qwen3-14B	37.03 / 37.56	77.65 / 77.44	50.15 / 50.59	9.82 / 9.76	8.30 / 8.26
	Qwen3-32B	43.21 / 43.68	69.74 / 68.50	53.36 / 53.34	8.16 / 8.03	6.95 / 6.77
FEVER	Llama3.1-8B	62.51 / 62.37	46.36 / 43.97	53.24 / 51.58	8.14 / 7.63	7.39 / 7.04
	Qwen3-8B	69.74 / 70.12	51.46 / 50.76	59.22 / 58.89	6.55 / 6.29	5.28 / 5.12
	Qwen3-14B	69.92 / 70.47	58.40 / 58.07	63.64 / 63.67	6.11 / 5.94	5.09 / 5.02
	Qwen3-32B	71.67 / 71.06	31.91 / 31.26	44.15 / 43.42	4.96 / 4.92	3.94 / 3.87

E Known-Rejection

Due to the fact that LLMs still select useful passages when the query can be answered without any passages, we all the known-rejection prompts after the definition in promot, i.e., "`If you can answer the question without the passages, all the passages do not have utility for you.`". The Table 8 shows the detailed utility alignment performance and numbers comparison. We can observe that: Though the over-select prompt can be mitigated, the recall metric is generally worse than not adding the prompt, resulting in the F1 performance being degraded. Therefore, it is a huge challenge for LLMs to reject other passages when the query is already known and select more useful passages when they cannot answer the queries.