EXAM++: LLM-based Answerability Metrics for IR Evaluation

Naghmeh Farzi[1], Laura Dietz[1]

[1]University of New Hampshire, 33 Academic Way, Durham, NH, USA

Abstract

Large language models provide an opportunity for reliable and efficient information retrieval evaluation methods. However, current evaluation metrics fall short in accurately assessing the information content of systems' responses—without resorting to expensive human judgments.

In contrast, the EXAM++ Answerability Metric leverages a bank of query-related exam questions to quantify relevant information content that is covered in the systems' responses. The process involves (1) decomposing the query into detailed questions, (2) checking each for answerability with passages in the system response, and (3) devising evaluation metrics based on this information. Using the TREC Complex Answer Retrieval benchmark, we demonstrate that our LLM-based EXAM++ approach works successfully, outperforming several established baselines. In particular, we take a deep dive into different approaches to determine the answerability of questions in a given passage, including the use of question answering systems with answer verification and self-rated answerability determination.[1]

Keywords

Information Retrieval Evaluation, Large Language Models

1. Introduction

Large Language Models (LLMs) can generate and/or retrieve responses for search queries, resulting in many systems that combine traditional retrieval with neural ranking and natural language generation. Ideally, the systems' responses cover relevant information content while being concise and complete. However, there is a need for convincing evaluation metrics to assess the accuracy and completeness of the information content in responses. This should be accomplished in a repeatable and reusable manner and without resorting to expensive human judgments.

To address this scenario, Sander and Dietz [1] proposed the EXAM Answerability Metric, which evaluates retrieval/generation systems based on whether they retrieve passages that answer a set of query-specific exam questions. Given a test bank of exam questions, they automate the work-intensive part of scanning each passage for answers using an automated question answering system.

[1]Data and code available at https://github.com/TREMA-UNH/exam_plusplus

LLM4Eval: The First Workshop on Large Language Models for Evaluation in Information Retrieval, 18 July 2024, Washington DC, United States

Naghmeh.Farzi@unh.edu (N. Farzi); dietz@cs.unh.edu (L. Dietz)

https://www.cs.unh.edu/~dietz (L. Dietz)

0009-0000-3297-8888 (N. Farzi); 0000-0003-1624-3907 (L. Dietz)

With EXAM++, we significantly expand on Sander's idea by

- supporting the development of exam question banks with prompt-based generation,
- modernizing the question answering system with the recently released FLAN-T5 family [2],
- exploiting abilities of modern LLMs to determine the answerability of questions,
- offering relevance labels that are inter-operable with commonly used evaluation tools (e.g. `trec_eval`).

A strength of EXAM++ is that, in contrast to other work on LLM-based relevance grading, we can readily integrate humans into the evaluation by having them manage the design of the test bank of exam questions. The test questions should be designed to cover all relevant facets of a query, so that the more questions are addressed, the more relevant a passage is. Based on the long history of classroom education and exam design, we argue it is more natural for human judges to control the design of exam questions than to directly provide relevance judgments.

By virtue of automating the grading of system responses, human judges are never required to perform passage-level relevance assessments. At the same time, humans are fully in control of defining which information content is relevant via the exam question bank.[1]

The evaluation approach yields reusable test collections that can be expanded by modifying the question bank at any point in the evaluation process, as the remaining pipeline is fully automated. The impact of a question bank modification can be directly observed by listing passages whose relevance grade would change.

Contributions. In this paper, we provide an in-depth study analyzing different choices of the EXAM++ approach: Automatic vs. manual test banks, predicted relevance labels with traditional evaluation metrics vs. coverage-based measures, impacts of fine-tuning vs. prompt engineering, grading via self-ratings vs. via question answering systems with different answer verification approaches. While EXAM++ is identical to the question-based RUBRIC evaluation method [4], in this paper we provide an in-depth comparison on different question-answering approaches. Additionally, we compare to the original EXAM method [1] and several direct grading prompts [5, 6, 7, 8].

2. Related Work

We focus on an approach that does not require passage-level relevance judgments or source texts. Our work is unique in this regard, but aspects relate to many active branches of research, which we detail below.

2.1. LLM-based Relevance Label Predictors

In contrast to our approach, several LLM-based evaluation approaches attempt to directly imitate the relevance judgment process.

Sun et al. [5] rerank passages using a simple LLM prompt "does the passage answer the query?" Faggioli et al. [6] conduct an early evaluation experiment by asking an LLM to judge

[1]We recently released a resource to support human judges in supervising this process [3] https://github.com/TREMA-UNH/rubric-grading-workbench.

the relevance of a passage. They design a simple prompt and a more elaborate multi-relevance few-shot prompt developed for the TREC Deep Learning track. Thomas et al. [8] compare the ability of LLMs to perform document-level relevance judgments in comparison to different groups of human annotators. In their study they use a detailed prompt that instructs the LLM to respond with a multi-level relevance grade. We include several of these prompts in our empirical evaluation.[2]

In 1SLs, MacAvaney and Soldaini [9] focus on evaluating passages with a DuoPrompt, that instructs an LLM to indicate which of two passages is more relevant for a query.

However, several critiques have been raised about using LLMs for producing relevance labels in general. Faggioli et al. [6, 10] elaborates a wide range of theoretical concerns, centered on questions of trustworthiness and reliability of LLMs now and in the future. Liu et al. [11] demonstrate that evaluator-LLMs assign a higher score to systems that use the same LLM model. Wang et al. [12] empirically demonstrate that LLMs exhibit unfair positional bias towards candidates displayed for evaluation. Fok and Weld [13] studies general issues of human over-reliance and under-reliance on LLMs. They elaborate why rationales produced by LLMs for human verification do not generally lead to improvements.

2.2. Evaluation with Test Questions

The idea of anchoring an evaluation on a bank of test questions has been widely discussed in literature on summarization [14], recently with automated question answering methods. Eyal et al. [15] suggest a system evaluation score that is based on the number of questions that a question answering system can correctly answer using the system response—a principle that both the original EXAM method and our approach follow.

Many approaches use a Cloze-style approach to generate questions from a given gold summary or source text. Questions can be in the form of multiple-choice questions [16], free text questions with exact-match answer verification [17], or be derived from extracted entities and relations [18, 15].

As it pertains to information retrieval evaluation, the problem with generating questions from a given source text or gold summary is that (1) such a gold standard is usually not available and (2) it is unclear which of these questions relate to *relevant information* in the gold summary (or source text).

The original EXAM method avoids this problem altogether by asking a human to design questions that address the search query. In contrast, we propose to automatically generate questions directly from the query, building on the world-knowledge of GPT [19]—with the intention of employing manual labor to verify or weight the question set.

3. Background: Original EXAM

The original EXAM method [1] uses a query-specific test bank of exam questions harvested from school textbooks in the Textbook Question Answering (TQA) dataset [20]—a dataset from

[2]All baseline prompts are provided in our online appendix.

which topics for the TREC CAR Y3 evaluation were derived. Furthermore, they use a custom question answering system that is optimized to answer multiple-choice questions in the style of TQA questions.

Their approach considers each passage retrieved by a system submitted to TREC CAR Y3 and uses the automated question answering system to extract answers for all test questions. Each of these answers is verified against the answer key for each exam question—tracking correctly answered questions. The system's evaluation score is based on the set of questions that is correctly addressed with any of the top 20 passages—averaged across all queries. The more questions can be answered, the higher the EXAM evaluation score for that system.

The original EXAM method relies solely on humans to design the exam, with the intent that only a human could identify the core questions that would need to be addressed in a relevant answer. This is in contrast to approaches that generate questions from a gold summary (detailed in Section 2.2), which might lead to questions derived from non-relevant aspects mentioned in relevant text.

4. Approach

In this work we explore a modernized version of Sander's EXAM Answerability metric, which we call EXAM++.[3] Akin to Sander's method, we use a bank of exam questions to grade systems based on the set of questions that can be correctly answered with information in the system's response. The more questions can be answered with the system's response, and the more passages answer questions well, the higher the EXAM evaluation score of the system. By automating the component that determines the answerability of passages, the evaluation paradigm becomes repeatable and reusable at a reasonable cost. As a result, it can be applied to systems that retrieve passages from a corpus as well as systems that generate content with LLMs.

Our EXAM++ evaluation system assumes the following inputs:

1. A set of queries, optionally with query subtopics.
2. A set of system responses, which can come in the form of a passage ranking or a set of generated passages.

Note that the exam questions are intended to be kept secret from the retrieval/generation system, only to be used for evaluation.

The EXAM++ evaluation approach is structured into the following three phases that we detail in the remainder of the section and depict in Figure 1.

1. **Obtaining an exam question bank:** A process of creating a test bank of query-specific exam questions.
2. **Grading system responses:** All passages in system responses are graded using an automated LLM-based system to determine which questions are answerable with the passage content. For each passage, the set of answerable questions is tracked along with grades that represent how relevant, complete, and accurate the provided answer is.

[3] An implementation of EXAM++ is available in the Autograding Workbench [3].

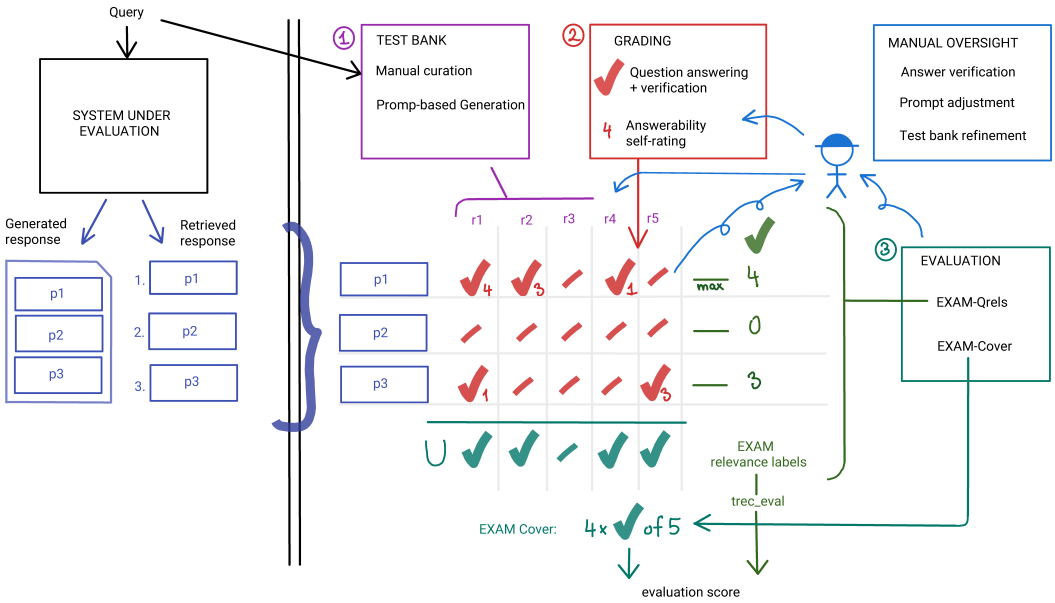

Figure 1: EXAM++ approach. Left: The system under evaluation retrieves or generates passages $p \in P$ in response to queries q (blue). The system does not have access to exam questions. Passages from all systems will be pooled for assessment, and additional passages can be added later as new systems are developed.
Right: The EXAM++ evaluation system uses three phases detailed in Section 4. For each query q, an exam question bank R_q is developed, which can be modified later in an iterative fashion (purple). All passages from the system response (e.g., p1, p2, p3) are graded based on which questions (r1, r2, ..., r5) can be correctly answered with the passage text (red). We support two modes: one where answers are verified against an answer key (depicted as check marks), or by having an LLM self-rate the answerability on a scale from 0 to 5. The EXAM++ evaluation scores are derived from these grades (green). The EXAM-Cover score is based on how many questions are covered, as binary verification or via a minimum self-rating level. For EXAM-Qrels a relevance file for `trec_eval` is derived, which is based on the coverage or best self-rating obtained by this passage in isolation.
A human-in-the-loop is ensuring that grade annotations correlate with relevant passages and will improve the test bank in response and adjust the grading system where necessary (cyan). We provide a worked example in Section 5.6.

3. **EXAM evaluation scoring:** We derive multiple evaluation scores. The more exam questions can be answered well with passages of a systems' response, the higher the system's EXAM-Cover score. The more passages address any of the exam questions well, the higher the system's precision-oriented EXAM score. By exporting passage-level relevance labels, any traditional evaluation metric can be incorporated (we refer to this evaluation score as EXAM-Qrels).

Our contribution differs from the original EXAM method in several important ways:

A. Obtaining an exam question bank: To obtain exam questions,

- The original EXAM method is based on manually created multiple-choice exam questions.

Table 1
Phase 1. Question bank generation prompt.

TREC CAR Y3 Question Bank Prompt

Explore the connection between '{query_title}' with a specific focus on the subtopic '{query_subtopic}'. Generate insightful questions that delve into advanced aspects of '{query_subtopic}', showcasing a deep understanding of the subject matter. Avoid basic or introductory-level inquiries. Give the question set in the following JSON format:

```
```json
{"questions":[question_text_1, question_text_2,...]}
```
```

- We propose to semi-automatically generate free-text questions for each query, as described in Section 4.1.1.

B. Grading system responses: To grade each passage via the answerability of exam questions,

- The original EXAM method uses a pre-neural multiple-choice question answering system with answer verification.
- First, we modernize the question answering system with an LLM-based approach (Section 4.2.1).
- Second, we explore the ability of LLMs to self-rate the answerability of a question with given context, without directly verifying the correctness of answer (Section 4.2.2).

C. EXAM-Cover evaluation: To evaluate each IR system,

- With EXAM-Cover, we follow the original EXAM method by evaluating systems according to the number of answerable exam questions (Section 4.3.1).
- To improve adoption, we add a variant "EXAM-Qrels" that implements a related idea so that it is inter-operable with the popular evaluation tool `trec_eval` (Section 4.3.2).

4.1. Phase 1: EXAM++ Question Banks

4.1.1. Generating Question Banks

We use a generative LLM, specifically ChatGPT, to automate the creation of free-text questions[4] that are directly tailored to the needs of information retrieval (IR) tasks and specific domain requirements. This approach allows a larger information to be broken down need into insightful and relevant questions that probe deeply into the nuances of each query, enhancing the depth and quality of the question banks. With application to TREC CAR Y3, a set of open-ended questions R_q are generated for each subtopic, via a zero-shot prompt as detailed in Table 1.

The goal during topic development is to have a human judge ensure that essential information about the query is covered by the question bank, and (if necessary) modify the questions accordingly.

[4]This step is identical to generating question-based RUBRICs in Farzi and Dietz [4].

Table 2
Phase 2. Grading prompts. The question answering prompt extracts an answer to an exam question from the passage, to be verified with word matching or the verification prompt. Alternatively, the answerability can be self-rated by the LLM without explicitly extracting the answer.

Question Answering Prompt provide a complete and concise answer to the question based on the context. Question: {question} Context: {context}
Optional: Answer Verification Prompt For the question "{question}" the correct answer is "{correct_answer}". Is "{answer}" an equally correct response to this question? Answer yes or no.
Self-rating Answerability Prompt Can the question be answered based on the available context? choose one: - 5: The answer is highly relevant, complete, and accurate. - 4: The answer is mostly relevant and complete but may have minor gaps or inaccuracies. - 3: The answer is partially relevant and complete, with noticeable gaps or inaccuracies. - 2: The answer has limited relevance and completeness, with significant gaps or inaccuracies. - 1: The answer is minimally relevant or complete, with substantial shortcomings. - 0: The answer is not relevant or complete at all. Question: {question} Context: {context}

4.1.2. Manual Question Banks

Alternatively, query-specific question banks R_q can be manually constructed from scratch. Optionally this can include a gold answer key for verification, as described in the original EXAM method, which uses such a test bank from the TQA dataset.

4.2. Phase 2: Automated EXAM++ Grading

The grading process leverages a state-of-the-art LLM, such as the `FLAN-T5-large` [2] model, chosen to trade-off processing speed and ability to understand complex queries and context. Prompts in Table 2 have been designed for reliable exam grading—especially so that the LLM focuses solely on the provided context rather than relying on its pre-trained knowledge. The LLM is queried separately for each passage to prevent positional biases, ensuring that each answer is contextually derived from the passage to which it corresponds.

Pre-processing system responses. Before grading, a judgment pool of all retrieved passages is created for efficient processing. Longer system responses are segmented into paragraph-sized passages P. Each passage is given a unique identifier (`passage_id`) to ensure that every part of the response can be individually traced throughout the grading process.

4.2.1. LLM-based Question Answering with Answer Checking

For every passage-question pair (p, r), we ask the LLM to extract a best effort answer from the passage. We use the prompt in Table 2 (top) with a text-to-text generation pipeline.

Once answers are extracted, they are verified for correctness against the answer key. The verification process will normalize the correct and predicted answers through lower-casing, stopword removal, and stemming. We then apply a heuristic matching function where a match is considered valid if the edit distance between normalized answers is less than 20% of the length of the longer string.

Occasionally the LLM will respond with an expression indicating that the question is unanswerable with the provided context. We count an answer as incorrectly answered (grade 0) when we encounter an ill-formed answer (such as "a." or "(iii)") or one of the following expressions: "unanswerable", "no", "no answer", "not enough information", "unknown", "it is not possible to tell", "it does not say", or "no relevant information".

Variation: SQuAD2 fine-tuning. We study the impacts of fine-tuning the question answering system using the SQuAD2 dataset [21]. SQuAD2 is comprised of questions in a similar style to TQA, to be answered in the context of a provided passage. SQuAD2 also includes many training examples where questions are unanswerable with the given context, which is essential to determine the answerability of questions for EXAM++.

Variation: Answer verification with LLMs. The implementation of the answer verification remains a technical challenge. Noticing that many correct answers are missed because they are phrased differently, we additionally explore asking the LLM to verify the answer match. We verify with the prompt in Table 2 (middle) by providing the extracted answer, the gold answer, and the question.

We manually analyzed the accuracy of this verification step, based on extracted and correct answers. To give an example from TQA exam question `L_0016/NDQ_000615` "During very wet times, the water table will..." for which the correct answer is "rise", this LLM-based process identifies additional answers including "increase", "rising", "be higher", "increase substantially" as well as answers that restate the question such as "During very wet times, the water table will rise."

4.2.2. Grading by Self-rating Answerability

Given the technical challenges of answer verification we explore an easier alternative. We use an answerability system introduced as RUBRIC in Farzi and Dietz [4], that self-rates whether the passage p answers the question $r \in R_q$, without first extracting the answer.

Given each passage-question pair, the LLM rates the answerability on a scale from 0 (worst) to 5 (best) using the prompt provided in Table 2, bottom. In cases where the LLM does not provide a numerical rating, we default to a rating of 1 for answered questions—with the exception of answers that denote unanswerability (as in Section 4.2.1) for which we assign a grade of 0.

This method enables an autonomous assessment of answerability and relevance, avoiding technical issues of answer verification when there are different ways to phrase a correct answer or if there are different answers that are equally correct. Moreover, this supports the use of open-ended questions for evaluation.

The output of the grading phase is, for each passage-question pair (p, r), a grade that represents the relevance, completeness, and accuracy with which the question is addressed. The grade is 0 if the passage does not address the answer. For question answering with answer verification, the grade is either 1 (if correct) or 0 otherwise. In addition, we track the extracted answer to support manual verification via human judges.

4.3. Phase 3: EXAM++ Evaluation

4.3.1. EXAM-Cover Evaluation

We incorporate a coverage-style evaluation metric as suggested by Sander et al. It quantifies the set of exam questions $r \in R_q$ for the query q that are covered in retrieved passages $p \in P$ with a minimum grade level τ, as defined by:

$$\text{EXAM-Cover}_\tau(P) = \frac{1}{|R_q|} \left| \bigcup_{p \in P} \{r | \text{grade}(p, r) \geq \tau, \forall r \in R_q\} \right| \tag{1}$$

To avoid gaming the evaluation metric with a very long system response, the size of the passage set P is limited to a fixed budget, e.g. $k = 20$ passages.

4.3.2. EXAM-Qrels Evaluation

Alternatively, we provide relevance labels for each passage facilitating compatibility with traditional IR evaluation metrics, such as implemented in the `trec_eval` tool. Passage-level relevance labels are obtained by mapping grades to a binary or multi-graded relevance label,

$$\text{EXAM-Label}(p) = \max_{r \in R_q} \text{grade}(p, r) \tag{2}$$

The EXAM-Label allows to use established IR evaluation metrics that incorporate multi-graded relevance labels (such as NDCG), or by choosing a minimum grade indicating relevance, τ, to control the leniency of the evaluation.

Like all relevance-label based approaches, the pool of graded passages may impact the evaluation results—therefore, as systems reveal unjudged passages, these should be graded to update the qrel files.

The downside of this EXAM-Qrels approach is that once a relevance label is determined, the evaluation metric is unaware of which exam questions were covered. To preserve this information, future work should explore integrating EXAM++ with intent-aware evaluation measures such as α-NDCG [22, 23].

Whether EXAM-Cover or EXAM-Qrels is a more appropriate evaluation measure depends on the goals of the information retrieval application. When users are expected to stop after the first relevant passage, then we suggest evaluating with EXAM-Qrels with mean reciprocal rank. When recall is a priority, we suggest using EXAM-Qrels with R-precision or (mean-) average precision (MAP). When the emphasis is on covering diverse facets of relevance, we suggest to use EXAM-Cover.

5. Experimental Evaluation

5.1. Experimental Setup

We experimentally compare variations of our EXAM++ system to the original EXAM method [1]. The evaluation uses queries, manual TREC judgments, and submitted systems from the third year of the TREC Complex Answer Retrieval track (TREC CAR Y3) [24],[5] as these align with manual test questions and results from Sander and Dietz [1]. Empirical results on other datasets are available in Farzi and Dietz [4].

In experiments with generated question banks, we follow "Phase 1" to obtain ten questions for each of the 721 query-subtopics across 131 queries in CAR Y3. In experiments that use the manual TQA question bank, we use all non-diagram questions with gold answer keys. In preparation for grading (Phase 2), we build a judgment pool of all passages in official judgments and the top 20 of all run submissions—a total of 85,329 passages.

For question generation we use `gpt-3.5-turbo-instruct`; for question verification and self-rating we use the `FLAN-T5-large` model with the `text2text-generation` pipeline from HuggingFace.[6] We also explore fine-tuning the FLAN-T5-large on the SQuAD2 dataset. The fine-tuned model is available on HuggingFace as `sjrhuschlee/flan-t5large-squad2` to be used with the extractive `question-answering` pipeline.

We compare the following variations of our approach.

EXAM++: Using our generated question banks and grading with self-ratings (Sections 4.1.1 and 4.2.2).

Manual EXAM++: As previous but using manual question banks from the TQA dataset [20] (Sections 4.1.2 and 4.2.2).

Manual-EXAM-QA: As previous but grading via question answering using prompts from Table 2 (top), with word-based answer checking (Sections 4.1.2 and 4.2.1).

Manual-EXAM-Squad2: As previous but fine-tuning the grading LLM on SQuAD2 and using the `question-answering` pipeline of a prompt.

LLM-verified Manual-EXAM-QA & Manual-EXAM-Squad2: Like the two previous but the extracted answers are verified with the `FLAN-T5-large` LLM using the answer verification prompt from Table 2 (middle).

For all these methods we compare both the EXAM-Qrels and EXAM-Cover evaluation approach. For EXAM-Qrels, we export passage-level EXAM++ relevance labels to be used with `trec_eval` on traditional evaluation measures. In this experiment we use measures used in the official TREC CAR Y3 evaluation, such as average precision (MAP), normalized cumulative discounted gain (NDCG@20), and R-precision (Rprec).

We compare to the following reference baselines.

Original EXAM: Using the results provided by Sander et al [1].

[5]The TREC CAR Y3 test set `benchmarkY3test` is available at http://trec-car.cs.unh.edu/datareleases/
[6]https://huggingface.co/google/flan-t5-large

Table 3
Rank correlations of each evaluation method with different minimum grades τ with the official TREC CAR Y3 leaderboard. S: Spearman's rank correlation. K: Kendall's Tau correlation. Best evaluation method in bold-italics. Equally good methods within (±0.05) marked in bold. Poor methods (obtaining less than 0.5) marked in grey. Leaderboards of selected methods (marked in blue) are presented in Figure 2. In one case, all systems obtained a perfect score (marked with †), therefore the rank correlation cannot be computed.

		Relevance-labels for `trec_eval`						Coverage	
		MAP		NDCG@20		RPrec		Cover@20	
Evaluation Method	τ	S	K	S	K	S	K	S	K
Self-rated Answerability									
EXAM++	1	**0.931**	0.808	***0.883***	***0.735***	**0.909**	0.764	0.852	0.694
	3	**0.933**	0.817	***0.883***	***0.735***	**0.910**	0.773	0.837	0.677
	4	**0.933**	0.817	***0.883***	***0.735***	**0.910**	0.773	0.830	0.659
Best overall→	5	***0.980***	***0.902***	***0.883***	***0.735***	***0.959***	***0.860***	0.904	0.738
Manual EXAM++	1	0.710	0.516	0.643	0.467	0.704	0.516	***0.959***	***0.852***
	3	0.704	0.524	0.643	0.467	0.688	0.513	**0.927**	0.782
	4	0.690	0.469	0.643	0.467	0.681	0.516	**0.932**	0.773
	5	0.435	0.266	0.643	0.467	0.432	0.298	0.865	0.738
QA + Answer Verification									
Manual-EXAM-QA		0.716	0.586	0.694	0.502	0.756	0.621	0.824	0.656
Manual-EXAM-Squad2		0.789	0.609	0.727	0.520	***0.858***	***0.711***	0.788	0.633
LLM-verified Manual-EXAM-QA		0.350	0.236	0.182	0.079	0.310	0.236	0.350	0.236
LLM-verified Manual-EXAM-Squad2		0.468	0.331	0.629	0.385	0.365	0.241	0.683	0.546
Original EXAM [1]								0.75	0.57
Direct LLM-based Grading									
Thomas [8]	1	0.666	0.576	0.640	0.537	0.646	0.561	†	†
FaggioliB [6]	1	0.588	0.443	0.582	0.432	0.685	0.511	0.702	0.541
HELM [7]	1	0.550	0.434	0.486	0.362	0.520	0.403	0.584	0.408
Sun [5]	1	0.655	0.510	0.627	0.511	0.677	0.544	0.759	0.590
FaggioliB_few [6]	1	0.284	0.179	0.409	0.293	0.320	0.219	0.646	0.500
Sun_few [5]	1	0.286	0.180	0.286	0.179	0.175	0.096	0.683	0.542

FaggioliB, Sun, HELM, Thomas: Using the same `FLAN-T5-large` LLM as for EXAM++ but obtaining relevance labels by directly asking whether a passage is relevant for a query. We use a set of established prompts [5, 6, 7, 8], listed in Appendix A.

FaggioliB_few, Sun_few: As previous but using few-shot prompts suggested for the TREC Deep Learning track [5, 6] to test their generalizability.

We measure the quality of our evaluation paradigm in two ways:

Leaderboard rank-correlation: The leaderboard of systems under the EXAM-Cover and EXAM-Qrels metric should be similar to the official TREC CAR Y3 leaderboard. This similarity is evaluated with two rank correlation measures: Spearman's rank correlation coefficient, which measures differences of a system's rank on the leaderboard, and Kendall's τ rank correlation which penalizes swaps of two systems on the leaderboard.

Inter-annotator agreement: High passage-level agreement between official judgments and

Table 4
Grade/judgment agreement for EXAM++.

Grade	Judgments		Total	Cohen's κ
	1–3	≤0		
4–5	**1910**	1117	3027	0.38
0–3	880	**2445**	3325	0.37

our predicted relevance labels. We provide count statistics and Cohen's κ inter-annotator agreement statistics.

Since Sander's work demonstrated that ROUGE metrics are uncorrelated with leaderboard rankings, we omit the comparison here.

Significance testing. We perform a standard-error bar overlap test for Figure 2 and only describe significant differences in the text. For leaderboard correlation results in Table 3, we consider results within ±0.05 as equally good.

5.2. Overall Results

EXAM++. Each evaluation method gives rise to a leaderboard of systems. Table 3 compares how well each leaderboard correlates with the official TREC leaderboard. Our proposed EXAM++ with minimum grade $\tau = 5$ obtains overall best results for EXAM-Qrels. In many cases this approach obtains near-perfect rank correlations above 0.9. For reference, rank correlation statistics are on a range from -1 to +1, with 0 indicating no correlation.

Table 4 presents the inter-annotator agreement between manual TREC judgments and predicted relevance labels. We see that especially high self-rating grades obtain a good correlation (Cohen's kappa of 0.38).

For leaderboards based on the EXAM-Cover metric, we also obtain strong results with EXAM++, but observe even better results using the manually created question bank (Manual EXAM++). We believe that the manual control in question bank design would only select vetted questions that represent relevance. We find that some of the generated questions are too broad, promoting systems that provide information that is not sufficiently specific. Future work should focus on adjusting the question bank generation prompt (Table 1) to obtain more focused questions.

QA + answer verification. Next, we turn to EXAM++ approaches that determine relevance by verifying extracted answers from passages against gold answer keys. We find that verifying extracted answers (Manual-EXAM-QA) obtains comparable results to the self-rated answerability approach (Manual EXAM++) when used to obtain relevance labels. However, it is slightly worse when used with coverage-based metrics.

In either case, all our proposed approaches outperform the original EXAM method [1] by using a strong LLM-based question answering method as opposed to a pre-neural question answering system.

LLM-based relevance label predictors. None of the direct grading prompts described in Section 2.1 work well on the TREC CAR Y3 dataset—in many cases obtaining weak rank correlations of below 0.5 (marked in grey). This is in contrast to findings on the TREC DL test collections [5, 6, 8], where these direct grading prompts perform extremely well (both when using GPT [6] and FLAN-T5-large [4]). We suspect that the exact prompts are designed for the for unambiguous narrowly specified question-style queries as found in the TREC DL collection ("When did rock'n'roll begin?") but struggle with the broad information needs in the TREC CAR Y3 collection (e.g., "the integumentary system").

Furthermore, the few shot examples designed for the TREC DL domain (used in `FaggioliB_few`, `Sun_few`) do not generalize to the broad information needs of the TREC CAR Y3 domain. We hope that future research analyzes which of the findings on the DL collection generalize to other information retrieval use cases.

5.3. Obtained System Leaderboards

Figure 2 presents the impact of different evaluation methods on how systems are ranked on the leaderboard. We choose three of the best performing evaluation methods, spanning across our different options (marked in blue in Table 3), namely:

EXAM++ MAP (grade>=5): Generated question bank, self-rated answerability EXAM++ with EXAM-Qrels, `trec_eval` using (mean) average precision, relevant grade ≥ 5.

Manual EXAM++ Cover (grade>=1): Manual question bank, self-rated answerability EXAM-Cover, relevant grade ≥ 1.

Manual-EXAM-Squad2: Manual question bank, using the question answering approach with answer verification on a fine-tuned LLM model, EXAM-Qrels, `trec_eval` using R-precision.

Orig EXAM [1]: As reported in Sander and Dietz [1] as "unnormalized", which is akin to EXAM-Cover.

Official leaderboard (MAP): Manual TREC judgments, (mean) average precision, relevant grade ≥ 1, as reported in the TREC CAR Y3 overview paper [24].

To make the system ranking behavior more visible, all systems' evaluation scores are renormalized so that the highest scores maps to 1.0, and the lowest to 0.0. Several systems use a similar approach, leading to near identical scores on all leaderboards (including the official CAR leaderboard).

We find that all evaluation methods track the official leaderboard. Self-rating-based EXAM++ follows the shape the best. However, the higher grade cutoff of $\tau = 5$ leads to much large error bars in contrast to $\tau = 4$. We find that the question answering-based method Manual-EXAM-Squad2 is too unspecific, assigning the same high score to two-thirds of all systems.

We find that coverage-based evaluation with Manual EXAM++ and original EXAM promote some of the low-ranking systems. With our experiment it is impossible to say whether this is due to a bias in the official leaderboard (which does not acknowledge coverage) or an issue with the coverage-based evaluation metric. However, the fact that two independent coverage-based implementations agree on assigning `ICT-DRMMTKS` a higher score, suggests that this system might indeed provide good coverage (albeit at lower precision).

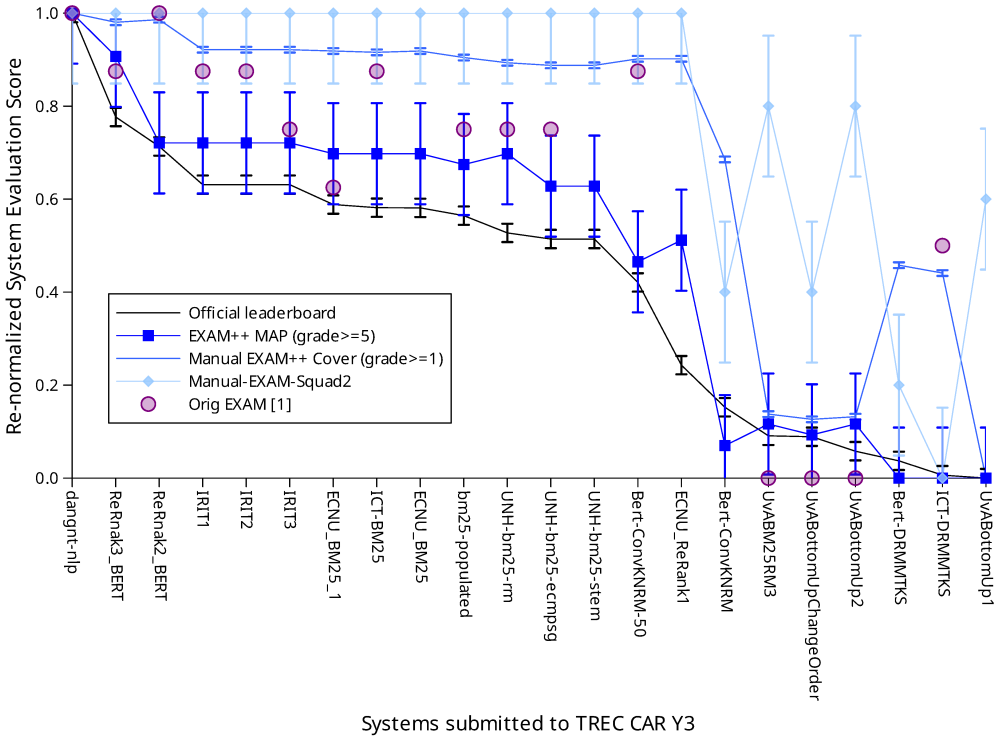

Figure 2: The goal is for leaderboards under different evaluation methods to track the official TREC CAR Y3 leaderboard. Selection of leaderboards of systems submitted to TREC CAR Y3 under three of the best correlating EXAM++ measures according to Table 3, with official leaderboard and Sander's original EXAM method (as available in Table 4 of their paper [1]). Systems are ordered by the official leaderboard (MAP). To make the general behavior more visible, we use a min/max re-normalization of all evaluation scores. Standard error bars adjusted accordingly.
We confirm that all methods roughly follow the official leaderboard, some submitted systems are very similar leading to similar evaluation scores under both the official leaderboard and our evaluation methods. Some leaderboards (e.g., Manual-EXAM-Squad2 Rprec) are not able to detect differences in performance.

5.4. Impact of Grade Cutoffs

While for generated banks of open-ended questions, a higher grade cutoff of $\tau = 5$ obtains stronger results, we observe the opposite for manual question banks taken from the TQA dataset, where a grade cutoff $\tau = 1$ produces best results.

In general, we remark that the appropriate self-rating levels depend on the difficulty of the question bank. Sander et al remark that questions of the TQA collection are often phrased in an obtuse way, as they are designed to encourage (human) students to closely read the text. As a result, too few passages obtain a high grade for most questions, which then results in evaluation scores that don't distinguish between systems.

For the open-ended questions from our generated test bank, it is generally easier to obtain a high self-rating grade—especially since multiple answers can be considered reasonably relevant.

Nevertheless, while for EXAM++ the grading cutoff of $\tau = 5$ obtains slightly better correlations than a cutoff of $\tau = 4$, the large error bars for cutoff $\tau = 5$ (cf. Figure 2) suggest that a lower cutoff might yield a more useful evaluation measure.

5.5. Self-rating vs. Answer Verification

We analyze the set of evaluation approaches that use the manual benchmark, i.e., Manual EXAM++ and methods under QA + Answer Verification. The best correlation is achieved with self-rating methods on EXAM-Cover, obtaining a 0.959 Spearman's rank correlation coefficient. However, when it is desired to integrate the evaluation into `trec_eval`, we find that answer verification approaches are strong contenders. Especially fine-tuning the `FLAN-T5-large` model on SQuAD2, obtains slightly better results than other methods.

Given that many correct answers are missed due to a different phrasing, we further explore LLM-based answer verification. However, this adaptation has strong negative effects on leader board correlation, in several cases obtaining a rank correlation of less than 0.5. We suspect that this assigns a relevant grade to too many non-relevant passages, resulting in a degradation of the leaderboard.

5.6. A Worked Example

We illustrate our EXAM++ method on an example from the TREC CAR Y3 dataset for query `tqa2:L_0384`. The passage presented below was retrieved at rank 1 by the `dangnt-nlp` system and was assessed by TREC judges as 'MUST be mentioned'.

> **Query title**: The Integumentary System
> **Query subtopic**: Structure of the Skin
> **Passage**:
> **ID:** `b95bf325b7fdacac183b1daf7c118be407f52a3a`
>
> The skin is the largest organ in the human body. Skin is made up of three layers, the epidermis, dermis and the fat layer, also called the hypodermis. The epidermis is the outer layer of skin that keeps vital fluids in and harmful bacteria out of the body. The dermis is the inner layer of skin that contains blood vessels, nerves, hair follicles, oil, and sweat glands. Severe damage to large areas of skin exposes the human organism to dehydration and infections that can result in death.
> **TREC judgment**: 3 (MUST be mentioned)

The TQA question `NDQ_007535` "Outer layer of the skin?" was correctly answered as "epidermis" by this passage (highlighted in text). Under the self-rating prompt, FLAN-T5 indicates that this question can be answered in a mostly relevant way but may have minor gaps (self-rated answerability grade of 4).

A generated exam question, "What are the main components of the epidermis and how do they contribute to the structure of the skin?", was also graded with a self-rating of 4. The corresponding extracted answer is "keeps vital fluids in and harmful bacteria out of the body" (highlighted in text).

Other generated questions for this query are:

1. What are the different layers of the skin and their respective functions?
2. How does the structure of the skin contribute to its various functions?
3. What is the role of dermal papillae in the structure of the skin?
4. How does the structure of the hypodermis differ from the other layers of the skin?
5. What structural changes occur in the skin due to aging?
6. How does the skin's structure contribute to its role in temperature regulation?
7. What role does the extracellular matrix play in the structure and function of the skin?
8. How does the structure of the skin influence its ability to prevent water loss and maintain hydration?
9. What structural adaptations exist in the skin of different animals and how do they serve their specific needs?

6. Conclusion

With EXAM++ we are proposing an alternative evaluation approach that does not merely outsource passage-level relevance determination to LLMs (or human judges). Instead, an exam question bank is created as part of topic development, envisioning that each question addresses an essential piece of information content for the query. As a result, whenever such questions are answerable with responses from a retrieval/generation system, we conclude that the system provides relevant information.

Using the TREC Complex Answer data set, we demonstrate that (1) our proposed approach can reproduce official TREC leaderboards nearly perfectly; and (2) we outperform several strong LLM-based relevance label predictors [5, 6, 8] that were developed in the context of other retrieval benchmarks. In contrast, EXAM++ offers a clear path towards integrating a human-in-the-loop, by supporting the refinement of the exam question banks, as a means for humans to define relevance.

We believe that more research will improve the question bank generation and LLM-based grading. Future work should study effects on the quality, cost, and satisfaction of human judges working with the EXAM++ approach in our Autograde software [3].

We hope that by integrating EXAM++ evaluation metric with `trec_eval`, we offer a system that can be easily adopted by future IR evaluation tracks, offering organizers an avenue to reduce assessment costs, obtain reusable test collections for generative information systems.

References

[1] D. P. Sander, L. Dietz, Exam: How to evaluate retrieve-and-generate systems for users who do not (yet) know what they want., in: DESIRES, 2021, pp. 136–146.

[2] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al., The flan collection: Designing data and methods for effective instruction tuning, arXiv preprint arXiv:2301.13688 (2023).

[3] L. Dietz, A workbench for autograding retrieve/generate systems, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24) - Resource and Reproducibility Papers, 2024. doi:https://doi.org/10.1145/3626772.3657871.

[4] N. Farzi, L. Dietz, Pencils down! Automatic rubric-based evaluation of retrieve/generate systems, in: Proceedings of the International Conference on the Theory of Information Retrieval, 2024.

[5] W. Sun, L. Yan, X. Ma, P. Ren, D. Yin, Z. Ren, Is chatgpt good at search? investigating large language models as re-ranking agent, arXiv e-prints (2023) arXiv–2304.

[6] G. Faggioli, L. Dietz, C. L. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, 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, pp. 39–50.

[7] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al., Holistic evaluation of language models, arXiv preprint arXiv:2211.09110 (2022).

[8] P. Thomas, S. Spielman, N. Craswell, B. Mitra, Large language models can accurately predict searcher preferences, 2023. arXiv:2309.10621.

[9] S. MacAvaney, L. Soldaini, One-shot labeling for automatic relevance estimation, arXiv preprint arXiv:2302.11266 (2023).

[10] G. Faggioli, L. Dietz, C. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, et al., Who determines what is relevant? humans or ai? why not both? a spectrum of human-ai collaboration in assessing relevance., Communications of the ACM (2024).

[11] Y. Liu, N. S. Moosavi, C. Lin, Llms as narcissistic evaluators: When ego inflates evaluation scores, arXiv preprint arXiv:2311.09766 (2023).

[12] P. Wang, L. Li, L. Chen, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, Z. Sui, Large language models are not fair evaluators, arXiv preprint arXiv:2305.17926 (2023).

[13] R. Fok, D. S. Weld, In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making, arXiv preprint arXiv:2305.07722 (2023).

[14] J. Clarke, M. Lapata, Discourse constraints for document compression, Computational Linguistics 36 (2010).

[15] M. Eyal, T. Baumel, M. Elhadad, Question answering as an automatic evaluation metric for news article summarization, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3938–3948. URL: https://www.aclweb.org/anthology/N19-1395. doi:10.18653/v1/N19-1395.

[16] L. Huang, L. Wu, L. Wang, Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020). URL: http://dx.doi.org/10.18653/v1/2020.acl-main.457. doi:10.18653/v1/2020.acl-main.457.

[17] D. Deutsch, T. Bedrax-Weiss, D. Roth, Towards question-answering as an automatic metric for evaluating the content quality of a summary, arXiv preprint arXiv:2010.00490

(2020).
[18] A. Wang, K. Cho, M. Lewis, Asking and answering questions to evaluate the factual consistency of summaries, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5008–5020. URL: https://www.aclweb.org/anthology/2020.acl-main.450. doi:10.18653/v1/2020.acl-main.450.
[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).
[20] A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, H. Hajishirzi, Are you smarter than a sixth grader? Textbook question answering for multimodal machine comprehension, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 5376–5384.
[21] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100, 000+ questions for machine comprehension of text, CoRR abs/1606.05250 (2016). URL: http://arxiv.org/abs/1606.05250. arXiv:1606.05250.
[22] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, I. MacKinnon, Novelty and diversity in information retrieval evaluation, in: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008, pp. 659–666.
[23] T. Sakai, M. P. Kato, Y.-I. Song, Overview of ntcir-9, in: Proceedings of the 9th NTCIR Workshop Meeting, 2011, 2011, pp. 1–7.
[24] L. Dietz, J. Foley, TREC CAR Y3: Complex Answer Retrieval overview, in: Proceedings of Text REtrieval Conference (TREC), 2019.

A. Appendix: Relevance Label Predictor Prompts

Thomas [8]: *As full prompt exceeds the token limitation, we use the following abridged prompt used in citing work:*

> Instruction: 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
> Question: {query_title}
> Passage: {context}
> Answer:

FaggioliB [6]: *Prompt designed for TREC DL:*

> Instruction: Indicate if the passage is relevant for the question. Respond with
> 'Yes' or 'No'.
> Question: {query_title}

Passage: {context}
Answer:

HELM [7]: *Prompt designed for evaluating LLMs on information retrieval:*
Instruction: Does the passage answer the query?
Respond with 'Yes' or 'No'.
Question: {query_title}
Passage: {context}
Answer:

Sun [5]: *Prompt designed for question-style queries:*
Instruction: Given a passage and a query, predict whether the passage includes an answer to the query by producing either "Yes" or "No".
Question: {query_title}
Passage: {context}
Answer:

FaggioliB_few [6]: *Prompt* `FaggioliB` *with additional few shot examples from the TREC DL collection:*
Instruction: Indicate if the passage is relevant for the question. Respond with 'Yes' or 'No'.

Passage: Its 25 drops per ml, you guys are all wrong. If it is water, the standard was changed 15 - 20 years ago to make 20 drops = 1mL. The viscosity of most things is temperature dependent, so this would be at room temperature. Hope this helps.
Question: how many eye drops per ml
Answer: Yes

Passage: RE: How many eyedrops are there in a 10 ml bottle of Cosopt? My Kaiser pharmacy insists that 2 bottles should last me 100 days but I run out way before that time when I am using 4 drops per day. In the past other pharmacies have given me 3 10-ml bottles for 100 days. E: How many eyedrops are there in a 10 ml bottle of Cosopt? My Kaiser pharmacy insists that 2 bottles should last me 100 days but I run out way before that time when I am using 4 drops per day.
Question: how many eye drops per ml
Answer: No

Passage: You can transfer money to your checking account from other Wells Fargo. accounts through Wells Fargo Mobile Banking with the mobile app, online, at any. Wells Fargo ATM, or at a Wells Fargo branch. 1 Money in — deposits.
Question: can you open a wells fargo account online
Answer: No

> Passage: You can open a Wells Fargo banking account from your home or even online. It is really easy to do, provided you have all of the appropriate documentation. Wells Fargo has so many bank account options that you will be sure to find one that works for you. They offer free checking accounts with free online banking.
> Question: can you open a wells fargo account online
> Answer: Yes
>
> Question: {query_title}
> Passage: {context}
> Answer:

Sun_few [5]: *Prompt Sun with the same few shot examples as `FaggioliB_few`.*