BIOBENCH: A Benchmark to Evaluate RAG-LLM Self-Explanations and Cited Passages in Biomedical Question Answering

Anonymous ACL submission

Abstract

Retrieval Augmented Generation (RAG) aims to improve the performance of Large Language Models (LLMs) by providing them with retrieved passages from trusted document collections. In the medical domain, benchmarks for RAG-enhanced LLMs (RAG-LLMs) were only recently published. However, they mainly consider *factuality*, i.e., the correctness of the final responses. We introduce BIOBENCH, an improved biomedical RAG-LLM question answering (QA) benchmark, which also evaluates the *plausibility* (convincing to experts) and *faithfulness* (truly reflecting the model's reasoning) of both *self-explanations* of the answers and *passages cited* by the RAG-LLMs as supporting their answers. Moreover, BIOBENCH assesses the *pertinence* of retrieved gold passages (the influence they had on the LLM) when generating answers or self-explanations. We also introduce a suite of evaluation measures, BIOBENCH-EVAL, including existing measures tailored for RAG-LLMs and new ones, for all the aforementioned dimensions. BIOBENCH includes data from four publicly available English biomedical QA datasets. To establish baselines and demonstrate the usage of BIOBENCH, we conduct experiments with two dense and one sparse retriever, with four open-source LLMs in zero- and few-shot settings.

1 Introduction

Retrieval Augmented Generation (RAG) (Lewis et al., 2020b; Gao et al., 2024; Xiong et al., 2024) is a popular extension of Large Language Models (LLMs) that grounds them on retrieved passages from trusted document collections to mitigate hallucinations (Shuster et al., 2021). RAG-enhanced LLMs (RAG-LLMs) can use up-to-date information, avoiding the cost of re-training the LLMs. There are still, however, problems with RAG, such as using information from an imperfect retriever, or failing to consider all the information, demonstrating a primacy effect (Wang et al., 2023b) and ignoring context in the middle of the prompt (Liu et al., 2023). Comprehensive and challenging benchmarks are, thus, required to monitor progress in RAG-LLMs.

In the biomedical domain, the first extensive benchmark for RAG-LLMs, MIRAGE (Xiong et al., 2024), was recently published. However, it only measures *factuality*, in terms of accuracy of the generated answers to multiple-choice questions. It requires *self-explanations*, meaning generated chain-of-thought (COT) (Kojima et al., 2024; Brown et al., 2020) explanations of the answers (Madsen et al., 2024; Randl et al., 2024; Huang et al., 2023), but it evaluates neither their *plausibility* (how convincing they are to experts) nor their *faithfulness* (if they truly reflect the model's reasoning) (Jacovi and Goldberg, 2020). Furthermore, it does not require RAG-LLMs to *cite retrieved passages* their answers are based on. Newer benchmarks are MEDRGB (Xiong et al., 2024; Ngo et al., 2024) and MEDEC (Abacha et al., 2025). MEDRGB only partially considers the faithfulness of cited passages (it measures only their sufficiency, §D.2), but it does not assess their plausibility, and does not consider self-explanations. MEDEC evaluates the plausibility of self-explanations, but not of cited passages, and does not evaluate faithfulness at all.

We introduce an improved biomedical question answering (QA) RAG-LLM benchmark, BIOBENCH, that apart from factuality, evaluates the plausibility and faithfulness of both self-explanations and cited passages. Moreover, it assesses the *pertinence* of gold retrieved passages, capturing whether the retrieved passages that mostly influenced the answers or self-explanations, respectively, of a RAG-LLM were known gold ones. BIOBENCH also evaluates retrievers on their own, decoupled from LLMs, to assess their contribution in RAG-LLMs.

While BIOBENCH targets the critical, knowledge-intensive biomedical domain, our introduced metrics are highly transferable. They can be readily

applied to evaluate RAG-LLMs in any high-stakes domain, such as biomedicine (Siriwardhana et al., 2023; Yu et al., 2024; Wu et al., 2024), law (Magesh et al., 2024; Pipitone and Alami, 2024) or finance (Li et al., 2024b; Islam et al., 2023; Zhao et al., 2025) to name a few, where RAG is particularly important since knowledge is continuously updated, trustworthiness and explainability are paramount.

To evaluate the plausibility of self-explanations and cited passages, we compare them to expert explanations and gold passages, respectively, using BERTSCORE (Zhang et al., 2020), BARTSCORE (Yuan et al., 2021), BLEURT (Sellam et al., 2020a), and (only for cited passages) F_1, Mean Average Precision (MAP) and Natural Language Inference (NLI) scores between the cited passages and the corresponding self-explanations, assessed by an LLM-judge. For faithfulness of self-explanations, we modify CC-SHAP (Parcalabescu and Frank, 2024), a consistency measure based on Shapley values (Lundberg and Lee, 2017), originally designed for LLMs without RAG, to operate on RAG-LLMs. For faithfulness of cited passages, we introduce a new measure, *faithful grounding*, which also uses Shapley values. We also explore *comprehensiveness*, *sufficiency*, and *sparsity*, commonly used to evaluate faithfulness in other tasks (DeYoung et al., 2020; Yu et al., 2019; Chalkidis et al., 2021b; Brinner and Zarrieß, 2023; Carton et al., 2020; Namazov et al., 2026; Chrysostomou and Aletras, 2022), but we explain and experimentally demonstrate (§D.2) that they are unsuitable for RAG-LLMs. For pertinence of gold retrieved passages, we propose another new measure, *contextual awareness*, also based on Shapley values. Lastly, to individually evaluate retrievers, we use Precision, Recall, MAP, normalized Discounted Cumulated Gain (DCG), and F_1, when gold relevant passages are available. Implementations of all evaluation measures used in the benchmark are included in BIOBENCH-EVAL, an evaluation suite that accompanies BIOBENCH.

We conduct experiments with four open-source chat LLMs: LLAMA 3.1 70B (Dubey et al., 2024); LLAMA 3.3 70B (Dubey et al., 2024); MED42-V2-70B (Wu et al., 2023), a medical LLM based on LLAMA 3 70B; and MIXTRAL-8×7B (Jiang et al., 2024a) to obtain reproducible baseline results and demonstrate the usage of BIOBENCH. We use these LLMs with or without RAG, in both cases using either zero-shot (Kojima et al., 2024) or few-shot (Brown et al., 2020) COT prompting. When RAG is enabled, we experiment with two dense retrievers, namely MEDCPT (Jin et al., 2023) and CONTRIEVER (Izacard et al., 2021), as well as BM25 (Robertson and Zaragoza, 2009).

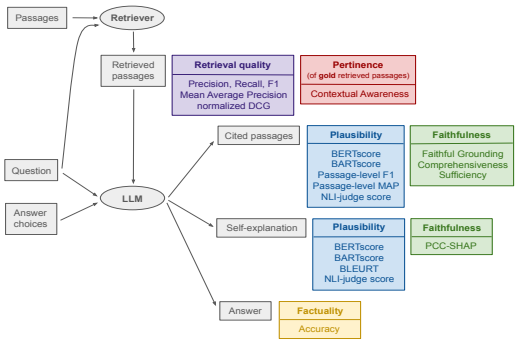

Figure 1: Illustration of BIOBENCH and its metrics.

Our contributions are: **(i)** BIOBENCH, the first benchmark for RAG-LLMs in biomedical QA that, apart from factuality, evaluates the plausibility and faithfulness of self-explanations and cited passages, as well the pertinence (new evaluation dimension) of gold retrieved passages; **(ii)** BIOBENCH-EVAL, a suite of metrics including modified versions of CC-SHAP and comprehensiveness, sufficiency, sparsity to evaluate the faithfulness of self-explanations and cited passages, respectively, but also two new measures, *faithful grounding* and *contextual awareness*, to evaluate the faithfulness of cited passages and pertinence of gold retrieved passages respectively; **(iii)** baseline results from four open-source LLMs and three open-source and widely used retrievers, in zero- and few-shot settings, as well as evaluation results for the retrievers themselves and LLMs without RAG; **(iv)** we make BIOBENCH and BIOBENCH-EVAL public, along with a leaderboard and code facilitating broader use of the benchmark, through testing of additional LLMs and retrievers.[1]

2 The new BIOBENCH benchmark

Each sample of BIOBENCH includes a question, a detailed long answer (treated as gold explanation), possibly a short multiple-choice answer, and gold relevant passages. Unlike others (Liévin et al., 2023; Lála et al., 2023; Narayanan et al., 2024; Skarlinski et al., 2024), who include the available multiple-choice answers in the queries to the retrievers, we follow MIRAGE and use only the questions as queries, making the task more challenging.

2.1 Datasets

BIOBENCH comprises selected data (appropriate for our goals) from four existing biomedical QA

[1] Our code and data will be released at the camera-ready.

datasets, namely PUBMEDQA (Jin et al., 2019), MEDMCQA (Pal et al., 2022), HEALTHFC (Vladika et al., 2024), as well as the yes/no and summary questions from BIOASQ-QA (Krithara et al., 2023). We use PUBMED[2] as the collection to retrieve passages from, which is the largest freely accessible collection of publications in biomedical and life sciences, with over 38 million scientific articles.

The four datasets of BIOBENCH were selected because **(a)** they are in English and publicly available, **(b)** they require multiple-choice answers (except for the summary questions of BIOASQ-QA §B.3), which facilitates factuality evaluation, **(c)** they provide gold textual explanations, similar to the required self-explanations of RAG-LLMs, **(d)** they contain biomedical questions requiring expert knowledge retrievable from documents, and **(e)** they cover different types of biomedical QA, including information extraction from PUBMED (PUBMEDQA, BIOASQ-QA), medical entrance examinations (MEDMCQA), and health-related fact checking (HEALTHFC). The latter type is not covered by other RAG-LLM benchmarks. Comprehensive statistics for the datasets used in BIOBENCH and filtering details are deferred to Appendix (§B), where indicative samples are also provided.

2.2 Retrievers

We use the MEDRAG toolkit (Xiong et al., 2024), which facilitates prompting RAG-LLMs with retrieved passages. It originally supported five corpora, four retrievers, and the LLMs used in MIRAGE, in zero-shot settings only. We extend it to support the additional LLMs (§1) and datasets (§2.1) of BIOBENCH, as well as few-shot settings.

We perform experiments with two dense and one sparse retriever supported by MEDRAG: MEDCPT (Jin et al., 2023), a biomedical neural retriever fine-tuned on PUBMED user logs; CONTRIEVER (Izacard et al., 2021), a generic neural retriever; as well as BM25 (Robertson and Zaragoza, 2009), a sparse retriever with strong performance in biomedical QA (Nentidis et al., 2024). We do not experiment with SPECTER (Cohan et al., 2020), also in MEDRAG, since it was consistently outperformed by the other retrievers in MIRAGE (Xiong et al., 2024). For details about exact hyper-parameters, indexing strategies, and prompt templates of our retrievers, we refer the reader to the Appendix (§C).

[2]https://pubmed.ncbi.nlm.nih.gov/

2.3 Evaluation measures

2.3.1 Plausibility of self-explanations

To evaluate the plausibility of self-explanations (how convincing they are to medical experts), we measure the similarity between self-explanations generated by the LLM, and the corresponding gold explanations of an expert; the latter are included in the datasets of BIOBENCH (§2.1). Following previous work on evaluating LLMs (Moreno and Bitterman, 2024; Dada et al., 2024; Ionescu et al., 2024) in medical and clinical text generation, we employ BERTSCORE (Zhang et al., 2020), BARTSCORE (Yuan et al., 2021), and BLEURT (Sellam et al., 2020a) to measure the similarity between each self-explanation and the corresponding gold one, and we then average the similarity scores over all test questions. We use BERTSCORE as our primary metric, as it measures directly the semantic similarity between the generated and gold self-explanations, without requiring fine-tuning. By contrast, BLEURT implementations are fine-tuned on human-annotations from other domains, which may not generalize well to the biomedical domain; and BARTSCORE evaluates the likelihood of the generated text, conditioned on the gold text, rather than directly assessing the semantic similarity of the two (see §D.3 for details). For completeness, we report scores for all three measures in §G.

2.3.2 Plausibility of cited passages

To evaluate the plausibility of cited passages, we compare the cited passages of each question to the gold passages of the question by computing F_1 and MAP at the passage level (e.g., a true positive is a passage cited by the LLM and included in the gold passages), and averaging over questions. These scores can be computed only in BIOASQ-QA and PUBMEDQA (§2.1), which include gold passages.

Even in those datasets, however, the gold passages may not be complete. There may be alternative passages, retrieved and cited by the LLM, that support the correct answer, even though they are not included in the gold passages. In such cases, the system is incorrectly penalized by the passage-level F_1 and MAP. To address this issue, we employ an LLM-judge to compute NLI scores between the cited passages of each question and the corresponding self-explanations, assessing to what extent an external observer (human or NLI system acting as a surrogate) would consider the passages as plausible support of the self-explanation. In addition, we

report BERTSCORE and BARTSCORE, as when evaluating the plausibility of self-explanations (§2.3.1), but this time comparing the cited passages to the corresponding gold ones. These measures are more sensitive to *semantic* similarity between passages.

Concretely, for every cited passage, we compute its BERTSCORE comparing to each gold passage, and keep the best score. We then average over the cited passages to obtain a *precision-oriented* BERTSCORE, which intuitively checks if the cited passages are all included in the set of gold ones. Similarly, for each gold passage, we compute its BERTSCORE comparing to each cited passage and keep the best score, averaging over gold passages. This *recall-oriented* BERTSCORE checks if the gold passages are included in the cited ones. We then report the F_1 of the precision- and recall-oriented BERTSCORE. This is again possible only when gold passages are available. The same process is followed for BARTSCORE. We do not employ BLEURT here, because it also considers fluency (§D.3), but retrieved passages are always fluent.

2.3.3 Faithfulness of self-explanations

Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017) systematically masks parts (e.g., features or tokens) of the input to a model, to determine the *contribution* that each part of the input had to the model's output (e.g., classification decision). Consistency Scores based on SHAP (CC-SHAP) (Parcalabescu and Frank, 2024) is an offsping of SHAP that provides a single contribution score per input token, when the output is an LLM-generated text, comprising an answer and a self-explanation. If the self-explanation is faithful, the token contributions should be similar when generating the self-explanation and the answer.

CC-SHAP was designed for LLMs without RAG. In RAG-LLMs, the input to the LLM is much longer, since it includes retrieved passages, and computing contribution scores for all the input tokens becomes infeasible. We introduce a new variant for RAG-LLMs, dubbed passage-level CC-SHAP (PCC-SHAP), which operates on retrieved passages rather than input tokens. PCC-SHAP compares to what extent a RAG-LLM considers each retrieved passage (the passage contributions) when producing the self-explanation vs. the answer itself. If the self-explanation is faithful, the contribution of each passage should be similar in both cases (see §D.1 for details). Computing a contribution score per retrieved passage rather than for every token of the retrieved passages, greatly speeds up calculations (see §B.5). It also provides contribution scores that correspond more directly to self-contained pieces of knowledge, i.e., entire passages, as opposed to individual tokens that usually express more fragmented information. Hence, passage-level contribution scores provide a higher-level, more coherent window to the model's reasoning.

2.3.4 Faithfulness of cited passages

To evaluate the faithfulness of cited passages, we introduce *faithful grounding*. For each question, we rank all the retrieved passages by decreasing c_{self} or c_{ans}; these are the contribution scores of each passage (computed during PCC-SHAP, §2.3.3, §D.1) that show how much the LLM based its self-explanation or answer, respectively, on the particular passage. We then use MAP, as in information retrieval, to evaluate the c_{self} or c_{ans} rankings (averaging over questions), treating cited passages as relevant, and the other retrieved passages as irrelevant. Faithful grounding (denoted FG_{self}, FG_{ans}) is, thus, high if the cited passages were indeed those that influenced mostly the LLM when generating the self-explanation or answer, respectively.

Apart from faithful grounding, we also considered *sufficiency* and *comprehensiveness* (Jacovi and Goldberg, 2020; DeYoung et al., 2020) to evaluate the faithfulness of cited passages. Sufficiency and comprehensiveness have been used to assess the faithfulness of *extractive rationales*, which are parts of the input (e.g., words or sentences) that mostly influenced the model's output (§A.2). We modified (§D.2) the definitions of sufficiency and comprehensiveness for RAG-LLMs. However, these metrics are problematic when systems are free to extract rationales of any size (or *sparsity*); a system that always selects the entire input as rationale (in our case, a RAG-LLM that cites all the retrieved passages) will have perfect sufficiency and comprehensiveness. We show experimentally (§D.2) that different RAG-LLMs cite very different numbers of retrieved passages, hence these metrics cannot reliably evaluate the faithfulness of cited passages, unless one forces the lists of cited passages to have the same lengths across LLMs, which is unrealistic.

2.3.5 Pertinence of gold retrieved passages

Recall that faithful grounding (§2.3.4) uses MAP to compare the passages cited by the LLM to the c_{self} or c_{ans}-based rankings of retrieved passages. Intuitively, it checks if the cited passages were indeed

those that mostly affected the self-explanations or answers, respectively. Similarly, we use MAP (again averaging over questions) to compare the set of *gold* relevant passages to the c_{self} or c_{ans}-based rankings of retrieved passages; we remove from the set of gold passages those that were not retrieved. The resulting measure, *contextual awareness* (denoted CA_{self}, CA_{ans}), assesses how much the self-explanations or answers, respectively, of the LLM are based on gold relevant passages, among those included in its context. This measure can again be computed only when gold relevant passages are available (PUBMEDQA, BIOASQ-QA). A caveat is that the LLM may base its self-explanation or answer to passages in its context that are semantically similar to gold relevant passages, but not the gold passages themselves, which may not have even been retrieved. Nevertheless, the measure provides an indication of whether or not LLMs can recognize known relevant gold passages in their context.

2.3.6 Independent evaluation of the retrievers

To evaluate each retriever independently (without considering its interaction with the LLM), we report MAP (overall and MAP@k), Precision (overall and Precision@k), Recall, F_1, and normalized DCG (Manning et al., 2008), again when gold relevant passages are available (PUBMEDQA, BIOASQ-QA).

3 Experimental results and discussion

3.1 Retrieval evaluation

Tables 1 and 2 show that the generic sparse BM25 retriever (Robertson and Zaragoza, 2009) is consistently the best performer in the datasets that include gold relevant passages for retrieval evaluation purposes (PUBMEDQA, BIOASQ-QA, see §2.1, §2.3.6), even with default hyper-parameter values.[3] The scores of BM25 may further improve by optimizing the hyper-parameter values for each dataset of BIOBENCH, as shown by previous work (Chalkidis et al., 2021a). In all three datasets, BM25 is clearly the best across all evaluation measures, outperforming both the domain-specific MEDCPT (Jin et al., 2023) and the generic CONTRIEVER (Izacard et al., 2021), which are both dense retrievers. Interestingly, MEDCPT performs better than CONTRIEVER on PUBMEDQA, whose questions are obtained from abstract titles (§B.1). MEDCPT is not, however, always better than the generic CONTRIEVER on the two BIOASQ datasets (e.g., see the MAP of Table 1), which contain real-life questions written by biomedical experts (§B.3). All results reported hereafter use BM25, the best-performing retriever, unless otherwise noted. Even BM25, however, performs poorly, especially in terms of Precision.

Dataset	Retriever	Precision (↑)	Recall (↑)	F_1 (↑)	MAP (↑)
PUBMEDQA	BM25	**0.031**	**0.984**	**0.060**	**0.930**
	MEDCPT	0.030	0.974	0.059	0.844
	CONTRIEVER	0.028	0.898	0.054	0.701
BIOASQ-QA (yes/no)	BM25	**0.132**	**0.446**	**0.204**	**0.592**
	MEDCPT	0.086	0.292	0.133	0.374
	CONTRIEVER	0.082	0.277	0.126	0.389
BIOASQ-QA (summary)	BM25	**0.138**	**0.427**	**0.209**	**0.459**
	MEDCPT	0.082	0.253	0.124	0.174
	CONTRIEVER	0.119	0.367	0.179	0.333

Table 1: Independent evaluation (§2.3.6) of the three baseline retrievers of BIOBENCH (§2.2), for the datasets that include gold relevant passages. See also Table 2.

The findings of Tables 1–2 contradict those of MIRAGE (Xiong et al., 2024), where MEDCPT performed better than the other two retrievers, when using the PUBMED repository (as we also do). However, MIRAGE ranks retrievers by downstream QA accuracy, whereas here we evaluate them independently of their interaction with the LLM. Assessing the retrievers solely based on downstream QA accuracy of RAG-LLMs can be misleading, as overall performance is influenced by other factors; the LLM's reasoning ability, prompt design, and generation strategy may amplify or mask retrieval quality.

3.2 Factuality evaluation

Table 3 shows that RAG consistently and substantially improves factuality compared to LLMs without RAG in PUBMEDQA and BIOASQ-QA.[4] These two datasets were specifically created (§B.1, §B.3) to check if systems can identify and report relevant information from PUBMED; each question is guaranteed to have relevant PUBMEDQA passages (the gold ones). By contrast, RAG degrades performance in MEDMCQA (medical entrance examinations, §B.2) and HEALTHFC (health-related fact checking, §B.4), which were not created specifically for PUBMED. These observations hold for all numbers k of retrieved passages (Table 3), for few-shot RAG, and when using dense retrievers (§G).

It is unlikely that a vast biomedical repository like PUBMED does not contain useful information

[3] We use Lucene (lucene.apache.org), with $k_1 = 1.5$, $b = 0.75$ as BM25 hyper-parameter values (Manning et al., 2008).

[4] The BIOASQ-QA scores are also on par with those reported in the 11th BIOASQ (Nentidis et al., 2023), where the BIOASQ-QA questions originate from. Table 3 does not include BIOASQ-QA summary questions, because they do not have multiple-choice answers and, hence, accuracy cannot be computed.

Dataset	Retriever	Precision (↑)				MAP (↑)				Normalized DCG (↑)			
		P@5	P@10	P@15	P@20	MAP@5	MAP@10	MAP@15	MAP@20	DCG@5	DCG@10	DCG@15	DCG@20
PUBMEDQA	BM25	**0.193**	**0.098**	**0.065**	**0.049**	**0.928**	**0.930**	**0.930**	**0.930**	**0.973**	**0.941**	**0.941**	**0.942**
	MEDCPT	0.186	0.095	0.064	0.048	0.841	0.843	0.844	0.844	0.863	0.869	0.873	0.875
	CONTRIEVER	0.155	0.081	0.057	0.043	0.698	0.702	0.706	0.706	0.717	0.729	0.740	0.743
BIOASQ-QA (yes/no)	BM25	**0.335**	**0.249**	**0.196**	**0.165**	**0.518**	**0.555**	**0.568**	**0.580**	**0.649**	**0.625**	**0.616**	**0.620**
	MEDCPT	0.202	0.150	0.120	0.106	0.317	0.350	0.359	0.367	0.397	0.404	0.400	0.411
	CONTRIEVER	0.207	0.160	0.127	0.103	0.356	0.373	0.380	0.385	0.405	0.396	0.399	0.405
BIOASQ-QA (summary)	BM25	**0.299**	**0.238**	**0.200**	**0.176**	**0.347**	**0.416**	**0.437**	**0.448**	**0.502**	**0.497**	**0.490**	**0.492**
	MEDCPT	0.154	0.123	0.104	0.100	0.121	0.149	0.160	0.167	0.225	0.228	0.230	0.239
	CONTRIEVER	0.241	0.205	0.165	0.146	0.267	0.304	0.319	0.325	0.399	0.392	0.396	0.399

Table 2: Independent evaluation (§2.3.6) of the three baseline retrievers of BIOBENCH (§2.2) in terms of Precision@k and MAP@k, for $k \in \{5, 10, 15, 20\}$. Overall Precision, Recall, F_1, MAP reported in Table 1.

Dataset	LLM	ZERO-SHOT no RAG	ZERO-SHOT with RAG $k=5$	$k=10$	$k=15$	$k=20$
PUBMEDQA	MIXTRAL	31.80	65.60	66.40	56.00	63.20
	LLAMA 3.1	59.80	78.80	<u>**80.00**</u>	**78.60**	**78.00**
	LLAMA 3.3	**60.80**	**79.00**	78.00	**78.60**	77.60
	MED42	19.80	42.40	40.40	42.20	44.60
BIOASQ-QA (yes/no)	MIXTRAL	81.40	87.21	89.53	82.56	84.88
	LLAMA 3.1	81.40	**91.86**	**91.86**	88.37	<u>**93.02**</u>
	LLAMA 3.3	**82.56**	89.53	**91.86**	**90.70**	89.53
	MED42	26.74	61.63	65.12	56.98	55.81
MEDMCQA	MIXTRAL	57.38	52.62	52.14	54.52	51.67
	LLAMA 3.1	77.38	70.00	71.43	74.05	72.14
	LLAMA 3.3	<u>**78.33**</u>	**73.81**	**73.57**	**75.71**	**73.10**
	MED42	35.00	13.10	43.57	46.19	38.33
HEALTHFC	MIXTRAL	<u>**56.93**</u>	**45.33**	**43.33**	**40.13**	**42.93**
	LLAMA 3.1	41.33	40.80	42.40	39.20	41.07
	LLAMA 3.3	41.47	42.53	42.40	39.20	40.13
	MED42	10.40	12.13	17.07	16.27	19.20

Table 3: Factuality (accuracy, %) zero-shot results, with/without RAG, for different numbers of retrieved passages (k). Underlined scores are the best per dataset. RAG results obtained here with the BM25 retriever.

for entrance exams or health-related fact checking. A more plausible explanation for the degradation (with RAG) in MEDMCQA and HEALTHFC is that their questions are more difficult to match to relevant PUBMED passages, because they were not written having particular PUBMED passages in mind, unlike PUBMEDQA and BIOASQ-QA. This causes the retrievers to pass on to the LLMs irrelevant passages, degrading their performance. Hence, we view MEDMCQA and HEALTHFC as more challenging: they require both better retrievers and LLMs capable of coping with irrelevant retrieved passages.

Notice also that MED42 is outperformed in factuality by all other LLMs (Table 3), highlighting that domain-specific models may lag behind improvements in generic ones (Dorfner et al., 2025).

In the following tables of the main text, we report results only for $k = 5$, because similar observations hold for the other values of k (see §G).

3.3 Plausibility evaluation

Table 4 shows that RAG overall improves the plausibility of self-explanations in PUBMEDQA and the two BIOASQ-QA datasets, especially coupled with few-shot examples; the latter presumably guide the LLMs to generate outputs that better resemble the style of expert explanations (§2.1, §B). In MEDMCQA and HEALTHFC, RAG harms the plausibility of self-explanations, but this is consistent with the negative effect RAG has in the factuality scores of those datasets (Table 3), since incorrect responses are also more unlikely to be judged as plausible. There is no clear winner in Table 4, but the two (generic) LLAMA models achieve the best results in the datasets where RAG is beneficial, whereas MIXTRAL prevails in MEDMCQA and HEALTHFC.

Dataset	LLM	ZERO-SHOT no RAG	ZERO-SHOT with RAG	FEW-SHOT with RAG
PUBMEDQA	MIXTRAL	**55.05**	57.76	56.77
	LLAMA 3.1	54.65	**60.95**	<u>**63.00**</u>
	LLAMA 3.3	54.09	58.67	61.44
	MED42	53.34	54.08	55.35
BIOASQ-QA (yes/no)	MIXTRAL	**58.48**	**56.79**	55.91
	LLAMA 3.1	56.01	57.32	59.10
	LLAMA 3.3	54.43	56.88	<u>**59.50**</u>
	MED42	54.76	52.20	53.55
BIOASQ-QA (summary)	MIXTRAL	**58.41**	58.71	44.03
	LLAMA 3.1	55.21	**65.57**	65.72
	LLAMA 3.3	57.49	63.77	<u>**65.94**</u>
	MED42	56.99	55.32	55.05
MEDMCQA	MIXTRAL	<u>**54.15**</u>	51.74	49.17
	LLAMA 3.1	53.20	**52.81**	**54.14**
	LLAMA 3.3	53.64	52.53	53.45
	MED42	53.25	51.53	52.36
HEALTHFC	MIXTRAL	<u>**56.76**</u>	53.55	–
	LLAMA 3.1	55.00	53.68	–
	LLAMA 3.3	54.33	**53.72**	–
	MED42	52.47	49.51	–

Table 4: Plausibility of self-explanations (BERTSCORE, %), zero-shot with/without RAG, and few-shot ($n = 5$, §2.2) with RAG. RAG uses BM25, $k = 5$. One retrieved passage included in each few-shot example. Underlined scores are the best per dataset. HEALTHFC contains no training data usable as few-shot examples (§B.4).

Moving on to the plausibility of cited passages, the most interesting observation in Table 5 is that all three measures produce very similar rankings, with LLAMA 3.1 being always the best LLM, except for its MAP in the yes/no questions of BIOASQ-QA. The low F_1 scores indicate that the passages cited by the RAG-LLMs are largely different from

those cited by the experts. The low F_1-oriented BERTSCORE results (§2.3.2) show that the RAG-LLMs largely also fail to cite passages that are semantically similar to the gold ones. The high MAP scores in PUBMEDQA are due to the fact that even though the RAG-LLMs fail to cite many of the gold passages, in that particular dataset they manage to place the gold passages that they do cite highly in their citation lists. This is observed only in PUBMEDQA and can be attributed to the questions of that dataset being titles of PUBMED abstracts (§B.1), which presumably makes identifying gold passages easier than in other datasets.

Dataset	LLM	ZERO-SHOT with RAG		
		Passage-level		F_1-oriented
		F_1 (↑)	MAP (↑)	BERTSCORE (↑)
PUBMEDQA	MIXTRAL	52.37	92.98	56.78
	LLAMA 3.1	**56.08**	**93.78**	**57.84**
	LLAMA 3.3	43.40	93.07	56.68
	MED42	50.88	93.45	57.05
BIOASQ-QA (yes/no)	MIXTRAL	27.25	48.66	55.01
	LLAMA 3.1	**30.23**	52.80	**55.69**
	LLAMA 3.3	29.89	51.62	55.65
	MED42	29.38	**54.95**	55.39
BIOASQ-QA (summary)	MIXTRAL	20.38	30.60	56.42
	LLAMA 3.1	**24.93**	**35.43**	**56.71**
	LLAMA 3.3	24.66	33.50	56.50
	MED42	18.54	30.20	56.66

Table 5: Plausibility of cited passages (%), zero-shot RAG, using BM25, $k = 5$. Measures applicable only to datasets providing gold passages, with RAG enabled. Few-shot, BARTSCORE, and NLI scores (for all datasets) from the LLM-judge results are similar (§G).

3.4 Faithfulness evaluation

Recall that we use PCC-SHAP (§2.3.3) to assess the faithfulness of self-explanations, checking if the contribution of each retrieved passage is the same when generating the self-explanation and the answer. We also use faithful grounding (§2.3.4) to assess the faithfulness of cited passages, checking if the cited passages are indeed those that influenced mostly the LLM when it generated the self-explanation (FG_{self}) or answer (FG_{ans}), respectively.

Table 6 shows that LLAMA 3.3 is overall the top performer in PCC-SHAP, with the only exception of BIOASQ-QA yes/no questions. Note, however, that PCC-SHAP is a cosine distance (§D.1.1), hence its values range in $[0, 2]$ and lower values are better. The PCC-SHAP scores are, therefore, overall quite poor, signaling that the LLMs do not base their self-explanations and answers to the same retrieved passages. On the other hand, the FG_{self} scores are very high, with LLAMA 3.3 being again the top performer overall, achieving near-perfect scores. (FG_{self} and FG_{ans} are MAP-based, their values range in $[0, 1]$, higher values are better.) LLAMA 3.3 is also the top-performer in FG_{ans}, but the FG_{ans} scores are overall much worse than the FG_{self} ones.

Dataset	LLM	ZERO-SHOT with RAG		
		PCC-SHAP (↓)	FG_{self} (↑)	FG_{ans} (↑)
PUBMEDQA	LLAMA 3.1	0.867	**0.999**	0.683
	LLAMA 3.3	**0.852**	0.997	**0.804**
	MED42	0.890	0.985	0.754
BIOASQ-QA (yes/no)	LLAMA 3.1	0.820	0.981	0.793
	LLAMA 3.3	0.925	**0.992**	**0.896**
	MED42	**0.711**	0.959	0.807
BIOASQ-QA (summary)	LLAMA 3.1	–	0.962	–
	LLAMA 3.3	–	**0.979**	–
	MED42	–	0.687	–
MEDMCQA	LLAMA 3.1	**0.807**	0.874	0.684
	LLAMA 3.3	0.987	**0.952**	**0.829**
	MED42	0.823	0.878	0.668
HEALTHFC	LLAMA 3.1	**0.756**	0.965	0.786
	LLAMA 3.3	0.841	**0.994**	**0.861**
	MED42	0.809	0.889	0.709

Table 6: Faithfulness of self-explanations (PCC-SHAP) and cited passages (FG_{self}, FG_{ans}). Zero-shot RAG, using BM25, $k = 5$. Metrics applicable only with RAG enabled. Results with dense retrievers are similar (§G). The summary questions of BIOASQ do not have exact answers (§B.3); PCC-SHAP and FG_{ans} cannot be computed. No scores for MIXTRAL are provided due to problems in obtaining its token probabilities (same in Table 7).

The FG_{self} and FG_{ans} scores show that there is good alignment between the cited passages and the retrieved passages ranked by c_{self} (contributions when generating self-explanations, §2.3.4), but the alignment between cited passages and retrieved passages ranked by c_{ans} (contributions when generating answers) is much worse. In other words, the cited passages faithfully explain which retrieved passages mostly influenced the generation of self-explanations, but not the answers. These findings are consistent with the low PCC-SHAP scores, which indicate a mismatch between the c_{self} and c_{ans} rankings. We suspect that the mismatch is due to the LLMs basing their answers on the self-explanations (which are then no longer explanations of how the answers were obtained from the retrieved passages), whereas c_{ans} contributions correctly show how much each retrieved passage directly influenced the answer (see §D.1.3, §F). We leave further investigation for future work.

3.5 Pertinence evaluation

Recall that pertinence is a new evaluation dimension that intuitively checks if LLMs can recognize known relevant gold passages in their context (§2.3.5). We measure pertinence with the new con-

textual awareness measure (CA_{self}, CA_{ans}), which assesses the alignment between gold retrieved passages and retrieved passages ranked by c_{self} or c_{ans}.

Dataset	LLM	ZERO-SHOT with RAG	
		CA_{ans} (↑)	CA_{self} (↑)
PUBMEDQA	LLAMA 3.1	0.485	**0.907**
	LLAMA 3.3	0.460	0.906
	MED42	**0.582**	0.893
BIOASQ (yes/no)	LLAMA 3.1	0.366	**0.568**
	LLAMA 3.3	0.396	0.527
	MED42	**0.440**	0.545
BIOASQ (summary)	LLAMA 3.1	–	**0.427**
	LLAMA 3.3	–	0.411
	MED42	–	0.372

Table 7: Pertinence of retrieved passages. Zero-shot RAG, BM25, $k = 5$. Applicable only to datasets with gold passages, with RAG enabled. The summary questions of BIOASQ do not have exact answers (§B.3).

Table 7 shows that in PUBMEDQA, the LLMs largely base their self-explanations on gold passages (high CA_{self}), which is consistent with the high MAP scores observed when assessing the plausibility of cited passages (§3.3, Table 5). Their answers, however, are much less based on gold passages (low CA_{ans}), which is again consistent with the low PCC-SHAP scores (mismatched c_{self} and c_{ans} contributions). In the other datasets of Table 7, we observe much lower CA_{self} and CA_{ans} scores, indicating that the LLMs are much less capable of recognizing gold passages in their context. This is again due to the fact that the BIOASQ-QA questions are more difficult to match to gold passages.

4 Related work

MIRAGE (Xiong et al., 2024) includes five biomedical QA datasets, four retrievers, general and domain-specific LLMs, both open-source and proprietary. However, it only measures the factuality of multiple-choice answers. It requires self-explanations, but does not evaluate them at all, and does not require the LLMs to cite passages. Its MEDRAG toolkit facilitates prompting RAG-LLMs with retrieved passages for the datasets used in MIRAGE, but only in zero-shot settings. We expanded it to support few-shot prompting, as well as the additional datasets and LLMs of BIOBENCH.

MEDRGB (Xiong et al., 2024; Ngo et al., 2024) measures the sufficiency of cited passages (§D.2) to check if RAG-LLMs can refrain from responding when not enough information is available, but does not assess their plausibility, and does not use the other faithfulness measures we employ for cited passages (comprehensiveness, faithful grounding). Also, it does not consider self-explanations at all.

MEDEC (Abacha et al., 2025) evaluates the ability of RAG-LLMs to identify and correct medical errors, focusing on US clinical notes. It measures factuality and the plausibility of self-explanations, but does not require citing passages, nor does it evaluate faithfulness at all. Multiple (proprietary only) RAG-LLMs are tested, with results showing that they are outperformed by human experts.

Recent work has also explored the use of LLMs as automated evaluators for clinical text generation (Croxford et al., 2025). In this LLM-as-a-judge paradigm, models assess generated outputs against reference reports along dimensions such as clinical correctness, omitted findings, and hallucinated observations. Unlike similarity-based metrics (e.g., BERTSCORE), LLM-judges can reason over medical statements to identify complex logical inconsistencies and factual errors. However, due to concerns regarding reproducibility, prompt sensitivity, and high inference costs, they are typically used as complementary evaluation tools rather than primary benchmark metrics. BIOBENCH employs an LLM-judge, alongside other metrics, to measure the plausibility of cited passages (§2.3.2).

5 Conclusions and future work

We presented BIOBENCH, the first RAG-LLM biomedical QA benchmark that, apart from factuality, assesses the plausibility and faithfulness of self-explanations and cited passages. We also introduced pertinence, a new evaluation dimension that assesses the influence of gold, retrieved passages in the generation of answers and self-explanations. We modified CC-SHAP and comprehensiveness, sufficiency, to evaluate the faithfulness of self-explanations and cited passages, respectively, but we also introduced two new measures, faithful grounding and contextual awareness, for the faithfulness of cited passages and pertinence. BIOBENCH includes data from four English biomedical QA datasets. To establish baselines and demonstrate the usage of BIOBENCH, we conducted experiments with three retrievers, evaluated independently and combined with four open-source LLMs in zero- and few-shot settings. Implementations of all evaluation measures are included in BIOBENCH-EVAL, for other researchers wishing to measure their effect on BIOBENCH. Future work could investigate further the mismatch between c_{self} and c_{ans} contributions, as well as the influence of self-explanations to answers (§3.4).

6 Limitations

Following previous research that evaluated the faithfulness of self-explanations (Madsen et al., 2024; Randl et al., 2024; DeYoung et al., 2020; Parcalabescu and Frank, 2024), we do not analyze the inner workings of LLMs to understand how they reach their decisions. By contrast, we employ post-hoc explanation methods to study the self-consistency of LLMs (e.g., comparing the contributions of retrieved passages when generating answers and self-explanations) observing their inputs and outputs. Understanding how LLMs internalize information provided in their context and how they process it to answer, produce self-explanations, and cite passages is an extremely challenging research question that we do not address.

Unlike the competitive MEDEC benchmark (Abacha et al., 2025), which compares multiple (proprietary only) RAG-LLMs to human experts, a limitation of BIOBENCH is the absence of human evaluation. Instead, we rely on automated metrics to evaluate RAG-LLMs (§2.3). We emphasize that high performance on these automated metrics does not guarantee efficacy or safety in real-world clinical applications. However, consistent with standard practices, automated benchmarking is a highly practical necessity for rapid development cycles, particularly in the medical domain where human expert evaluation is prohibitively expensive. Once models achieve a reasonable baseline performance, which our current results show the RAG-LLMs we experimented with have not yet reached, subsequent and more resource-intensive evaluations with medical experts will be required.

Furthermore, prior work on medical text generation highlights a key limitation: semantic similarity metrics alone cannot reliably assess clinical correctness. Metrics such as BERTSCORE, which we use for assessing the plausibility of self-explanations (§2.3.1), capture semantic similarity but often fail to penalize factual errors or hallucinated findings. Recent evaluation frameworks adopt hybrid approaches that combine more tightly semantic similarity metrics with factuality-oriented assessments. These methods explicitly verify clinical entities and their relations, often using knowledge bases such as the Unified Medical Language System (UMLS)[5] or structured schemas like RADGRAPH (Jain et al., 2021), thereby evaluating both semantic similarity and clinical accuracy. The use of hybrid metrics of this kind, which are more often used in multimodal LLMs (Yu et al., 2023; Hoque et al., 2025), has not been explored in BIOBENCH.

Finally, a practical limitation of our work lies in the extensive computational resources required to fully replicate our baseline experiments. To establish robust and reproducible baselines, we evaluated several large-scale models (LLAMA 3.1 70B (Dubey et al., 2024), LLAMA 3.3 70B (Dubey et al., 2024), MED42-V2-70B (Wu et al., 2023), and MIXTRAL-8×7B (Jiang et al., 2024a)) which necessitated a dedicated cluster of 8×A100 GPUs (with 40GB of memory each). For future researchers evaluating their own models, the inference phase may similarly incur high computational costs, as computing the contribution ratios for PCC-SHAP requires three separate inference repetitions per instance (see Appendix, §D.1.3 and Figure 4). However, once model inference is complete, calculating the metrics themselves is highly resource-efficient. The only computationally demanding component of the evaluation pipeline is the LLM-judge (GPT-OSS-120B (OpenAI, 2025)) used to compute the NLI scores when assessing the plausibility of cited passages (§2.3.2). However, we deliberately selected this locally deployable model to ensure the benchmark remains sustainable, reproducible, and fully accessible to academic institutions, as we also discuss in the Appendix (§G).

[5] https://www.nlm.nih.gov/research/umls/index.html

References

Asma Ben Abacha, Wen wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2025. MEDEC: A Benchmark for Medical Error Detection and Correction in Clinical Notes. *Preprint*, arXiv:2412.19260.

Abdelrahman Abdallah, Jamshid Mozafari, Bhawna Piryani, and Adam Jatowt. 2025. DeAR: Dual-stage document reranking with reasoning agents via LLM distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5710–5723, Suzhou, China. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Henning Bartsch, Ole Jorgensen, Domenic Rosati, Jason Hoelscher-Obermaier, and Jacob Pfau. 2023. Self-consistency of large language models under ambiguity. *Preprint*, arXiv:2310.13439.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. *Preprint*, arXiv:2112.04426.

Marc Brinner and Sina Zarrieß. 2023. Model interpretability and rationale extraction by input mask optimization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13722–13744, Toronto, Canada. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *International Conference on Learning Representations*.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalou, and Prodromos Malakasiotis. 2021a. Regulatory compliance through doc2doc information retrieval: A case study in eu/uk legislation where text similarity has limitations. *Preprint*, arXiv:2101.10726.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021b. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Aaron Chan, Shaoliang Nie, Liang Tan, Xiaochang Peng, Hamed Firooz, Maziar Sanjabi, and Xiang Ren. 2022. Frame: Evaluating rationale-label consistency metrics for free-text rationales.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023. REV: Information-theoretic evaluation of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2007–2030, Toronto, Canada. Association for Computational Linguistics.

Peter Ch'en, Wesley Day, Ryan Pekson, Juan Barrientos, William Burton, Allison Ludwig, Sunit Jariwala, and Todd Cassese. 2025. Gpt-4 generated answer rationales to multiple choice assessment questions in undergraduate medical education. *BMC Medical Education*, 25.

George Chrysostomou and Nikolaos Aletras. 2022. Flexible instance-specific rationalization of nlp models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:10545–10553.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, Matthew M Churpek, Anoop Mayampurath, Frank Liao, Cherodeep Goswami, Karen K. Wong, Brian W. Patterson, and Majid Afshar. 2025. Automating evaluation of ai text generation in healthcare with a large language model (llm)-as-a-judge. *medRxiv*.

Amin Dada, Marie Bauer, Amanda Butler Contreras, Osman Alperen Koracs, Constantin Seibold, Kaleb E. Smith, and Jens Kleesiek. 2024. Does biomedical training lead to better medical performance?

Michiel de Jong, Yury Zemlyanskiy, Joshua Ainslie, Nicholas FitzGerald, Sumit Sanghai, Fei Sha, and William Cohen. 2023. FiDO: Fusion-in-decoder optimized for stronger performance and faster inference. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11534–11547, Toronto, Canada. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Lisa C Adams, and Keno K Bressem. 2025. Evaluating the effectiveness of biomedical fine-tuning for large language models on clinical tasks. *Journal of the American Medical Informatics Association*, page ocaf045.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2288–2292, New York, NY, USA. Association for Computing Machinery.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4957, Brussels, Belgium. Association for Computational Linguistics.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? In *Annual Meeting of the Association for Computational Linguistics*.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Mahmudul Hoque, Raisa Nusrat Chowdhury, Md Rakibul Hasan, Ojonugwa Oluwafemi Ejiga Peter, Fahmi Khalifa, and Md Mahmudur Rahman. 2025. An empirical evaluation of low-rank adapted vision–language models for radiology image captioning. *Bioengineering*, 12(12).

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. 2023. Can Large Language Models Explain Themselves? A Study of LLM-Generated Self-Explanations. *Preprint*, arXiv:2310.11207.

Bogdan Ionescu, Henning Müller, Ana-Maria Drăgulinescu, Johannes Rückert, Asma Ben Abacha, Alba Garcıa Seco de Herrera, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia Sabrina Schmidt, Tabea M.G. Pakull, Hendrik Damm, Benjamin Bracke, Christoph M. Friedrich, Alexandra-Georgiana Andrei, Yuri Prokopchuk, Dzmitry Karpenka, Ahmedkhan Radzhabov, Vassili Kovalev, Cécile Macaire, Didier Schwab, Benjamin Lecouteux, Emmanuelle Esperança-Rodier, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, Steven A. Hicks, Michael A. Riegler, Vajira Thambawita, Andrea Storås, Pål Halvorsen, Maximilian Heinrich, Johannes Kiesel, Martin Potthast, and Benno Stein. 2024. Overview of ImageCLEF 2024: Multimedia retrieval in medical applications. In *Experimental IR Meets Multilinguality, Multimodality,*

and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Grenoble, France. Springer Lecture Notes in Computer Science LNCS.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *Preprint*, arXiv:2311.11944.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised Dense Information Retrieval with Contrastive Learning.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2024. Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24(1).

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Matthew Lungren, Andrew Ng, Curtis Langlotz, and Pranav Rajpurkar. 2021. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *PhysioNet*. Version 1.0.0.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024a. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Zhengping Jiang, Yining Lu, Hanjie Chen, Daniel Khashabi, Benjamin Van Durme, and Anqi Liu. 2024b. Rora: Robust free-text rationale evaluation. *Preprint*, arXiv:2402.18678.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Shikai Jing, Jihong Liu, and Hongfei Zhan. 2013. A new design rationale knowledge evaluation method. In *2013 International Conference on Computer-Aided Design and Computer Graphics*, pages 385–386.

Akira Kawabata and Saku Sugawara. 2023. Evaluating the rationale understanding of critical reasoning in logical reading comprehension. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Akira Kawabata and Saku Sugawara. 2024. Rationale-aware answer verification by pairwise self-evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16178–16196, Miami, Florida, USA. Association for Computational Linguistics.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations (ICLR)*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2024. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. *Scientific Data*, 10:170.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson E. Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, John Kernion, Kamil.e Lukovsiut.e, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy D. Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Janina Brauner, Sam Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. *ArXiv*, abs/2307.13702.

Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. 2017. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126, Copenhagen, Denmark. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024a. Long-context LLMs Struggle with Long In-context Learning. *Preprint*, arXiv:2404.02060.

Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, and Jun Huang. 2024b. AlphaFin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 773–783, Torino, Italia. ELRA and ICCL.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400, Singapore. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the Middle: How Language Models Use Long Contexts. *Preprint*, arXiv:2307.03172.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2023. Can large language models reason about medical questions? *Preprint*, arXiv:2207.08143.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 295–337, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Preprint*, arXiv:2405.20362.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Yuki Momii, Tetsuya Takiguchi, and Yasuo Ariki. 2024. RAG-fusion based information retrieval for fact-checking. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 47–54, Miami, Florida, USA. Association for Computational Linguistics.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. *Layer-Wise Relevance Propagation: An Overview*, pages 193–209. Springer International Publishing, Cham.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2019. Interrogating the explanatory power of attention in neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 221–230, Hong Kong. Association for Computational Linguistics.

Amy C. Moreno and Danielle S. Bitterman. 2024. Toward clinical-grade evaluation of large language models. *International Journal of Radiation Oncology Biology Physics*, 118(4):916–920.

Mahammad Namazov, Tomáš Koref, and Ivan Habernal. 2026. Legal experts disagree with rationale extraction techniques for explaining ecthr case outcome classification. *Preprint*, arXiv:2601.12419.

Siddharth Narayanan, James D. Braza, Ryan-Rhys Griffiths, Manu Ponnapati, Albert Bou, Jon Laurent, Ori Kabeli, Geemi Wellawatte, Sam Cox, Samuel G. Rodriques, and Andrew D. White. 2024. Aviary: training language agents on challenging scientific tasks. *arXiv preprent arXiv:2412.21154*.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Salvador Lima López, Eulália Farré-Maduell, Luis Gasco, Martin Krallinger, and Georgios Paliouras. 2023. Overview of BioASQ 2023: The Eleventh BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18–21, 2023, Proceedings*, page 227–250, Berlin, Heidelberg. Springer-Verlag.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Salvador Lima-López, Eulàlia Farré-Maduell, Martin Krallinger, Natalia Loukachevitch, Vera Davydova, Elena Tutubalina, and Georgios Paliouras. 2024. Overview of BioASQ 2024: The Twelfth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part II*, page 3–27, Berlin, Heidelberg. Springer-Verlag.

Nghia Trung Ngo, Chien Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2024. Comprehensive and Practical Evaluation of Retrieval-Augmented Generation Systems for Medical Question Answering. *Preprint*, arXiv:2411.09213.

Shuai Niu, Jing Ma, Liang Bai, Zhihua Wang, Yida Xu, Yunya Song, and Xian Yang. 2024. Multimodal clinical reasoning through knowledge-augmented rationale generation.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. 2018. BioRead: A new dataset for biomedical reading comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. BioMRC: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149, Online. Association for Computational Linguistics.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089, Bangkok, Thailand. Association for Computational Linguistics.

Nicholas Pipitone and Ghita Houir Alami. 2024. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. *Preprint*, arXiv:2408.10343.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2024. Evaluating the Reliability of Self-Explanations in Large Language Models. *Preprint*, arXiv:2407.14487.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. 3(4):333–389.

Haggai Roitman. 2018. Enhanced Performance Prediction of Fusion-based Retrieval. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '18, page 195–198, New York, NY, USA. Association for Computing Machinery.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020a. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020b. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

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. *Preprint*, arXiv:2301.12652.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Suzanna Sia, Anton Belyy, Amjad Almahairi, Madian Khabsa, Luke Zettlemoyer, and Lambert Mathias. 2022. Logical satisfiability of counterfactuals for faithful explanations in nli.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. *Preprint*, arXiv:1312.6034.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprent arXiv:2409.13740*.

Tingyu Song, Yilun Zhao, Siyue Zhang, Chen Zhao, and Arman Cohan. 2025. LimRank: Less is more for reasoning-intensive information reranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20625–20639, Suzhou, China. Association for Computational Linguistics.

Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62, Florence, Italy. Association for Computational Linguistics.

Yumeng Sun, Renhan Zhang, Renzi Meng, Lian Lian, Heyi Wang, and Xuehui Quan. 2025. Fusion-Based Retrieval-Augmented Generation for Complex Question Answering with LLMs. In *2025 8th International Conference on Computer Information Science and Application Technology (CISAT)*, pages 116–120.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Juraj Vladika and Florian Matthes. 2024. Comparing knowledge sources for open-domain scientific claim verification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2103–2114, St. Julian's, Malta. Association for Computational Linguistics.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

Boxin Wang, Wei Ping, Peng Xu, Lawrence McAfee, Zihan Liu, Mohammad Shoeybi, Yi Dong, Oleksii Kuchaiev, Bo Li, Chaowei Xiao, Anima Anandkumar, and Bryan Catanzaro. 2023a. Shall we pretrain autoregressive language models with retrieval? a comprehensive study. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7763–7786, Singapore. Association for Computational Linguistics.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. 2023b. Primacy Effect of ChatGPT. *Preprint*, arXiv:2310.13206.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2020. Measuring association between labels and free-text rationales. In *Conference on Empirical Methods in Natural Language Processing*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *Preprint*, arXiv:2304.14454.

Junde Wu, Jiayuan Zhu, and Yunli Qi. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *Preprint*, arXiv:2408.04187.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Lawrence K. Q. Yan, Qian Niu, Ming Li, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Benji Peng, Ziqian Bi, Pohsun Feng, Keyu Chen, Tianyang Wang, Yunze Wang, Silin Chen, Ming Liu, and Junyu Liu. 2024a. Large Language Model Benchmarks in Medical Tasks. *Preprint*, arXiv:2410.21348.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024b. Corrective retrieval augmented generation. *Preprint*, arXiv:2401.15884.

Qinyuan Ye, Iz Beltagy, Matthew Peters, Xiang Ren, and Hannaneh Hajishirzi. 2023. FiD-ICL: A fusion-in-decoder approach for efficient in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8158–8185, Toronto, Canada. Association for Computational Linguistics.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs. *Preprint*, arXiv:2407.02485.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, Hawaii. Association for Computational Linguistics.

Kerem Zaman and Shashank Srivastava. 2025. A causal lens for evaluating faithfulness metrics. *Preprint*, arXiv:2502.18848.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722, St. Julian's, Malta. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Suifeng Zhao, Zhuoran Jin, Sujian Li, and Jun Gao. 2025. Finragbench-v: A benchmark for multimodal rag with visual citation in the financial domain. *Preprint*, arXiv:2505.17471.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2308–2313, New York, NY, USA. Association for Computing Machinery.

Appendix

A Additional related work

A.1 Retrieval Augmented Generation (RAG)

RAG (Lewis et al., 2020b; Gao et al., 2024; Xiong et al., 2024) is a popular extension of LLMs. Roughly speaking, given a query, it retrieves passages from trusted document collections, then tries to make the LLM base its response on the retrieved passages. In the simplest case, RAG simply appends the retrieved passages to the prompt of the LLM, but more elaborate approaches have also been proposed (Lin et al., 2023; Formal et al., 2021; Khandelwal et al., 2020; de Jong et al., 2023; Ye et al., 2023; Shi et al., 2023; Borgeaud et al., 2022; Wang et al., 2023a; Izacard et al., 2024). RAG-LLMs can use up-to-date information from evolving document collections, avoiding the cost of re-training the LLMs. RAG may also reduce hallucinations (Shuster et al., 2021), increasing trustworthiness, and may contribute to explainability, if RAG-LLMs are allowed to self-reflect on retrieved passages (Asai et al., 2024; Yan et al., 2024b).

A.2 Attributive explanations

Attributive explanation methods aim to highlight the parts of the input (e.g., tokens) that mostly affected the model's decision. Attention (Bahdanau et al., 2014) has been used to produce explanations of this kind (Lee et al., 2017; Ghaeini et al., 2018). However, there is criticism that attention scores do not provide meaningful explanations of the inner workings of models (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019; Brunner et al., 2020; Moradi et al., 2019; DeYoung et al., 2020), and that they can be manipulated to mislead (Pruthi et al., 2020). Hence, they are unsuitable for generating faithful explanations, even if attention-based explanations may appear plausible to users (Zhang et al., 2024). Post-hoc attributive explanation methods, on the other hand, seek to determine the parts of the input that *truly* (faithfully) influenced most the model's response. They include gradient-based techniques (Simonyan et al., 2014; Selvaraju et al., 2017; Sundararajan et al., 2017; Montavon et al., 2019) and perturbation tests, such as LIME (Ribeiro et al., 2016) and Shapley values (Lundberg and Lee, 2017).

Rationale extraction (i.e., identifying the most influential spans of tokens in the model input) can also be performed using attributive explanation methods. It was initially proposed to explain model predictions in classification tasks (DeYoung et al., 2020; Zaidan and Eisner, 2008; Rajani et al., 2019; Strout et al., 2019; Yu et al., 2019; Chalkidis et al., 2021b). Since then, a lot of research has been conducted towards rationale evaluation (Kawabata and Sugawara, 2024; Chen et al., 2023; Kawabata and Sugawara, 2023; Jiang et al., 2024b; Jing et al., 2013; Brinner and Zarrieß, 2023; Carton et al., 2020; Namazov et al., 2026; Chrysostomou and Aletras, 2022), as well as their application in medical QA (Ch'en et al., 2025; Niu et al., 2024). It is often not entirely clear, however, if the goal is to evaluate the plausibility (convincing to experts) or faithfulness of extractive rationales. The ERASER benchmark (DeYoung et al., 2020), for example, proposed several measures, including comprehensiveness and sufficiency, to evaluate the alignment between automatically and manually extracted rationales (in effect, plausibility), but also to measure to what extent the rationales influence the decisions of the models (mostly related to faithfulness).

BIOBENCH uses methods from rationale extraction (comprehensiveness, sufficiency, and sparsity) to evaluate faithfulness of cited passages (§2.3.4, §D.2), but these passages are selected (as citations included in the response) by LLMs, not via attention or post-hoc attribution. We also evaluate faithfulness of cited passages by combining passage contribution scores based on Shapley values and MAP in our new *faithful grounding* method (§2.3.4).

A.3 Chain-of-thought explanations

Unlike attributive explanations, self-explanations are COT-like texts (Kojima et al., 2024; Brown et al., 2020), generated by LLMs. Their plausibility can be automatically evaluated by comparing to expert explanations (§2.3.1), but their faithfulness is more difficult to assess. Self-consistency methods (Parcalabescu and Frank, 2024; Madsen et al., 2024; Bartsch et al., 2023) use scores from attributive methods to examine if a model focuses on the same input parts (e.g., tokens) when generating its self-explanation as when it produces an answer.[6] We build upon CC-SHAP, which employs Shapley values when evaluating the faithfulness of self-explanations (Parcalabescu and Frank, 2024). Although CC-SHAP considers LLMs without RAG, we modify it for RAG-LLMs to operate on retrieved

[6] For methods of this kind to work correctly, the self-explanation must not be visible to the model when computing the contribution scores for the answer (see §D.1.3).

passages (PCC-SHAP, §2.3.3, §D.1).

A.4 Faithfulness evaluation

It is often important to understand why a machine learning model makes a decision. Towards this direction, faithfulness is defined as the degree to which an explanation accurately reflects a model's true internal decision-making process (Jacovi and Goldberg, 2020). To measure faithfulness, baseline approaches such as simulatability (testing if an explanation alone yields the original prediction (Chan et al., 2022; Wiegreffe et al., 2020; Hase and Bansal, 2020; Hase et al., 2020; Doshi-Velez and Kim, 2017)), COT corruptions (degrading reasoning steps to observe prediction stability (Lanham et al., 2023)), and counterfactual or random input edits (checking if explanations accurately reflect targeted or random semantic perturbations (Atanasova et al., 2023; Sia et al., 2022)) test the sufficiency and the necessity of a model-generated explanation by perturbing inputs or reasoning steps and observing changes in the answer. However, by treating LLMs as black boxes, these generation-based baselines inherently conflate lack of faithfulness with a model's sensitivity to prompt phrasing (Turpin et al., 2023).

Furthermore, faithfulness is usually conflicted with plausibility, which merely measures how convincing an explanation appears to an external observer. This distinction is critical given the recent semantic drift in frameworks like RAGAS[7], used to evaluate RAG-LLMs, where faithfulness is co-opted to describe factual grounding, assessing whether an output is supported by retrieved data rather than evaluating true reasoning-to-explanation fidelity. We use the prompt templates from RAGAS[8], assessing to what extent an external observer (LLM-judge) would consider the cited passages by RAG-LLMs as plausible support of their self-explanations as an additional measure for plausibility (§2.3.2, §G).

Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017) bypasses the structural confounders of structural baselines by operating directly within the feature attribution space. In particular, SHAP systematically masks parts (e.g., tokens) of the input to a model, to determine the *contribution* that each part of the input had to the model's output (e.g., classification decision). Consistency Scores based on SHAP (CC-SHAP) (Parcalabescu and Frank, 2024) is an offsping of SHAP that provides a single contribution ratio per input token, when the output is an entire generated text, comprising an answer and a self-explanation. If the self-explanation is faithful, then the contributions of the input tokens should be similar when generating the self-explanation and the answer.

To rigorously prove the reliability, or diagnosticity, of metrics like CC-SHAP against structural baselines, recent frameworks employ causal interventions (Zaman and Srivastava, 2025). By using techniques such as in-context or parameter-based knowledge editing to alter a model's internal facts, one can establish a controlled ground truth. This causal lens allows for a systematic evaluation of how accurately and reliably CC-SHAP can distinguish genuinely faithful reasoning from unfaithful, post-hoc rationalizations compared to its black-box predecessors. Based on the experimental results, CC-SHAP consistently outperforms the structural baselines in accurately diagnosing faithful vs. unfaithful explanations. We introduce a new variant of CC-SHAP for self-explanations generated by RAG-LLMs (PCC-SHAP, §2.3.3, D.1), which is based on the same idea, but operates at a coarser granularity, in particular on retrieved passages rather than their individual tokens.

B BIOBENCH dataset details

BIOBENCH benchmark contains selected data from four biomedical QA datasets in English, namely PUBMEDQA (Jin et al., 2019), yes/no and summary questions from BIOASQ-QA (Krithara et al., 2023), MEDMCQA (Pal et al., 2022), and HEALTHFC (Vladika et al., 2024). By using these four datasets, we consider three task types in biomedical QA, namely information extraction from PUBMED articles (PUBMEDQA, BIOASQ-QA), medical entrance examinations (MEDMCQA), and fact check-

[7] https://docs.ragas.io/

[8] Note that even though the measure is called *faithfulness*, it simply checks whether the claims made in the generated answer can be directly inferred from the retrieved context, but without examining the actual inferencing of the model, i.e., if/how the retrieved context causally affects the model's answer, which makes it a *plausibility* metric using the terminology adopted by previous work (Jacovi and Goldberg, 2020). See https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/faithfulness/. Consider, for example, a case where the passages include the gold ones (supporting the correct answer) and the model has generated the correct answer, but relying entirely on its parametric memory, ignoring its cited passages. The LLM-judge would give a high score, whereas PCC-SHAP would give a low score. Both would be correct, because they measure different things (*plausible* support vs. *causal* support, *plausibility* vs. *faithfulness*). However, as we also mention in the limitations of BIOBENCH (§6), we as well do not analyze the inner workings of LLMs to understand how they reach their decisions, which is an extremely challenging research question that remains unanswered.

ing health-related claims (HEALTHFC). These tasks reflect real-world biomedical contexts.

With the exception of the first task, which is arguably the most important for biomedical researchers, we included only one representative dataset per task, to reduce the required computing resources. Therefore, for medical entrance examinations, we have not included MMLU (Hendrycks et al., 2021) and MED-QA (Jin et al., 2021) in BIOBENCH, although they were contained in MIRAGE. MEDMCQA (Pal et al., 2022) was preferred to the MMLU and MED-QA, because it includes human-authored explanations, similar to the self-explanations required by LLMs.

On the other hand, we included in BIOBENCH the HEALTHFC dataset on fact-checking (Vladika et al., 2024), a task setting not covered at all by MIRAGE. Furthermore, we also incorporated the summary questions from BIOASQ-QA (Krithara et al., 2023), which were not included in MIRAGE (only the yes/no questions of BIOASQ-QA were included in MIRAGE). All the selected datasets include detailed explanations and require systems to retrieve and combine information from trusted sources in order to generate plausible self-explanations (§2.3.1). Moreover, we avoided synthetic datasets, such as BIOREAD (Pappas et al., 2018) and BIOMRC (Pappas et al., 2020) to avoid evaluating RAG-LLMs in factually incorrect information. BIOBENCH is the only current biomedical QA benchmark evaluating RAG-LLMs in all our three selected task settings.

B.1 The PUBMEDQA dataset

PUBMEDQA (Jin et al., 2019) requires answering biomedical research questions with 'yes', 'no', or 'maybe', by retrieving information from PUBMED abstracts. It contains articles whose titles are questions and whose abstracts are well-structured, including an introduction, method description, most important results, conclusion. It comprises 273,500 samples, but only 1,000 of them, the PQA-L (labeled) subset, have been validated by human experts to contain conclusions answering the questions of the titles. Hence, systems are typically evaluated only on PQA-L. Examples are shown in Table 8. We use the same 500 PQA-L samples as MIRAGE for evaluation. The remaining PQA-L samples are used for few-shot learning (§2.2).

B.2 The MEDMCQA dataset

MEDMCQA (Pal et al., 2022) is a multiple-choice QA dataset with questions (in English) from the Indian medical entrance exams, covering 2,400 healthcare topics from 21 medical subjects (e.g., diagnostic reasoning and treatment, surgery, pharmacology). It contains 181,155 training samples, 6,000 validation, 6,000 test samples, with four possible answers per question and a single correct one. We considered the 4,813 samples included in the test set of MIRAGE. We discarded, however, samples that did not include explanations of their answers or topic labels; samples of the latter kind very often contained explanations that were mere rephrases of the answers, much shorter than the self-explanations we expect the LLMs to generate (Table 9). This left 420 samples, which we use for evaluation. Samples from the training set satisfying the same criteria are used for few-shot learning.

Examples	
Question:	Do patients with rheumatoid arthritis established on methotrexate and folic acid 5 mg daily need to continue folic acid supplements long term?
Answer:	Yes.
Explanation:	It is important to continue FA supplementation over the long term in patients on methotrexate and FA in order to prevent them discontinuing treatment because of mouth ulcers or nausea and vomiting. Our data suggest that FA supplementation is also helpful in preventing neutropenia, with very little loss of efficacy of methotrexate.
Question:	Can decisional algorithms replace global introspection in the individual causality assessment of spontaneously reported ADRs?
Answer:	No.
Explanation:	Decisional algorithms are sensitive methods for the detection of ADRs, but they present poor specificity. A reference method was not identified. Algorithms do not replace GI and are not definite alternatives in the individual causality assessment of suspected ADRs.
Question:	Does ultrasound imaging before puncture facilitate internal jugular vein cannulation?
Answer:	Maybe.
Explanation:	Prepuncture ultrasound evaluation did not improve the result of right internal jugular vein cannulation compared with the respiratory jugular venodilation-guided approach. When the landmark was not observed, however, the prepuncture ultrasound guidance was helpful in facilitating the cannulation.

Table 8: Three examples with different answers from the PQA-L (labeled) subset of PUBMEDQA.

B.3 The BIOASQ-QA dataset

The BIOASQ-QA dataset (Krithara et al., 2023), from the 11th BIOASQ (Nentidis et al., 2023), comprises 4,719 training and 330 test questions curated by biomedical experts. There are four types of

Examples	
Question:	A patient presented with flat vtopped iolaceous papules, thinning of nails and onycholysis. What is the microscopic finding observed in this patient?
Answer:	Basal cell degeneration
Explanation:	Basal degeneration
Question:	Which vitamin is required for glycogen Phosphorylase?
Answer:	PLP.
Explanation:	Glycogen phosphorylase is the rate limiting enzyme of glycogenolysis. And it requires PLP. The active form of vitamin B6 is the coenzyme pyridoxal phosphate (PLP). PLP can be synthesized from the three compounds pyridoxine, pyridoxal and pyridoxamine. This PLP for this enzyme Glycogen phosphorylase is not required as co-enzyme, but it acts as a phosphate donor. Enzyme glycogen phosphorylase will cut glycogen a (1-4) bond apa and the glucose released are transferred in Glucose-1-phosphate and that phosphate is taken from PLP.

Table 9: Two MEDMCQA samples. The first one is not included in BIOBENCH, because the explanation is a rephrase of the answer. The second one is included.

questions: yes/no, factoid, list, summary. The first three types require both 'exact' answers (e.g., entity names) and 'ideal' answers (longer answers, similar to the self-explanations we require LLMs to generate). Summary questions do not admit exact answers (e.g., "Why/how...?") and require only 'ideal' answers, typically longer than those of the other types. They also necessitate combining evidence from multiple passages retrievable from PUBMED. Examples of all types are shown in Table 10. The ground truth of each question includes relevant PUBMED articles and passages.

We have included yes/no and summary questions in BIOBENCH. The former are used because their binary candidate responses facilitate automatic evaluation of factuality; for the same reason, they are also leveraged in previous work (Xiong et al., 2024; Thakur et al., 2021; Vladika and Matthes, 2024; Yan et al., 2024a). Summary questions are included because of their longer 'ideal' answers. BIOBENCH includes the 82 yes/no and 79 summary questions of the test set for evaluation. The yes/no and summary questions of the training set are utilized for few-shot learning (§2.2).

B.4 The HEALTHFC dataset

HEALTHFC (Vladika et al., 2024) is a benchmark for evidence-based medical fact-checking. It includes 750 health-related claims, each presented in English and German. Each claim is accompanied by trustworthy evidence sources (e.g., clinical trials, systematic reviews) and medical expert veracity labels, namely 'supported' (202 claims), 'refuted' (125), and 'not enough information' (423). Claims labeled as 'supported' or 'refuted' are further annotated on a three-point scale reflecting the strength and frequency of the supporting evidence, but we do not use this information. The veracity label of each claim is accompanied by an expert-authored concise justification (40 words on average). Examples are shown in Table 11. We use all 750 claims (in English) for evaluation in BIOBENCH. Since

Examples	
Question:	Is levosimendan effective for amyotrophic lateral sclerosis?
Type:	Yes/No
Answer:	No.
Explanation:	No. Levosimendan was not superior to placebo in maintaining respiratory function in a broad population with amyotrophic lateral sclerosis. Although levosimendan was generally well tolerated, increased heart rate and headache occurred more frequently with levosimendan than with placebo.
Question:	What is the alternative microexon splicing by RBFOX2 and PTBP1 associated with?
Type:	Factoid
Answer:	Colorectal cancer metastasis.
Explanation:	Alternative microexon splicing by RBFOX2 and PTBP1 is associated with metastasis in colorectal cancer.
Question:	What is the triad of the Eagle-Barrett Syndrome?
Type:	List
Answer:	Absent or defective abdominal wall muscles, urinary tract abnormalities, and bilateral cryptorchidism.
Explanation:	Eagle-Barrett Syndrome (EBS) is a rare congenital condition characterized by the triad of absent or defective abdominal wall muscles, urinary tract abnormalities, and bilateral cryptorchidism.
Question:	What is Cinryze?
Type:	Summary
Answer:	Intravenous nanofiltered human C1 inhibitor (C1-INH NF) concentrate (Cinryze®) is used as a direct replacement of deficient levels of plasma C1 inhibitor in patients with hereditary angioedema (HAE). In the EU, C1-INH NF concentrate 1000 U is indicated in the treatment, pre-procedural prevention, and routine prevention of angioedema attacks in adults and adolescents with HAE. Intravenous C1-INH NF concentrate 1000U effectively relieved angioedema attacks in patients with HAE.

Table 10: Four examples of different types from the test set of BIOASQ-QA. In summary questions, the explanations are shown as answers.

there is no training set in HEALTHFC, few-shot learning is not possible and therefore, we evaluate RAG-LLMs only in zero-shot settings.

Examples	
Question:	Can masks reduce corona infections when worn by a large proportion of the population?
Answer:	Supported.
Explanation:	International studies suggest that the number of corona infections decreases when many people wear masks - whether fabric, surgical or FFP2 masks. However, it is not possible to precisely quantify how great the protective effect actually is. It probably depends on the type of masks and the proportion of people who wear them. Studies show slow spread After three years of pandemic, there are now relatively meaningful data that masks can slow down the spread of the Corona virus. Surgical masks, in turn, seem to reduce the risk of infecting themselves with the corona virus.
Question:	Does taking probiotics shorten the duration of acute infectious diarrhea?
Answer:	Refuted.
Explanation:	According to current studies, the duration of disease cannot be reduced by taking probiotics. This assessment is reasonably well supported. Thus, at least in children, taking live probiotics is unlikely to affect whether they are still ill after 48 hours. Thus, a probiotic apparently has no clear influence on the resolution of the disease.
Question:	Does oil pulling promote dental health?
Answer:	Not enough information.
Explanation:	Studies conducted to date have been poorly done and are not conclusive. Thus, they cannot answer this question. However, the investigations are too short, they have too many shortcomings and the results are too inaccurate to prove that. The effect of oil pulling on dental health remains unclear – at least as long as there are no meaningful studies that have been carried out according to strict scientific criteria. However, this does not prove that oil pulling prevents gum inflammation as well as antibacterial chlorhexidine. Otherwise, the studies are too inadequate to be meaningful. The real question remains unanswered.

Table 11: Three examples from HEALTHFC, corresponding to various degrees of evidential substantiation.

B.5 Dataset statistics

When evaluating faithfulness of self-explanations, PCC-SHAP compares to what extent the RAG-LLMs consider every retrieved passage (as captured by their contribution ratios $c_{self,i}$ and $c_{ans,i}$, see §D.1.1) when producing the self-explanation of an answer vs. the answer itself. Assessing the contributions of the retrieved passages rather than their tokens, greatly speeds up the calculations. To motivate this choice, we provide token/passage statistics for all different configurations of LLMs, retrievers, and datasets of BIOBENCH (Tab. 12, 13, 14, 15, 16).

LLM Family	Retriever	Passages (k)	Tokens	Avg. Tokens (per passage)
LLAMA 3	MEDCPT	5	1743	348.60
		10	3370	337.00
		15	4766	317.73
		20	6616	330.80
	CONTRIEVER	5	1539	307.80
		10	2949	294.90
		15	4131	275.40
		20	5354	267.70
	BM25	5	1965	393.00
		10	3697	369.70
		15	5666	377.73
		20	7333	366.65
MIXTRAL	MEDCPT	5	2016	403.20
		10	3873	387.30
		15	5412	360.80
		20	7493	374.65
	CONTRIEVER	5	1819	363.80
		10	3488	348.80
		15	4877	325.13
		20	6394	319.70
	BM25	5	2331	466.20
		10	4334	433.40
		15	6625	441.67
		20	8594	429.70

Table 12: Statistics from the PUBMEDQA dataset across different LLM families and retrievers, from the test set of BIOBENCH (see §B.1 for details).

C BIOBENCH retriever details

We use the MEDRAG toolkit (Xiong et al., 2024), which facilitates prompting RAG-LLMs with retrieved passages, after extending it to support the additional LLMs (§1) and datasets (§2.1) of BIOBENCH, as well as few-shot settings. We use PUBMED as the collection to retrieve passages from, which is required for PUBMEDQA (Jin et al., 2019) and BIOASQ-QA (Krithara et al., 2023). Each passage in BIOBENCH is the abstract of a scientific article from the open-access subset of PUBMED. For each combination of LLM, retriever, and dataset, following MIRAGE we originally retrieve 32 passages, of which we keep the k most relevant ones per question. We experiment with different values of $k \in \{5, 10, 15, 20\}$; we do not use larger values since no improvements were observed in preliminary experiments. This was expected, since the performance of most LLMs has been shown to drop dramatically with contexts exceeding 20K tokens in other domains as well (Li et al., 2024a).

For few-shot training, we employ the particular retriever of the RAG-LLM being tested, and use as

LLM Family	Retriever	Passages (k)	Tokens	Avg. Tokens (per passage)
LLaMA 3	MEDCPT	5	2104	420.80
		10	3517	351.70
		15	5238	349.20
		20	7141	357.05
	CONTRIEVER	5	1518	303.60
		10	2782	278.20
		15	3887	259.13
		20	5103	255.15
	BM25	5	1642	328.40
		10	3130	313.00
		15	4770	318.00
		20	6642	332.10
MIXTRAL	MEDCPT	5	2351	470.20
		10	3937	393.70
		15	5917	394.47
		20	8055	402.75
	CONTRIEVER	5	1766	353.20
		10	3284	328.40
		15	4588	305.87
		20	5968	298.40
	BM25	5	1963	392.60
		10	3694	369.40
		15	5648	376.53
		20	7880	394.00

Table 13: Statistics from the MEDMCQA dataset across different LLM families and retrievers, from the test set of BIOBENCH (see §B.2 for details).

LLM Family	Retriever	Passages (k)	Tokens	Avg. Tokens (per passage)
LLaMA 3	MEDCPT	5	1797	359.40
		10	3652	365.20
		15	5337	355.80
		20	7219	360.95
	CONTRIEVER	5	1101	220.20
		10	2233	223.30
		15	4188	279.20
		20	5437	271.85
	BM25	5	1729	345.80
		10	3956	395.60
		15	5889	392.60
		20	7791	389.55
MIXTRAL	MEDCPT	5	2102	420.40
		10	4281	428.10
		15	6241	416.07
		20	8463	423.15
	CONTRIEVER	5	1306	261.20
		10	2683	268.30
		15	5108	340.53
		20	6639	331.95
	BM25	5	2075	415.00
		10	4773	477.30
		15	7138	475.87
		20	9482	474.10

Table 14: Statistics from the HEALTHFC dataset across different LLM families and retrievers, from the test set of BIOBENCH (see §B.4 for details).

LLM Family	Retriever	Passages (k)	Tokens	Avg. Tokens (per passage)
LLaMA 3	MEDCPT	5	2090	418.00
		10	4170	417.00
		15	5944	396.27
		20	8037	401.85
	CONTRIEVER	5	1455	291.00
		10	2935	293.50
		15	4921	328.07
		20	6454	322.70
	BM25	5	1659	331.80
		10	3145	314.50
		15	4779	318.60
		20	6640	332.00
MIXTRAL	MEDCPT	5	2489	497.80
		10	4909	490.90
		15	6970	464.67
		20	9438	471.90
	CONTRIEVER	5	1661	332.20
		10	3343	334.30
		15	5640	376.00
		20	7449	372.45
	BM25	5	1912	382.40
		10	3691	369.10
		15	5590	372.67
		20	7702	385.10

Table 15: Statistics from the BIOASQ-QA (yes/no questions) across different LLM families and retrievers, from the test set of BIOBENCH (see §B.3 for details).

LLM Family	Retriever	Passages (k)	Tokens	Avg. Tokens (per passage)
LLaMA 3	MEDCPT	5	939	187.80
		10	1933	193.30
		15	2810	187.34
		20	4090	204.50
	CONTRIEVER	5	1196	239.20
		10	2132	213.20
		15	3147	209.80
		20	4383	219.15
	BM25	5	1288	257.60
		10	3013	301.30
		15	4729	315.27
		20	6099	304.95
MIXTRAL	MEDCPT	5	1076	215.20
		10	2253	225.30
		15	3264	217.60
		20	4732	236.60
	CONTRIEVER	5	1374	274.80
		10	2436	243.60
		15	3576	238.40
		20	5013	250.65
	BM25	5	1376	275.20
		10	3276	327.60
		15	5173	344.87
		20	6717	335.85

Table 16: Statistics from BIOASQ-QA (summary questions) across different LLM families and retrievers, from the test set of BIOBENCH (see §B.3 for details).

few-shot examples the n training instances with the most similar questions to the current question being answered. Question similarity is computed as the inner product of question embeddings (for dense retrievers), or as the BM25 score between the questions. Again, we consider different values of $n \in \{1, 2, 3, 4, 5\}$. When RAG is not used, we employ the MEDCPT retriever to obtain few-shot examples, because it performed better in MIRAGE when using PUBMED (Xiong et al., 2024).

We do not include reranking (Nogueira et al., 2020; Abdallah et al., 2025; Song et al., 2025; Zhuang et al., 2023; BehnamGhader et al., 2024) or fusion-based retrieval (Momii et al., 2024; Sun et al., 2025; Roitman, 2018) in the baseline experiments, but they can be easily added, along with additional retrievers, by researchers wishing to measure their effect on BIOBENCH.

D Evaluation metrics details

D.1 Passage-level CC-SHAP (PCC-SHAP)

We introduce a new variant of CC-SHAP (§A.4) for self-explanations generated by RAG-LLMs, dubbed passage-level CC-SHAP (PCC-SHAP). PCC-SHAP is based on the same ideas as CC-SHAP, but operates at a coarser granularity, in particular retrieved passages rather than input tokens in order to evaluate RAG-LLMs. Assessing the contributions of the retrieved passages rather than their (many more) tokens, greatly speeds up calculations (see the token/passage statistics in §B.5). It also provides contribution scores that correspond more directly to self-contained pieces of knowledge, commonly expressed by passages (e.g., recommended usage of a medicine, symptoms of a disease), as opposed to individual tokens that usually express more fragmented information. Hence, considering the contribution of each retrieved passage provides a higher-level, more coherent window to the model's reasoning (the knowledge it is based on) than when assessing the contribution of individual tokens.

D.1.1 Method details

Given the set $G^{(n)}$ of all passages retrieved for a particular question $\mathbf{x}^{(n)}$, we assess the contribution of each passage $g_i^{(n)} \in G^{(n)}$ (when generating the answer or self-explanation) by considering all possible subsets (*coalitions* in SHAP terminology), $S_i^{(n)} \subseteq G^{(n)} \setminus \{g_i^{(n)}\}$ of passages that do not include $g_i^{(n)}$. Then, for each $S_i^{(n)}$, we examine the change in the behavior of the LLM when its input is $S_i^{(n)}$, with $g_i^{(n)}$ added or not.

Let $\mathcal{L}_{self}^{(n)}$ be a $d_{self}^{(n)}$-dimensional vector containing the log-probabilities (as computed by the LLM) of the tokens in the *self-explanation* for question $\mathbf{x}^{(n)}$, where $d_{self}^{(n)}$ is the length of the self-explanation (in tokens). When $g_i^{(n)}$ is added to $S_i^{(n)}$, we measure the change in $\mathcal{L}_{self}^{(n)}$, $\Delta\mathcal{L}_{self,i}^{(n)}$, defined below. This is the *marginal contribution* of $g_i^{(n)}$ with respect to $S_i^{(n)}$. The same process is followed to compute the marginal contributions of the passages $g_i^{(n)}$ when generating the answer to $\mathbf{x}^{(n)}$, now measuring the change in $\mathcal{L}_{ans}^{(n)}$, $\Delta\mathcal{L}_{ans,i}^{(n)}$, defined similarly.

The Shapley value (importance) $\phi_{self,i}^{(n)}$ of a retrieved passage $g_i^{(n)} \in G^{(n)}$ when generating the self-explanation for question $\mathbf{x}^{(n)}$ is the weighted average of the marginal contributions of $g_i^{(n)}$ w.r.t. all passage subsets $S_i^{(n)} \subseteq G^{(n)} \setminus \{g_i^{(n)}\}$ that do not include it. The weight $w_i^{(n)}$ of each contribution depends on the number of elements in $S_i^{(n)}$ and the total number of passages retrieved $\left(|G_i^{(n)}|\right)$. More concretely, $w_i^{(n)}$ is the binomial coefficient that corresponds to a set of $|G^{(n)}| - 1$ elements (passages) from which we only keep $|S_i^{(n)}|$. Formally:

$$w_i^{(n)} = \frac{|S_i^{(n)}|!\,\left(|G^{(n)}| - |S_i^{(n)}| - 1\right)!}{|G^{(n)}|!}, \quad (1)$$

$$\Delta\mathcal{L}_{self,i}^{(n)} = \left(\mathcal{L}_{self,\,S_i^{(n)} \cup \{g_i^{(n)}\}}^{(n)} - \mathcal{L}_{self,S_i^{(n)}}^{(n)}\right), \quad (2)$$

$$\phi_{self,i}^{(n)} = \sum_{S_i^{(n)}} w_i^{(n)} \Delta\mathcal{L}_{self,i}^{(n)}, \quad (3)$$

where $\mathcal{L}_{self,\,S_i^{(n)} \cup \{g_i^{(n)}\}}^{(n)}$ and $\mathcal{L}_{self,S_i^{(n)}}^{(n)}$ denote $\mathcal{L}_{self}^{(n)}$ when the input to the LLM is $S_i^{(n)}$ with $g_i^{(n)}$ included or not, respectively. The Shapley values $\phi_{ans,i}^{(n)}$ of the passages when generating the answer are defined similarly, now using $\mathcal{L}_{ans}^{(n)}$. Notice that $\phi_{self,i}^{(n)}$ and $\phi_{ans,i}^{(n)}$ are $d_{self}^{(n)}$-dimensional and $d_{ans}^{(n)}$-dimensional vectors, respectively.

For all subsets $S_i^{(n)}$ and $S_i^{(n)} \cup \{g_i^{(n)}\}$, we retain the original answers and self-explanations of the LLM, i.e., those obtained when all the passages of $G^{(n)}$ were given to the LLM; what changes when feeding the LLM with particular subsets of passages instead of the entire $G^{(n)}$ are the log-probabilities of the output tokens of the original answers and self-explanations. Hence, the log-probabilities in $\mathcal{L}_{self}^{(n)}$ will always be for the same tokens (the output

tokens of the original self-explanation), regardless of the subset of passages we are considering; and similarly for the log-probabilities in $\mathcal{L}^{(n)}_{ans}$.

Similarly to CC-SHAP (Parcalabescu and Frank, 2024), we normalize the Shapley values $\phi^{(n)}_{self,i}$ and $\phi^{(n)}_{ans,i}$ of every retrieved passage $g^{(n)}_i$, obtained for the answer and self-explanation respectively, to ensure they are comparable for self-explanations vs. answers, which have different lengths. We define the resulting normalized Shapley values $\rho_{self,i}$ for self-explanations hereunder; the normalized Shapley values $\rho_{ans,i}$ for answers are defined similarly. We use the normalized values $\rho_{self,i}$ to compute the aggregated (over $d^{(n)}_{self}$ dimensions) contribution ratio $c_{self,i}$ of each passage when generating self-explanations; and similarly for $c_{ans,i}$ when generating answers. We finally calculate the average distance between the contribution ratios for the self-explanations (c_{self}) and answers (c_{ans}) across all questions ($\mathcal{Q}$), obtaining our PCC-SHAP metric for faithfulness of self-explanations.

$$\rho^{(n)}_{self,i} = \frac{\phi^{(n)}_{self,i}}{\sum_{j=1}^{|G^{(n)}|} |\phi^{(n)}_{self,j}|}, \quad (4)$$

$$c^{(n)}_{self,i} = \frac{\sum_{k=1}^{d^{(n)}_{self}} \rho^{(n)}_{self,i,k}}{|d^{(n)}_{self}|}, \quad (5)$$

$$\text{PCC-SHAP} = 1 - \frac{1}{|\mathcal{Q}|} \sum_{n=1}^{|\mathcal{Q}|} 1 - \cos\left(c^{(n)}_{ans}, c^{(n)}_{self}\right) \quad (6)$$

D.1.2 Approximate PCC-SHAP

Computing one contribution ratio per retrieved passage (§D.1.1), rather than per individual token of the retrieved passages, greatly speeds up the calculations of PCC-SHAP. However, evaluating all possible coalitions $S^{(n)}_i \subseteq G^{(n)} \setminus \{g^{(n)}_i\}$ of retrieved passages that do not include the passage $g^{(n)}_i$ being scored remains impractical for large datasets. As a shortcut, it is common in SHAP-based methods to *sample* coalitions $S^{(n)}_i$ from $G^{(n)} \setminus \{g^{(n)}_i\}$ when computing the contribution of $g^{(n)}_i$. In our case, when sampling coalitions $S^{(n)}_i$ of a passage $g^{(n)}_i$, the weights $w^{(n)}_i$ of the contributions ($\Delta\mathcal{L}^{(n)}_{ans,i}$, $\Delta\mathcal{L}^{(n)}_{self,i}$) of the coalitions are initially computed as when considering all possible coalitions (§D.1.1), but are then normalized to form a probability distribution over the sampled coalitions of $g^{(n)}_i$.

To further speed-up the calculations for PCC-SHAP, we use a single maximal coalition $S^{(n)}_i = G^{(n)} \setminus \{g^{(n)}_i\}$. In effect, we assess the contribution of passage $g^{(n)}_i$ by observing the effect of removing it from the retrieved passages. We show experimentally (Fig. 2) that PCC-SHAP scores remain similar even when sampling more (smaller) coalitions $S^{(n)}_i$, provided that very small coalitions (e.g., including only $g^{(n)}_i$ and another retrieved passage) are avoided. We argue that coalitions of the latter kind should be avoided anyway, since they ignore most of the other retrieved passages. When using SHAP for traditional feature selection, it makes sense to consider even small coalitions, since we assess the predictive power of each feature. In our case, however, the goal is to assess how much the model's output is affected by each retrieved passage $g^{(n)}_i$, not to assess the predictive power of g by itself.

We also compute faithful grounding (FG) scores (§2.3.4) using the approximate contribution ratios, for coalitions of different sizes. Figure 3 shows that the scores are not affected by the coalition sizes and, hence, a single maximal coalition suffices.

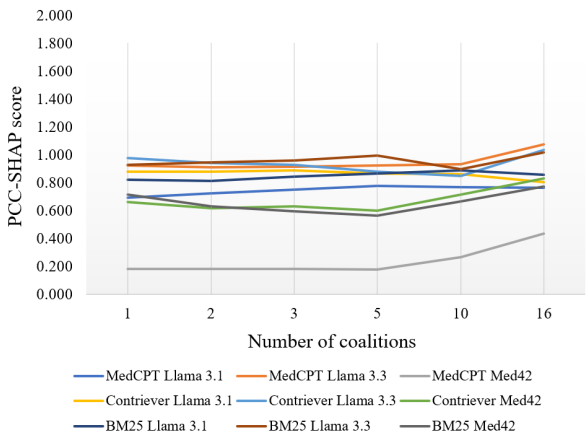

Figure 2: PCC-SHAP scores as cosine distances between contribution ratios for self-explanations and answers (§D.1.1), for a single maximal coalition (left) or increasingly more (and increasingly smaller) coalitions (right). Indicative results for the yes/no questions of BIOASQ-QA with 5 retrieved passages. The maximum number of possible coalitions is 16. The scores are reasonably stable when only large coalitions are used, but gradually change when small coalitions are also considered.

D.1.3 Collecting probabilities in PCC-SHAP

Given a question, RAG retrieves passages from trusted document collections, then it appends them to the prompt of the LLM along with instructions to ground its response to the retrieved passages

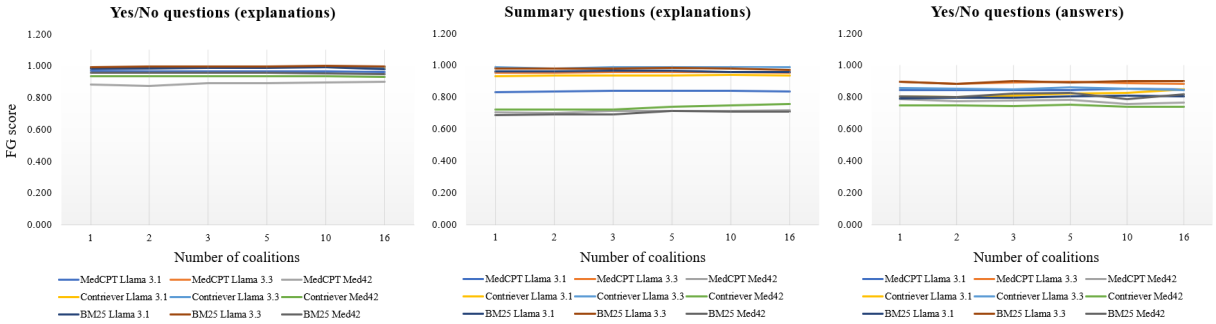

Figure 3: Faithful Grounding (FG) scores (MAP) for self-explanations or answers (§2.3.4), for a single maximal coalition (left) or increasingly more (and smaller) coalitions (right). Indicative results for the yes/no and summary questions of BIOASQ-QA with 5 retrieved passages. The maximum number of possible coalitions is 16.

(§A.1). In BIOBENCH, we require the response to contain a list of citations (pointing to retrieved passages that support the answer), a self-explanation (COT-like) of the answer, and the answer itself, in that order. However, this way the LLM can see (via masked multi-head attention) which passages it has cited, when generating the self-explanation and answer. Similarly, it can see its self-explanation when generating its answer. By contrast, for c_{self} to correctly indicate which retrieved passages mostly influenced the self-explanation, the citations must not be visible when computing c_{self}. Similarly, both the citations and the self-explanation must not be visible when computing c_{ans}.

To address this issue, we perform inference three times per question, to calculate the marginal contributions $\left(\phi^{(n)}_{self,i}, \phi^{(n)}_{ans,i}\right)$ of PCC-SHAP (§D.1.1). Specifically, we initially prompt the RAG-LLM to generate the citations, self-explanation, and answer, in that order. We retain the retrieved passages and the tokens of the generated citations, self-explanation and answer, then perform inference two additional times with the same retrieved passages to obtain log-probabilities for the *original* (from the initial inference) tokens of the self-explanation and answer, respectively. In the first inference repetition, we instruct the LLM to omit the citations, to obtain log-probabilities for the tokens of the original self-explanation unaffected by the citations. In the second repetition, we also instruct the LLM to omit the self-explanation, to obtain log-probabilities for the tokens of the answer unaffected by the citations and self-explanation. The prompts used and detailed examples are shown in §F. Figure 4 provides an illustration of the different inference repetitions performed to collect the log-probabilities $\mathcal{L}^{(n)}_{self}$ and $\mathcal{L}^{(n)}_{self}$ for the self-explanations and answers respectively.

D.2 Comprehensiveness, sufficiency, sparsity

To assess the faithfulness of cited passages, we additionally used *comprehensiveness* and *sufficiency* (Jacovi and Goldberg, 2020; DeYoung et al., 2020), previously used for gradient- and attention-based rationales (§A.2), to evaluate the faithfulness of the cited passages. We also employed a *sparsity* measure, similar in spirit to sparsity constraints of previous work (Lei et al., 2016; Chalkidis et al., 2021b), which we properly modified for RAG-LLMs.

To measure *sufficiency*, we again perform an inference repetition (§D.1.3), now feeding the LLM with only the cited passages of each question, removing the instructions about generating citations. We compare the accuracy of the LLM's answers when fed with all the retrieved passages per question vs. only the cited passages. If the cited passages are sufficient (they adequately support the answer), the accuracy of the LLM should not deteriorate when fed only with them; it might even improve, since irrelevant passages may have been filtered out. Hence, we measure sufficiency as the decrease in accuracy (compared to the original accuracy), when only cited passages are provided.

Similarly, to measure *comprehensiveness* we provide the same questions to the LLM, now excluding from the retrieved passages the cited ones. If all the passages that support the answer have been cited, then the accuracy of the LLM should deteriorate substantially. Again, we measure the decrease, compared to the initial accuracy with all passages; this time the larger the decrease the better.

Following previous work (Carton et al., 2020; Namazov et al., 2026), we use normalized versions of comprehensiveness and sufficiency. Given the set $G^{(n)}$ of all passages retrieved for a question $\mathbf{x}^{(n)}$ and the set $R^{(n)} \subseteq G^{(n)}$ of only the cited retrieved passages, let $f\left(G^{(n)}\right)$ and $f\left(R^{(n)}\right)$ be the

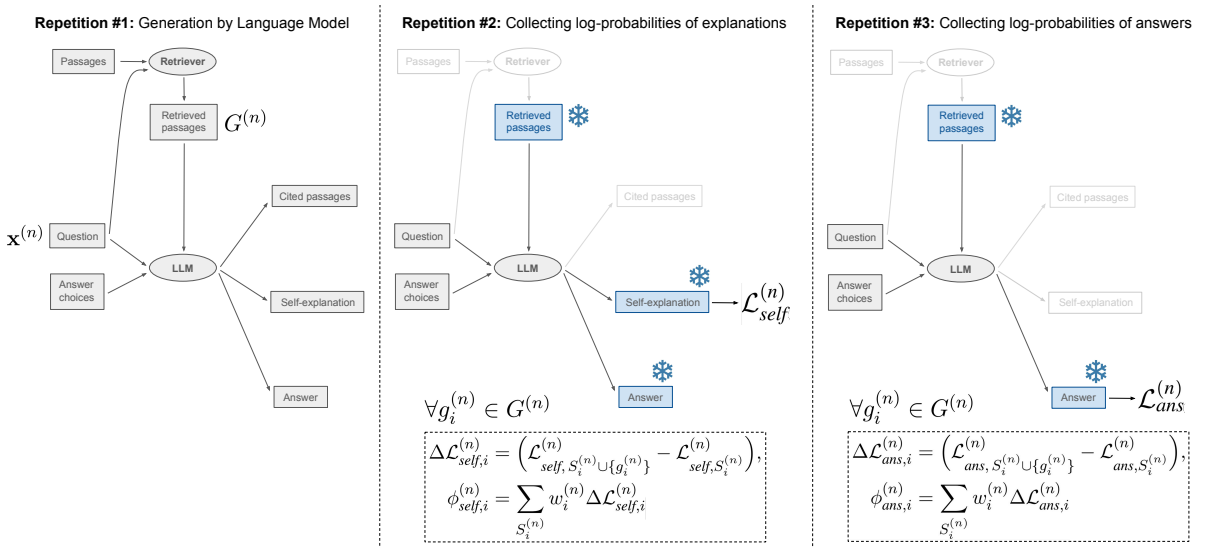

Figure 4: Illustration of the inference repetitions performed to collect the log-probabilities $\mathcal{L}^{(n)}_{self}$ and $\mathcal{L}^{(n)}_{self}$ for the self-explanations and answers respectively. The coalition weights are defined in Eq. 1 and are normalized (in approximate PCC-SHAP) to form a valid probability distribution, as explained in §D.1.2. PCC-SHAP compares to what extent the RAG-LLMs consider every retrieved passage (as captured by the aggregated, normalized contribution ratios, $c_{self,i}$ and $c_{ans,i}$ of each passage, defined in Eq. 5), when producing the self-explanation of an answer vs. the answer itself. Faithful Grounding (§2.3.2) and Contextual Awareness (§2.3.5) are also based on $c_{self,i}$ and $c_{ans,i}$.

accuracy of the LLM when the passages in $G^{(n)}$ or $R^{(n)}$ are included in the prompt, respectively. Then, for question $\mathbf{x}^{(n)}$, normalized sufficiency and comprehensiveness are defined as follows.

$$\text{NS}(n) = 1 - \frac{\text{ReLU}\left[f\left(G^{(n)}\right) - f\left(R^{(n)}\right)\right]}{f\left(G^{(n)}\right)} \quad (7)$$

$$\text{NC}(n) = 1 - \frac{\text{ReLU}\left[f\left(G^{(n)}\right) - f\left(G^{(n)} \setminus R^{(n)}\right)\right]}{f\left(G^{(n)}\right)} \quad (8)$$

The same formulae for $\text{NS}(n)$, $\text{NC}(n)$ also apply when using the (averaged over questions) semantic similarity of the generated answers to the ground truth ones as $f(\cdot)$, instead of accuracy. We use BERTSCORE (§2.3.1) as semantic similarity because it provides a highly intuitive 0-to-1 scoring scale, unlike the unbounded or negative scores of alternative metrics included in BIOBENCH-EVAL.

Unfortunately, sufficiency and comprehensiveness are problematic when systems are free to extract rationales of any size; a system that always selects the entire input as rationale (in our case, a RAG-LLM that cites all the retrieved passages) will have perfect sufficiency and comprehensiveness. Hence, sufficiency and comprehensiveness scores are not comparable across RAG-LLMs, unless one forces them to cite the same number of retrieved passages per question.[9] However, this would be an unrealistic constraint, since different RAG-LLMs may be basing their answers on different numbers of retrieved passages. To investigate this issue, we define an additional measure, *sparsity*, as the percentage of retrieved passages the LLM does *not* cite. A low sparsity score indicates the LLM may be citing irrelevant passages. More importantly, very different sparsity scores across RAG-LLMs signal that sufficiency and comprehensiveness are not comparable across RAG-LLMs. Table 17 shows that the sparsity scores of the RAG-LLMs we experimented with are indeed very different. Hence, we recommend using faithful grounding (Table 6) instead of normalized sufficiency (Eq. 7) and comprehensiveness (Eq. 8), even though we report (Table 17) them as well for completeness.

D.3 BERTSCORE, BARTSCORE, BLEURT

BERTSCORE computes the similarity between a machine-generated and a reference text, using context-aware token embeddings. It provides three metrics: precision, recall, F_1. We use F_1, which combines the other two. We feed BERTSCORE with context-aware embeddings from DEBERTA XL (He et al., 2021), a combination found to correlate well with human judgments in medical image captioning (Ionescu et al., 2024). BARTSCORE computes

[9] Previous work on extractive rationales includes similar sparsity constraints, e.g., in the training loss of a classifier whose decisions are explained by the rationales (Lei et al., 2016; Chalkidis et al., 2021b).

Dataset	RAG-LLM	ZERO-SHOT with RAG				
		Normalized Sufficiency		Normalized Comprehensiveness		Sparsity (↑)
		Accuracy (↑)	BERTSCORE (↑)	Accuracy (↑)	BERTSCORE (↑)	
PUBMEDQA	MIXTRAL	47.56	**100.00**	93.73	**100.00**	**51.24**
	LLAMA 3.1	65.99	**100.00**	89.40	**100.00**	31.16
	LLAMA 3.3	65.32	**100.00**	93.34	**100.00**	44.84
	MED42	**67.45**	**100.00**	94.91	**100.00**	49.12
BIOASQ-QA (yes/no)	MIXTRAL	74.67	**100.00**	99.75	**100.00**	32.09
	LLAMA 3.1	86.08	**100.00**	99.41	98.99	21.40
	LLAMA 3.3	93.51	98.70	99.68	**100.00**	34.19
	MED42	**98.12**	84.91	**100.00**	**100.00**	**38.14**
BIOASQ-QA (summary)	MIXTRAL	–	**99.11**	–	**100.00**	34.94
	LLAMA 3.1	–	86.34	–	99.04	27.85
	LLAMA 3.3	–	94.20	–	**100.00**	**47.59**
	MED42	–	98.01	–	**100.00**	43.80
MEDMCQA	MIXTRAL	84.61	89.59	99.25	99.71	47.86
	LLAMA 3.1	96.26	99.31	**100.00**	99.47	25.86
	LLAMA 3.3	98.06	99.67	**100.00**	**100.00**	49.29
	MED42	**100.00**	**100.00**	99.75	**100.00**	**49.57**
HEALTHFC	MIXTRAL	**100.00**	**100.00**	**100.00**	**100.00**	34.32
	LLAMA 3.1	**100.00**	**100.00**	99.94	**100.00**	20.16
	LLAMA 3.3	**100.00**	97.18	**100.00**	**100.00**	40.29
	MED42	**100.00**	**100.00**	**100.00**	**100.00**	**43.49**

Table 17: Plausibility of cited passages (%, secondary metrics: normalized comprehensiveness, sufficiency, and sparsity). Zero-shot results with RAG, obtained with the BM25 retriever. We use $k = 5$ passages per question.

the log-likelihood of the machine-generated text conditioned on the reference, using BART (Lewis et al., 2020a). BLEURT (Sellam et al., 2020b) returns a score measuring the paraphrase quality (in terms of fluency, semantics, grammar, and style) of a generated text against a reference.

E Oracle retrieval ablation

To evaluate LLMs without requiring them to identify gold passages, an ability they appear to lack according to the low CA_{self} and CA_{ans} scores from pertinence evaluation (§3.5), we conduct an ablation on BIOASQ-QA. In this setting, the context only contains gold relevant passages selected by biomedical experts (§B.3). We compare the original accuracy (factuality) and BERTSCORE (plausibility of self-explanations, §2.3.1) to those obtained under this setting and notice significant improvements in both metrics, as expected (Table 18). Assessing the plausibility and faithfulness of cited passages or pertinence of gold passages is not meaningful here, since all the provided passages are relevant.

F Prompting RAG-LLMs in BIOBENCH

When querying RAG-LLMs, similarly to MIRAGE (Xiong et al., 2024), we instruct them to generate their responses in `json` format, including citations to the cited passages (most influential), self-explanations and answers (in that order). In addition, similarly to MIRAGE (Xiong et al., 2024), we set the temperature to 0 for deterministic outputs. Specifically, we use the following instructions in the input prompt to the LLMs.

- **Zero-shot COT with LLMs (no RAG):** "You are a helpful medical expert, and your task is to answer a multi-choice medical question. Please first think step-by-step and then choose the answer from the provided options. Organize your output in a json formatted as Dict{"step_by_step_thinking": Str(explanation), "answer_choice": Str{A/B/C/...}}. Use the exact keys mentioned ("step_by_step_thinking" and "answer_choice") in the dictionary. Your responses will be used for research purposes only, so please have a definite answer."
- **Zero-shot COT with RAG-LLMs:** "You are a helpful medical expert, and your task is to answer a multi-choice medical question using the relevant documents provided. First mention only the numbers of the relevant documents that you used to answer. Separate the numbers between them using semicolons (e.g 0;1;2). Then please think step-by-step and then choose the answer from the provided options. Do not provide citations in your explanation, which needs to be short and precise. Organize your output in a json formatted as Dict{"useful_relevant_documents": Str(documents), "step_by_step_thinking": Str(explanation), "answer_choice": Str{A/B/C/...}}. Use the exact keys

mentioned ("useful_relevant_documents", "step_by_step_thinking", and "answer_choice") in the dictionary. Your responses will be used for research purposes only, so please have a definite answer."

For the summary questions of BIOASQ-QA, where there are only long (summary) gold answers, we treat the long answers as gold explanations and modify the input prompt appropriately as follows.

- **Zero-shot COT with LLMs for summary questions (no RAG):** "You are a helpful medical expert, and your task is to answer a multi-choice medical question. Please think step-by-step. Organize your output in a json formatted as Dict{"step_by_step_thinking": Str(explanation)}. Use the exact keys mentioned ("step_by_step_thinking") in the dictionary. Your responses will be used for research purposes only, so please have a definite answer."

- **Zero-shot COT with RAG-LLMs for summary questions:** "You are a helpful medical expert, and your task is to answer a medical question using the relevant documents provided. First mention only the numbers of the relevant documents that you used to answer. Separate the numbers between them using semicolons (e.g 0;1;2). Then please think step-by-step and summarize your thoughts in a short response. Do not provide citations in your response, which needs to be short and precise. Organize your output in a json formatted as Dict{"useful_relevant_documents": Str(documents), "step_by_step_thinking": Str(response) }. Use the exact keys mentioned ("useful_relevant_documents", and "step_by_step_thinking") in the dictionary. Your responses will be used for research purposes only, so please have a definite answer."'

When few-shot demonstrators are included in the input prompt, they are formatted according to the instructions provided to the RAG-LLMs, which they are required to follow in their responses. The prompts are adjusted accordingly.

- **Zero-shot COT with LLMs (no RAG):** "You are a helpful medical expert, and your task is to answer a multi-choice medical question using the relevant documents provided. Please think step-by-step and then choose the answer from the provided options. Your explanation needs to be short and precise. Organize your output in a json formatted as Dict{"useful_relevant_documents": Str(documents), "step_by_step_thinking": Str(explanation), "answer_choice": Str{A/B/C/...}}. Use the exact keys mentioned ("step_by_step_thinking", and "answer_choice") in the dictionary. One example of a question-answer pair is provided below, formatted exactly as it is required, followed by a question for you to answer. Your responses will be used for research purposes only, so please have a definite answer."

- **Zero-shot COT with RAG-LLMs:** "You are a helpful medical expert, and your task is to answer a multi-choice medical question using the relevant documents provided. First mention only the numbers of the relevant documents that you used to answer. Separate the numbers between them using semicolons (e.g 0;1;2). Then please think step-by-step and then choose the answer from the provided options. Do not provide citations in your explanation, which needs to be short and precise. Organize your output in a json formatted as Dict{"useful_relevant_documents": Str(documents), "step_by_step_thinking": Str(explanation), "answer_choice": Str{A/B/C/...}}. Use the exact keys mentioned ("useful_relevant_documents", "step_by_step_thinking", and "answer_choice") in the dictionary. One example of a question-answer pair is provided below, formatted exactly as it is required, followed by a question for you to answer. Your responses will be used for research purposes only, so please have a definite answer."

- **Zero-shot COT with LLMs for summary questions (no RAG):** "You are a helpful medical expert, and your task is to answer a medical question using the relevant documents provided. Please think step-by-step and summarize your thoughts in a short response. Your response needs to be short and precise. Organize your output in a json

Dataset	RAG-LLM	ZERO-SHOT with oracle RAG		FEW-SHOT with oracle RAG, BM25		FEW-SHOT with oracle RAG, MEDCPT	
		Accuracy (↑)	BERTSCORE (↑)	Accuracy (↑)	BERTSCORE (↑)	Accuracy (↑)	BERTSCORE (↑)
BIOASQ-QA (yes/no)	MIXTRAL	**95.35**	60.29	73.26	58.56	19.77	47.88
	LLAMA 3.1	**95.35**	60.86	**90.70**	62.55	51.16	70.79
	LLAMA 3.3	94.19	**62.93**	89.53	**63.64**	**52.32**	**71.71**
	MED42	77.91	55.95	73.26	56.97	15.12	62.30
BIOASQ-QA (summary)	MIXTRAL	–	66.98	–	47.35	–	36.21
	LLAMA 3.1	–	**71.98**	–	**68.44**	–	**73.52**
	LLAMA 3.3	–	70.21	–	67.94	–	72.77
	MED42	–	61.17	–	56.37	–	63.31

Table 18: Factuality and plausibility of self-explanations (%, accuracy and BERTSCORE respectively). Zero-shot and few-shot results with oracle RAG. We use all gold relevant passages and $n = 5$ few-shot examples (with all their corresponding gold relevant passages) per question, selected with either the BM25 or the MEDCPT retriever (§2.2). Accuracy is not reported for summary questions because they do not admit exact answers (§B.3).

formatted as Dict{"step_by_step_thinking": Str(response)}. Use the exact keys mentioned ("step_by_step_thinking") in the dictionary. One example of a question-answer pair is provided below, formatted exactly as it is required, followed by a question for you to answer. Your responses will be used for research purposes only, so please have a definite answer."

- **Zero-shot COT with RAG-LLMs for summary questions:** "You are a helpful medical expert, and your task is to answer a medical question using the relevant documents provided. First mention only the numbers of the relevant documents that you used to answer. Separate the numbers between them using semicolons (e.g 0;1;2). Then please think step-by-step and summarize your thoughts in a short response. Do not provide citations in your response, which needs to be short and precise. Organize your output in a json formatted as Dict{"useful_relevant_documents": Str(documents), "step_by_step_thinking": Str(response)}. Use the exact keys mentioned ("useful_relevant_documents", and "step_by_step_thinking") in the dictionary. One example of a question-answer pair is provided below, formatted exactly as it is required, followed by a question for you to answer. Your responses will be used for research purposes only, so please have a definite answer."

F.1 Prompting to collect probabilities

The instructions in the input prompts of the three steps for PUBMEDQA and BIOASQ-QA yes/no questions are presented hereunder, accompanied by the responses of LLAMA 3.3 with $k = 20$ and using the BM25 retriever in example `63fa13da201352f04a000001` from BIOASQ-QA. The passages are truncated and replaced by [...] but the complete response is available online.[10]

1. "You are a helpful medical expert, and your task is to answer a multi-choice medical question using the relevant documents provided. First mention only the numbers of the relevant documents that you used to answer. Separate the numbers between them using semicolons (e.g 0;1;2). Then please think step-by-step and then choose the answer from the provided options. Do not provide citations in your explanation, which needs to be short and precise. Organize your output in a json formatted as Dict{"useful_relevant_documents": Str(documents), "step_by_step_thinking": Str(explanation), "answer_choice": Str{A/B/C/...}}. Use the exact keys mentioned ("useful_relevant_documents", "step_by_step_thinking", and "answer_choice") in the dictionary. Your responses will be used for research purposes only, so please have a definite answer."

   Sample input: Here are the relevant documents: Document [0] (Title: Spliceosomal component PRP-40 is a central regulator of microexon splicing.) [...] Document [1] (Title: A highly conserved program of neuronal microexons is misregulated in autistic brains.) [..] Document [2] (Title: MicroExonator enables systematic discovery and quantification of microexons across mouse embryonic development.) [...] Document [3] (Title: Neuronal-specific microexon splicing of TAF1 mRNA is directly regulated by

[10] https://1drv.ms/w/c/250a81d7582e167b/EeWN7C7q4cIIvjIkWa_YNnUB8wFV3hDnkNfTaZbHP8-2ig?e=lvWUWm

SRRM4/nSR100.) [...] Document [4] (Title: Microexon-based regulation of ITSN1 and Src SH3 domains specificity relies on introduction of charged amino acids into the interaction interface.) [...] Document [5] (Title: A novel protein domain in an ancestral splicing factor drove the evolution of neural microexons.) [...] Document [6] (Title: Silencing of SRRM4 suppresses microexon inclusion and promotes tumor growth across cancers.) [...] Document [7] (Title: Pervasive misannotation of microexons that are evolutionarily conserved and crucial for gene function in plants.) [...] Document [8] (Title: Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development.) [...] Document [9] (Title: Differential muscle-type expression of the Drosophila troponin T gene. A 3-base pair microexon is involved in visceral and adult hypodermic muscle specification.) [...] Document [10] (Title: Ribosome profiling at isoform level reveals evolutionary conserved impacts of differential splicing on the proteome.) [...] Document [11] (Title: Prediction of functional microexons by transfer learning.) [...] Document [12] (Title: Temporal regulation of axonal repulsion by alternative splicing of a conserved microexon in mammalian Robo1 and Robo2.) [...] Document [13] (Title: Specialization of the photoreceptor transcriptome by Srrm3-dependent microexons is required for outer segment maintenance and vision.) [...] Document [14] (Title: Alternative microexon splicing by RBFOX2 and PTBP1 is associated with metastasis in colorectal cancer.) [...] Document [15] (Title: Microexon alternative splicing of small GTPase regulators: Implication in central nervous system diseases.) [...] Document [16] (Title: Developmental Attenuation of Neuronal Apoptosis by Neural-Specific Splicing of Bak1 Microexon.) [...] Document [17] (Title: Autism-Misregulated eIF4G Microexons Control Synaptic Translation and Higher Order Cognitive Functions.) [...] Document [18] (Title: Microexons: at the nexus of nervous system development, behaviour and autism spectrum disorder.) [...] Document [19] (Title: HGF-induced migration depends on the PI(3,4,5)P₃-binding microexon-spliced variant of the Arf6 exchange factor cytohesin-1.) [...] Here are the potential choices: A. yes B. no Please think step-by-step and generate your output in json:

RAG-LLM original response (free-text generation): {"useful_relevant_documents": "0", "step_by_step_thinking": "PRP-40 is a core component of the U1 snRNP and is required for inclusion of alternatively spliced, but not constitutively spliced, exons. PRP-40 is particularly required for inclusion of neuronal microexons. Knockdown of the orthologous PRPF40A in mouse neuroblastoma cells causes widespread dysregulation of microexons but not conventionally sized exons. This suggests that PRP-40 regulation of microexons is conserved across species.", "answer_choice": "A"}"

2. "You are a helpful medical expert, and your task is to answer a multi-choice medical question using the relevant documents provided. Please think step-by-step and then choose the answer from the provided options. Your explanation needs to be short and precise. Organize your output in a json formatted as Dict{"step_by_step_thinking": Str(explanation), "answer_choice": Str{A/B/C/...}}. Use the exact keys mentioned ("step_by_step_thinking", and "answer_choice") in the dictionary. Your responses will be used for research purposes only, so please have a definite answer."

   RAG-LLM response to compute probabilities for the COT self-explanation (citations are not included; the generation is controlled using a custom `LogitsProcessor`): {"step_by_step_thinking": "PRP-40 is a core component of the U1 snRNP and is required for inclusion of alternatively spliced, but not constitutively spliced, exons. PRP-40 is particularly required for inclusion of neuronal microexons. Knockdown of the orthologous PRPF40A in mouse neuroblastoma cells causes widespread dysregulation of microexons but not conventionally sized exons. This suggests that PRP-40 regulation of microexons is conserved across species.", "answer_choice": "A"}"

3. "You are a helpful medical expert, and your task is to answer a multi-choice medical question. choose the answer from the pro-

vided options. Organize your output in a json formatted as Dict{"answer_choice": Str{A/B/C/...}}. Use the exact keys mentioned ("answer_choice") in the dictionary. Your responses will be used for research purposes only, so please have a definite answer."

RAG-LLM response to compute probabilities for the answer (now both the citations and the COT self-explanation are not included; the generation is controlled again using the same custom `LogitsProcessor`): {"useful_relevant_documents": "0", "step_by_step_thinking": "PRP-40 is a core component of the U1 snRNP and is required for inclusion of alternatively spliced, but not constitutively spliced, exons. PRP-40 is particularly required for inclusion of neuronal microexons. Knockdown of the orthologous PRPF40A in mouse neuroblastoma cells causes widespread dysregulation of microexons but not conventionally sized exons. This suggests that PRP-40 regulation of microexons is conserved across species.", "answer_choice": "A"}"

For the summary questions, which do not include "exact" answers, we need two steps. The respective prompts are below, with the responses of LLAMA 3.3 with $k = 20$ and using the BM25 retriever in example `64089b21201352f04a000021` from BIOASQ-QA. The passages are truncated and replaced by [...] but the complete response is available online.[11]

1. "You are a helpful medical expert, and your task is to answer a medical question using the relevant documents provided. First mention only the numbers of the relevant documents that you used to answer. Separate the numbers between them using semicolons (e.g 0;1;2). Then please think step-by-step and summarize your thoughts in a short response. Do not provide citations in your response, which needs to be short and precise. Organize your output in a json formatted as Dict{"useful_relevant_documents": Str(documents), "step_by_step_thinking": Str(response)}. Use the exact keys mentioned ("useful_relevant_documents", and "step_by_step_thinking") in the dictionary. One example of a question-answer pair is provided below, formatted exactly as it is required, followed by a question for you to answer. Your responses will be used for research purposes only, so please have a definite answer."

Sample input: Here are the relevant documents: Document [0] (Title: The emerging picture of CDK9/P-TEFb: more than 20 years of advances since PITALRE.) [...] Document [1] (Title: CDK9 as an Appealing Target for Therapeutic Interventions.) [...] Document [2] (Title: Targeting cyclin-dependent kinase 9 in cancer therapy.) Cyclin-dependent kinase (CDK) [...] Document [3] (Title: Cyclin-dependent kinase-9: an RNAPII kinase at the nexus of cardiac growth and death cascades.) [...] Document [4] (Title: CDK9 as a Valuable Target in Cancer: From Natural Compounds Inhibitors to Current Treatment in Pediatric Soft Tissue Sarcomas.) [...] Document [5] (Title: CDK9 a potential target for drug development.) [...] Document [6] (Title: Cyclin-dependent kinase 9 (Cdk9) of fission yeast is activated by the CDK-activating kinase Csk1, overlaps functionally with the TFIIH-associated kinase Mcs6, and associates with the mRNA cap methyltransferase Pcm1 in vivo.) [...] Document [7] (Title: The PP2A-Integrator-CDK9 axis fine-tunes transcription and can be targeted therapeutically in cancer.) [...] Document [8] (Title: Novel CDK9 inhibitor oroxylin A promotes wild-type P53 stability and prevents hepatocellular carcinoma progression by disrupting both MDM2 and SIRT1 signaling.) [...] Document [9] (Title: MALAT1 induces osteosarcoma progression by targeting miR-206/CDK9 axis.) [...] Document [10] (Title: Novel, selective CDK9 inhibitors for the treatment of HIV infection.) [...] Document [11] (Title: Deregulations in the cyclin-dependent kinase-9-related pathway in cancer: implications for drug discovery and development.) [...] Document [12] (Title: Design, synthesis and biological evaluation of methylenehydrazine-1-carboxamide derivatives with (5-((4-(pyridin-3-yl)pyrimidin-2-yl)amino)-1H-indole scaffold: Novel potential CDK9 inhibitors.) [...] Document [13] (Title: Functionally active cyclin-dependent kinase 9 is essential for porcine reproductive

[11] https://1drv.ms/w/c/250a81d7582e167b/EVKBYwlqvUZGs9k5OqscKe0B9OFEn0VNXqQNAQassEgjwA?e=KVHcAa

and respiratory syndrome virus subgenomic RNA synthesis.) [...] Document [14] (Title: Regulation of P-TEFb elongation complex activity by CDK9 acetylation.) [...] Document [15] (Title: Functional inactivation of Cdk9 through oligomerization chain reaction.) [...] Document [16] (Title: Cyclin-dependent kinase 9 (CDK9) is a novel prognostic marker and therapeutic target in osteosarcoma.) [...] Document [17] (Title: Antiapoptotic function of Cdk9 (TAK/P-TEFb) in U937 promonocytic cells.) [...] Document [18] (Title: A role for CDK9-cyclin K in maintaining genome integrity.) [...] Document [19] (Title: CDK9 keeps RNA polymerase II on track.) [...] Here is the question: Can you summarize the function of CDK9? Please think step-by-step and generate your output in json:

RAG-LLM original response (free-text generation): {"useful_relevant_documents": "0;1;2;3;4;5;6;7;8;9;10;11;12;13;14;15;16;17;18;19", "step_by_step_thinking": "CDK9 is a cyclin-dependent kinase that plays a crucial role in the regulation of transcription elongation by RNA polymerase II. It is the catalytic subunit of the positive transcription elongation factor b (P-TEFb) complex and phosphorylates the C-terminal domain of RNAP II, releasing it from promoter-proximal pausing and allowing it to enter productive elongation. CDK9 is involved in various cellular processes, including cell cycle regulation, transcription, and mRNA processing. Its dysfunction has been linked to several diseases, including cancer, HIV, and cardiac hypertrophy. CDK9 inhibitors have been developed as potential therapeutic agents for these diseases. Overall, CDK9 is a key regulator of gene expression and its dysregulation can have significant consequences for cellular function and disease progression."}

2. "You are a helpful medical expert, and your task is to answer a multi-choice medical question using the relevant documents provided. Please think step-by-step and summarize your thoughts in a short response. Do not provide citations in your response, which needs to be short and precise. Organize your output in a json formatted as Dict{"step_by_step_thinking": Str(response) }. Use the exact keys mentioned ("step_by_step_thinking") in the dictionary. Your responses will be used for research purposes only, so please have a definite answer."

RAG-LLM response to compute probabilities for the COT self-explanation (citations are not included; the generation is controlled using a custom `LogitsProcessor`): {"step_by_step_thinking": "CDK9 is a cyclin-dependent kinase that plays a crucial role in the regulation of transcription elongation by RNA polymerase II. It is the catalytic subunit of the positive transcription elongation factor b (P-TEFb) complex and phosphorylates the C-terminal domain of RNAP II, releasing it from promoter-proximal pausing and allowing it to enter productive elongation. CDK9 is involved in various cellular processes, including cell cycle regulation, transcription, and mRNA processing. Its dysfunction has been linked to several diseases, including cancer, HIV, and cardiac hypertrophy. CDK9 inhibitors have been developed as potential therapeutic agents for these diseases. Overall, CDK9 is a key regulator of gene expression and its dysregulation can have significant consequences for cellular function and disease progression."}

G Additional results

We present baseline results from four open-source LLMs and three open-source and widely used retrievers, in zero- and few-shot settings, in order to demonstrate the use of BIOBENCH and its metrics. In addition to the scores reported in §3, here we provide few-shot results with/without RAG, as well as using dense retrievers included in BIOBENCH (§2.2). In terms of factuality, we observe that the observations presented in §3.2 also hold for few-shot prompting of RAG-LLMs (Table 19).

The same conclusions are also reached when using different retrievers, namely MEDCPT (Table 20) and CONTRIEVER (Table 21). Specifically, RAG improves factuality only in datasets that were created (PUBMEDQA, BIOASQ-QA) to assess information extraction from PUBMED (§B.1, §B.3).

In MEDMCQA and HEALTHFC, RAG also harms the plausibility of self-explanations, when using few-shot RAG or dense retrievers. This is consistent with the negative effect RAG has in factuality for those datasets (Tables 19, 20, 21), as wrong answers are more unlikely to be judged as plausible.

Setting	Dataset	LLM	Accuracy (↑)
Zero-shot no RAG	PUBMEDQA	MIXTRAL	31.80%
		LLAMA 3.1	59.80%
		LLAMA 3.3	**60.80%**
		MED42	19.80%
	BIOASQ (yes/no)	MIXTRAL	81.40%
		LLAMA 3.1	81.40%
		LLAMA 3.3	**82.56%**
		MED42	26.74%
	MEDMCQA	MIXTRAL	57.38%
		LLAMA 3.1	77.38%
		LLAMA 3.3	**78.33%**
		MED42	35.00%
	HEALTHFC	MIXTRAL	**56.93%**
		LLAMA 3.1	41.33%
		LLAMA 3.3	41.47%
		MED42	10.40%
Zero-shot with RAG	PUBMEDQA	MIXTRAL	65.60%
		LLAMA 3.1	78.80%
		LLAMA 3.3	**79.00%**
		MED42	42.40%
	BIOASQ (yes/no)	MIXTRAL	87.21%
		LLAMA 3.1	**91.86%**
		LLAMA 3.3	89.53%
		MED42	61.63%
	MEDMCQA	MIXTRAL	52.62%
		LLAMA 3.1	70.00%
		LLAMA 3.3	**73.81%**
		MED42	13.10%
	HEALTHFC	MIXTRAL	**45.33%**
		LLAMA 3.1	40.80%
		LLAMA 3.3	42.53%
		MED42	12.13%
Few-shot no RAG	PUBMEDQA	MIXTRAL	4.80%
		LLAMA 3.1	61.40%
		LLAMA 3.3	**61.60%**
		MED42	58.80%
	BIOASQ (yes/no)	MIXTRAL	11.63%
		LLAMA 3.1	**93.02%**
		LLAMA 3.3	89.53%
		MED42	84.88%
	MEDMCQA	MIXTRAL	10.95%
		LLAMA 3.1	79.76%
		LLAMA 3.3	**80.95%**
		MED42	49.52%
Few-shot with RAG	PUBMEDQA	MIXTRAL	68.20%
		LLAMA 3.1	76.40%
		LLAMA 3.3	**77.20%**
		MED42	2.80%
	BIOASQ (yes/no)	MIXTRAL	82.56%
		LLAMA 3.1	90.70%
		LLAMA 3.3	**91.86%**
		MED42	4.65%
	MEDMCQA	MIXTRAL	44.29%
		LLAMA 3.1	74.76%
		LLAMA 3.3	**78.10%**
		MED42	18.33%

Table 19: Factuality (accuracy, %) zero-shot with/ without RAG, and few-shot ($n = 5$, §2.2) with/without RAG results. RAG uses BM25, $k = 5$. One retrieved passage included in each few-shot example. HEALTHFC contains no training data usable as few-shot examples (§B.4).

Setting	Dataset	LLM	Accuracy (↑)
Zero-shot no RAG	PUBMEDQA	MIXTRAL	31.80%
		LLAMA 3.1	59.80%
		LLAMA 3.3	**60.80%**
		MED42	19.80%
	BIOASQ (yes/no)	MIXTRAL	81.40%
		LLAMA 3.1	81.40%
		LLAMA 3.3	**82.56%**
		MED42	26.74%
	MEDMCQA	MIXTRAL	57.38%
		LLAMA 3.1	77.38%
		LLAMA 3.3	**78.33%**
		MED42	35.00%
	HEALTHFC	MIXTRAL	**56.93%**
		LLAMA 3.1	41.33%
		LLAMA 3.3	41.47%
		MED42	10.40%
Zero-shot with RAG	PUBMEDQA	MIXTRAL	64.20%
		LLAMA 3.1	80.00%
		LLAMA 3.3	**80.40%**
		MED42	42.40%
	BIOASQ (yes/no)	MIXTRAL	87.21%
		LLAMA 3.1	**93.02%**
		LLAMA 3.3	91.86%
		MED42	46.51%
	MEDMCQA	MIXTRAL	49.05%
		LLAMA 3.1	**73.10%**
		LLAMA 3.3	71.67%
		MED42	42.20%
	HEALTHFC	MIXTRAL	**46.13%**
		LLAMA 3.1	42.93%
		LLAMA 3.3	43.33%
		MED42	14.40%
Few-shot no RAG	PUBMEDQA	MIXTRAL	4.80%
		LLAMA 3.1	61.40%
		LLAMA 3.3	**61.60%**
		MED42	58.80%
	BIOASQ (yes/no)	MIXTRAL	11.63%
		LLAMA 3.1	**93.02%**
		LLAMA 3.3	89.53%
		MED42	84.88%
	MEDMCQA	MIXTRAL	10.95%
		LLAMA 3.1	79.76%
		LLAMA 3.3	**80.95%**
		MED42	49.52%
Few-shot with RAG	PUBMEDQA	MIXTRAL	66.80%
		LLAMA 3.1	78.40%
		LLAMA 3.3	**78.80%**
		MED42	3.80%
	BIOASQ (yes/no)	MIXTRAL	81.40%
		LLAMA 3.1	**91.86%**
		LLAMA 3.3	90.70%
		MED42	4.65%
	MEDMCQA	MIXTRAL	45.71%
		LLAMA 3.1	73.57%
		LLAMA 3.3	**76.43%**
		MED42	19.76%

Table 20: Factuality (accuracy, %) zero-shot with/ without RAG, and few-shot ($n = 5$, §2.2) with/without RAG results. RAG now uses the **MEDCPT retriever**, $k = 5$. One retrieved passage included in each few-shot example. HEALTHFC contains no training data usable as few-shot examples (§B.4).

As already discussed, these datasets were not written having particular PUBMED passages in mind, unlike PUBMEDQA and BIOASQ-QA. Hence, their questions are more difficult to match to relevant PUBMED passages by the retrievers, degrading the performance of RAG-LLMs (§3.2).

We use BERTSCORE as our primary metric, as it measures directly the semantic similarity between the generated and gold explanations (§2.3.1). We show results using BERTSCORE for zero-shot and few-shot RAG, using BM25 (Table 22), MEDCPT (Table 23) and CONTRIEVER (Table 24). For com-

Setting	Dataset	LLM	Accuracy (↑)
Zero-shot no RAG	PUBMEDQA	MIXTRAL	31.80%
		LLAMA 3.1	59.80%
		LLAMA 3.3	**60.80%**
		MED42	19.80%
	BIOASQ (yes/no)	MIXTRAL	81.40%
		LLAMA 3.1	81.40%
		LLAMA 3.3	**82.56%**
		MED42	26.74%
	MEDMCQA	MIXTRAL	57.38%
		LLAMA 3.1	77.38%
		LLAMA 3.3	**78.33%**
		MED42	35.00%
	HEALTHFC	MIXTRAL	**56.93%**
		LLAMA 3.1	41.33%
		LLAMA 3.3	41.47%
		MED42	10.40%
Zero-shot with RAG	PUBMEDQA	MIXTRAL	62.80%
		LLAMA 3.1	75.60%
		LLAMA 3.3	**76.80%**
		MED42	42.60%
	BIOASQ (yes/no)	MIXTRAL	81.40%
		LLAMA 3.1	86.05%
		LLAMA 3.3	**87.21%**
		MED42	48.84%
	MEDMCQA	MIXTRAL	53.10%
		LLAMA 3.1	**73.57%**
		LLAMA 3.3	73.33%
		MED42	21.67%
	HEALTHFC	MIXTRAL	47.47%
		LLAMA 3.1	**43.20%**
		LLAMA 3.3	42.27%
		MED42	19.73%
Few-shot no RAG	PUBMEDQA	MIXTRAL	4.80%
		LLAMA 3.1	61.40%
		LLAMA 3.3	**61.60%**
		MED42	58.80%
	BIOASQ (yes/no)	MIXTRAL	11.63%
		LLAMA 3.1	**93.02%**
		LLAMA 3.3	89.53%
		MED42	84.88%
	MEDMCQA	MIXTRAL	10.95%
		LLAMA 3.1	79.76%
		LLAMA 3.3	**80.95%**
		MED42	49.52%
Few-shot with RAG	PUBMEDQA	MIXTRAL	58.20%
		LLAMA 3.1	74.00%
		LLAMA 3.3	**75.20%**
		MED42	8.20%
	BIOASQ (yes/no)	MIXTRAL	79.09%
		LLAMA 3.1	**90.70%**
		LLAMA 3.3	88.37%
		MED42	8.14%
	MEDMCQA	MIXTRAL	48.57%
		LLAMA 3.1	72.86%
		LLAMA 3.3	**73.10%**
		MED42	16.19%

Table 21: Factuality (accuracy, %) zero-shot with/ without RAG, and few-shot ($n = 5$, §2.2) with/without RAG results. RAG now uses the **CONTRIEVER retriever**, $k = 5$. One retrieved passage included in each few-shot example. HEALTHFC contains no training data usable as few-shot examples (§B.4).

pleteness, we also report scores for BARTSCORE and BLEURT (Table 25), in addition to BERTSCORE (Table 4). The two generic LLAMA models achieve the best results with respect to the other two plausibility metrics as well, except for very few cases, in which MED42 excels in terms of BLEURT.

Setting	Dataset	LLM	BERTSCORE (↑)
Zero-shot no RAG	PUBMEDQA	MIXTRAL	**55.05%**
		LLAMA 3.1	54.65%
		LLAMA 3.3	54.09%
		MED42	53.34%
	BIOASQ (yes/no)	MIXTRAL	**58.48%**
		LLAMA 3.1	56.01%
		LLAMA 3.3	54.43%
		MED42	54.76%
	BIOASQ (summary)	MIXTRAL	**58.41%**
		LLAMA 3.1	55.21%
		LLAMA 3.3	57.49%
		MED42	56.99%
	MEDMCQA	MIXTRAL	**54.15%**
		LLAMA 3.1	53.20%
		LLAMA 3.3	53.64%
		MED42	53.25%
	HEALTHFC	MIXTRAL	**56.76%**
		LLAMA 3.1	55.00%
		LLAMA 3.3	54.33%
		MED42	52.47%
Zero-shot with RAG	PUBMEDQA	MIXTRAL	57.76%
		LLAMA 3.1	**60.95%**
		LLAMA 3.3	58.67%
		MED42	54.08%
	BIOASQ (yes/no)	MIXTRAL	56.79%
		LLAMA 3.1	**57.32%**
		LLAMA 3.3	56.88%
		MED42	52.20%
	BIOASQ (summary)	MIXTRAL	58.71%
		LLAMA 3.1	**65.57%**
		LLAMA 3.3	63.77%
		MED42	55.32%
	MEDMCQA	MIXTRAL	51.74%
		LLAMA 3.1	**52.81%**
		LLAMA 3.3	52.53%
		MED42	51.53%
	HEALTHFC	MIXTRAL	53.55%
		LLAMA 3.1	53.68%
		LLAMA 3.3	**53.72%**
		MED42	49.51%
Few-shot no RAG	PUBMEDQA	MIXTRAL	59.13%
		LLAMA 3.1	**60.57%**
		LLAMA 3.3	59.86%
		MED42	60.69%
	BIOASQ (yes/no)	MIXTRAL	43.39%
		LLAMA 3.1	**65.34%**
		LLAMA 3.3	63.87%
		MED42	64.20%
	BIOASQ (summary)	MIXTRAL	38.89%
		LLAMA 3.1	**66.08%**
		LLAMA 3.3	64.87%
		MED42	63.69%
	MEDMCQA	MIXTRAL	40.26%
		LLAMA 3.1	**55.51%**
		LLAMA 3.3	55.26%
		MED42	54.64%
Few-shot with RAG	PUBMEDQA	MIXTRAL	56.77%
		LLAMA 3.1	**63.00%**
		LLAMA 3.3	61.44%
		MED42	55.35%
	BIOASQ (yes/no)	MIXTRAL	55.91%
		LLAMA 3.1	59.10%
		LLAMA 3.3	**59.50%**
		MED42	53.55%
	BIOASQ (summary)	MIXTRAL	44.03%
		LLAMA 3.1	65.72%
		LLAMA 3.3	**65.94%**
		MED42	55.05%
	MEDMCQA	MIXTRAL	49.17%
		LLAMA 3.1	54.14%
		LLAMA 3.3	**53.45%**
		MED42	52.36%

Table 22: Plausibility of self-explanations (BERTSCORE, %), zero-shot with/without RAG, and few-shot with RAG. Same configuration with Table 19.

Setting	Dataset	LLM	BERTSCORE (↑)
Zero-shot no RAG	PUBMEDQA	MIXTRAL	**55.05%**
		LLAMA 3.1	54.65%
		LLAMA 3.3	54.09%
		MED42	53.34%
	BIOASQ (yes/no)	MIXTRAL	**58.48%**
		LLAMA 3.1	56.01%
		LLAMA 3.3	54.43%
		MED42	54.76%
	BIOASQ (summary)	MIXTRAL	**58.41%**
		LLAMA 3.1	55.21%
		LLAMA 3.3	57.49%
		MED42	56.99%
	MEDMCQA	MIXTRAL	**54.15%**
		LLAMA 3.1	53.20%
		LLAMA 3.3	53.64%
		MED42	53.25%
	HEALTHFC	MIXTRAL	**56.76%**
		LLAMA 3.1	55.00%
		LLAMA 3.3	54.33%
		MED42	52.47%
Zero-shot with RAG	PUBMEDQA	MIXTRAL	57.71%
		LLAMA 3.1	**60.04%**
		LLAMA 3.3	58.29%
		MED42	54.03%
	BIOASQ (yes/no)	MIXTRAL	57.28%
		LLAMA 3.1	**57.71%**
		LLAMA 3.3	57.42%
		MED42	52.18%
	BIOASQ (summary)	MIXTRAL	58.44%
		LLAMA 3.1	**62.14%**
		LLAMA 3.3	60.62%
		MED42	55.55%
	MEDMCQA	MIXTRAL	51.67%
		LLAMA 3.1	**52.91%**
		LLAMA 3.3	52.81%
		MED42	51.08%
	HEALTHFC	MIXTRAL	53.57%
		LLAMA 3.1	**53.87%**
		LLAMA 3.3	53.80%
		MED42	49.40%
Few-shot no RAG	PUBMEDQA	MIXTRAL	40.80%
		LLAMA 3.1	60.57%
		LLAMA 3.3	59.86%
		MED42	**60.69%**
	BIOASQ (yes/no)	MIXTRAL	43.39%
		LLAMA 3.1	**65.34%**
		LLAMA 3.3	63.87%
		MED42	64.20%
	BIOASQ (summary)	MIXTRAL	38.89%
		LLAMA 3.1	**66.08%**
		LLAMA 3.3	64.87%
		MED42	63.69%
	MEDMCQA	MIXTRAL	40.26%
		LLAMA 3.1	**55.51%**
		LLAMA 3.3	55.26%
		MED42	54.64%
Few-shot with RAG	PUBMEDQA	MIXTRAL	56.86%
		LLAMA 3.1	**63.44%**
		LLAMA 3.3	61.27%
		MED42	55.43%
	BIOASQ (yes/no)	MIXTRAL	55.60%
		LLAMA 3.1	58.76%
		LLAMA 3.3	**60.13%**
		MED42	53.63%
	BIOASQ (summary)	MIXTRAL	43.97%
		LLAMA 3.1	62.98%
		LLAMA 3.3	**63.12%**
		MED42	55.93%
	MEDMCQA	MIXTRAL	50.83%
		LLAMA 3.1	**54.16%**
		LLAMA 3.3	53.52%
		MED42	52.21%

Table 23: Plausibility of self-explanations (BERTSCORE, %), zero-shot with/without RAG, and few-shot with RAG. Same configuration with Table 20.

Setting	Dataset	LLM	BERTSCORE (↑)
Zero-shot no RAG	PUBMEDQA	MIXTRAL	**55.05%**
		LLAMA 3.1	54.65%
		LLAMA 3.3	54.09%
		MED42	53.34%
	BIOASQ (yes/no)	MIXTRAL	**58.48%**
		LLAMA 3.1	56.01%
		LLAMA 3.3	54.43%
		MED42	54.76%
	BIOASQ (summary)	MIXTRAL	**58.41%**
		LLAMA 3.1	55.21%
		LLAMA 3.3	57.49%
		MED42	56.99%
	MEDMCQA	MIXTRAL	**54.15%**
		LLAMA 3.1	53.20%
		LLAMA 3.3	53.64%
		MED42	53.25%
	HEALTHFC	MIXTRAL	**56.76%**
		LLAMA 3.1	55.00%
		LLAMA 3.3	54.33%
		MED42	52.47%
Zero-shot with RAG	PUBMEDQA	MIXTRAL	57.73%
		LLAMA 3.1	**60.03%**
		LLAMA 3.3	58.33%
		MED42	54.46%
	BIOASQ (yes/no)	MIXTRAL	57.04%
		LLAMA 3.1	**57.34%**
		LLAMA 3.3	56.76%
		MED42	52.05%
	BIOASQ (yes/no)	MIXTRAL	60.74%
		LLAMA 3.1	**65.52%**
		LLAMA 3.3	65.06%
		MED42	56.98%
	MEDMCQA	MIXTRAL	51.71%
		LLAMA 3.1	**52.73%**
		LLAMA 3.3	52.71%
		MED42	51.33%
	HEALTHFC	MIXTRAL	54.34%
		LLAMA 3.1	**54.49%**
		LLAMA 3.3	54.40%
		MED42	50.34%
Few-shot no RAG	PUBMEDQA	MIXTRAL	59.13%
		LLAMA 3.1	60.57%
		LLAMA 3.3	59.86%
		MED42	**60.69%**
	BIOASQ (yes/no)	MIXTRAL	43.39%
		LLAMA 3.1	**65.34%**
		LLAMA 3.3	63.87%
		MED42	64.20%
	BIOASQ (summary)	MIXTRAL	38.89%
		LLAMA 3.1	**66.08%**
		LLAMA 3.3	64.87%
		MED42	63.69%
	MEDMCQA	MIXTRAL	40.26%
		LLAMA 3.1	**55.51%**
		LLAMA 3.3	55.26%
		MED42	54.64%
Few-shot with RAG	PUBMEDQA	MIXTRAL	56.52%
		LLAMA 3.1	**62.35%**
		LLAMA 3.3	61.36%
		MED42	55.36%
	BIOASQ (yes/no)	MIXTRAL	56.62%
		LLAMA 3.1	58.47%
		LLAMA 3.3	**59.02%**
		MED42	53.87%
	BIOASQ (yes/no)	MIXTRAL	50.04%
		LLAMA 3.1	**66.94%**
		LLAMA 3.3	66.67%
		MED42	57.33%
	MEDMCQA	MIXTRAL	50.30%
		LLAMA 3.1	**54.02%**
		LLAMA 3.3	53.60%
		MED42	52.15%

Table 24: Plausibility of self-explanations (BERTSCORE, %), zero-shot with/without RAG, and few-shot with RAG. Same configuration with Table 21.

Dataset	RAG-LLM	ZERO-SHOT RAG BARTSCORE (↑)	ZERO-SHOT RAG BLEURT (↑)	FEW-SHOT RAG BARTSCORE (↑)	FEW-SHOT RAG BLEURT (↑)
PUBMEDQA	MIXTRAL	−0.681	−3.675	−0.755	−3.890
	LLAMA 3.1	**−0.544**	**−3.142**	−0.470	**−2.785**
	LLAMA 3.3	−0.575	−3.449	**−0.444**	−3.382
	MED42	−0.765	−3.543	−0.747	−3.624
BIOASQ-QA (yes/no)	MIXTRAL	−0.822	−3.246	−0.841	−3.342
	LLAMA 3.1	**−0.705**	−3.096	−0.649	**−3.095**
	LLAMA 3.3	−0.693	**−3.047**	**−0.615**	−3.096
	MED42	−0.956	−3.128	−0.912	−3.157
BIOASQ-QA (summary)	MIXTRAL	−0.668	−3.231	−1.071	−4.574
	LLAMA 3.1	**−0.369**	**−3.197**	−0.380	−3.285
	LLAMA 3.3	−0.413	−3.294	**−0.338**	−3.327
	MED42	−0.748	−3.241	−0.741	**−3.242**
MEDMCQA	MIXTRAL	−0.880	−5.013	−0.971	−5.186
	LLAMA 3.1	−0.740	−4.876	**−0.662**	**−4.816**
	LLAMA 3.3	**−0.736**	−4.879	−0.673	−4.823
	MED42	−0.893	**−4.851**	−0.863	−4.853
HEALTHFC	MIXTRAL	−0.816	−4.169	–	–
	LLAMA 3.1	−0.737	−4.117	–	–
	LLAMA 3.3	**−0.718**	**−4.073**	–	–
	MED42	−0.929	−4.127	–	–

Table 25: Plausibility of self-explanations (BARTSCORE, BLEURT), zero-shot with/without RAG, and few-shot ($n = 5$) with RAG (BM25, $k = 5$). One retrieved passage per few-shot example. HEALTHFC contains no training data usable as few-shot examples (§B.4).

Moving on to the plausibility of cited passages, we report results using the passage-level scores (F_1, MAP) for BM25, MEDCPT and CONTRIEVER (Tables 26, 28, 30 respectively) and similarly for the F_1 oriented scores (BERTSCORE, BARTSCORE) for all three retrievers of BIOBENCH (Tables 27, 29, 31 respectively). We observe that all evaluation measures produce very similar rankings, with LLAMA 3.1 achieving consistently higher scores as in Table 5), except for very few cases, for example in MEDMCQA, where MIXTRAL consistently obtains high scores from the LLM-judge. While the default LLM-judge in RAGAS is GPT-4O-MINI (OpenAI, 2024), we instead employ GPT-OSS-120B (OpenAI, 2025), a recent open-weight model released by the same provider. This deliberate substitution ensures our benchmark remains sustainable and reproducible. By utilizing a locally deployable model, we ensure the evaluation framework remains freely accessible to academic institutions with decent resources.

The low passage-level F_1 scores observed consistently indicate that the passages cited by the RAG-LLMs are largely different from those cited by the experts. The high MAP scores in PUBMEDQA are due to the fact that even though the RAG-LLMs fail to cite many of the gold passages, in that particular dataset they manage to place the gold passages that they do cite highly in their citation lists (§3.3). The low F_1-oriented BERTSCORE and BARTSCORE results show that the RAG-LLMs largely also fail to cite passages semantically similar to the gold ones.

Setting	Dataset	LLM	Passage-level scores F_1 (↑)	Passage-level scores MAP (↑)	LLM-judge (↑)
Zero-shot with RAG	PUBMEDQA	MIXTRAL	52.37%	92.98%	74.80%
		LLAMA 3.1	**56.08%**	**93.78%**	**82.08%**
		LLAMA 3.3	43.40%	93.07%	76.55%
		MED42	50.88%	93.45%	80.92%
	BIOASQ (yes/no)	MIXTRAL	27.25%	48.66%	67.29%
		LLAMA 3.1	**30.23%**	**52.80%**	78.95%
		LLAMA 3.3	29.89%	51.62%	72.05%
		MED42	29.38%	54.95%	**80.07%**
	BIOASQ (summary)	MIXTRAL	20.38%	30.60%	75.06%
		LLAMA 3.1	**24.93%**	**35.43%**	**86.30%**
		LLAMA 3.3	24.66%	33.50%	80.95%
		MED42	18.54%	30.20%	72.38%
	MEDMCQA	MIXTRAL	–	–	**44.49%**
		LLAMA 3.1	–	–	36.13%
		LLAMA 3.3	–	–	35.88%
		MED42	–	–	38.96%
	HEALTHFC	MIXTRAL	–	–	77.40%
		LLAMA 3.1	–	–	**76.50%**
		LLAMA 3.3	–	–	75.92%
		MED42	–	–	76.33%
Few-shot with RAG	PUBMEDQA	MIXTRAL	47.55%	93.75%	59.91%
		LLAMA 3.1	**53.59%**	**94.05%**	**85.69%**
		LLAMA 3.3	46.28%	93.45%	79.26%
		MED42	51.05%	92.94%	64.76%
	BIOASQ (yes/no)	MIXTRAL	28.47%	56.41%	65.52%
		LLAMA 3.1	**29.57%**	54.45%	**83.64%**
		LLAMA 3.3	29.08%	51.45%	83.43%
		MED42	29.40%	**58.58%**	57.83%
	BIOASQ (summary)	MIXTRAL	9.93%	26.92%	46.11%
		LLAMA 3.1	**25.53%**	36.89%	**84.45%**
		LLAMA 3.3	25.32%	34.82%	84.29%
		MED42	21.18%	**37.07%**	59.48%
	MEDMCQA	MIXTRAL	–	–	**36.91%**
		LLAMA 3.1	–	–	30.91%
		LLAMA 3.3	–	–	31.66%
		MED42	–	–	28.02%

Table 26: Plausibility of cited passages (passage-level F_1, MAP, LLM-judge scores), zero-shot with/without RAG, and few-shot ($n = 5$) with RAG (BM25, $k = 5$). One retrieved passage per few-shot example. Passage-level measures applicable only to datasets that provide gold passages, with RAG enabled.

Setting	Dataset	LLM	F_1-oriented scores BERTSCORE (↑)	F_1-oriented scores BARTSCORE (↑)
Zero-shot with RAG	PUBMEDQA	MIXTRAL	56.78%	−1.361
		LLAMA 3.1	**57.84%**	**−1.247**
		LLAMA 3.3	56.68%	−1.343
		MED42	57.05%	−1.317
	BIOASQ (yes/no)	MIXTRAL	55.01%	−2.890
		LLAMA 3.1	**55.69%**	**−2.742**
		LLAMA 3.3	55.65%	−2.776
		MED42	55.39%	−2.806
	BIOASQ (summary)	MIXTRAL	56.42%	−2.918
		LLAMA 3.1	**56.71%**	**−2.845**
		LLAMA 3.3	56.50%	−2.907
		MED42	56.66%	−2.909
Few-shot with RAG	PUBMEDQA	MIXTRAL	56.61%	−1.528
		LLAMA 3.1	**57.36%**	**−1.293**
		LLAMA 3.3	56.63%	−1.344
		MED42	57.15%	−1.367
	BIOASQ (yes/no)	MIXTRAL	55.21%	−2.832
		LLAMA 3.1	**55.57%**	**−2.720**
		LLAMA 3.3	55.53%	−2.799
		MED42	54.77%	−2.745
	BIOASQ (summary)	MIXTRAL	55.27%	−3.296
		LLAMA 3.1	**57.01%**	−2.864
		LLAMA 3.3	56.61%	−2.892
		MED42	56.98%	**−2.843**

Table 27: Plausibility of cited passages (F_1-oriented BERTSCORE, BARTSCORE), zero-shot with/without RAG, and few-shot ($n = 5$) with RAG (BM25, $k = 5$). One retrieved passage per few-shot example. Measures applicable only to datasets that provide gold passages, with RAG enabled.

Setting	Dataset	LLM	Passage-level scores F_1 (↑)	MAP (↑)	LLM-judge (↑)
Zero-shot with RAG	PUBMEDQA	MIXTRAL	49.29%	85.59%	70.74%
		LLAMA 3.1	**51.27%**	**87.28%**	**80.17%**
		LLAMA 3.3	40.17%	84.65%	69.91%
		MED42	46.48%	86.73%	78.65%
	BIOASQ (yes/no)	MIXTRAL	14.69%	31.16%	72.40%
		LLAMA 3.1	**17.75%**	35.19%	**81.30%**
		LLAMA 3.3	16.71%	32.29%	72.67%
		MED42	17.07%	**36.75%**	70.61%
	BIOASQ (summary)	MIXTRAL	12.45%	**14.84%**	70.19%
		LLAMA 3.1	**13.64%**	14.81%	**78.51%**
		LLAMA 3.3	13.04%	12.71%	76.28%
		MED42	11.24%	12.19%	68.30%
	MEDMCQA	MIXTRAL	–	–	**40.16%**
		LLAMA 3.1	–	–	33.44%
		LLAMA 3.3	–	–	33.17%
		MED42	–	–	38.19%
	HEALTHFC	MIXTRAL	–	–	75.89%
		LLAMA 3.1	–	–	73.34%
		LLAMA 3.3	–	–	72.74%
		MED42	–	–	**78.16%**
Few-shot with RAG	PUBMEDQA	MIXTRAL	45.29%	86.79%	57.08%
		LLAMA 3.1	**51.79%**	**88.18%**	**84.80%**
		LLAMA 3.3	43.27%	85.43%	78.56%
		MED42	47.71%	86.12%	67.13%
	BIOASQ (yes/no)	MIXTRAL	14.85%	33.02%	58.06%
		LLAMA 3.1	**17.81%**	37.80%	83.81%
		LLAMA 3.3	16.87%	33.48%	**87.31%**
		MED42	15.55%	**39.78%**	62.52%
	BIOASQ (summary)	MIXTRAL	2.95%	4.01%	48.49%
		LLAMA 3.1	**13.63%**	14.89%	72.44%
		LLAMA 3.3	13.23%	13.52%	**79.55%**
		MED42	13.27%	**16.95%**	54.10%
	MEDMCQA	MIXTRAL	–	–	**34.98%**
		LLAMA 3.1	–	–	29.19%
		LLAMA 3.3	–	–	28.25%
		MED42	–	–	26.12%

Table 28: Plausibility of cited passages (passage-level F_1, MAP, LLM-judge scores), zero-shot with/without RAG, and few-shot ($n = 5$) with RAG (MEDCPT, $k = 5$). One retrieved passage per few-shot example. Passage-level measures applicable only to datasets that provide gold passages, with RAG enabled.

Setting	Dataset	LLM	F_1-oriented scores BERTSCORE (↑)	BARTSCORE (↑)
Zero-shot with RAG	PUBMEDQA	MIXTRAL	56.73%	−1.511
		LLAMA 3.1	**57.62%**	**−1.399**
		LLAMA 3.3	56.51%	−1.497
		MED42	56.77%	−1.473
	BIOASQ (yes/no)	MIXTRAL	55.66%	−3.208
		LLAMA 3.1	**55.85%**	**−3.030**
		LLAMA 3.3	55.55%	−3.197
		MED42	55.39%	−3.138
	BIOASQ (summary)	MIXTRAL	56.93%	−3.655
		LLAMA 3.1	**57.84%**	**−3.432**
		LLAMA 3.3	57.15%	−3.602
		MED42	56.76%	−3.657
Few-shot with RAG	PUBMEDQA	MIXTRAL	56.36%	−1.705
		LLAMA 3.1	**57.29%**	**−1.436**
		LLAMA 3.3	56.46%	−1.507
		MED42	56.81%	−1.586
	BIOASQ (yes/no)	MIXTRAL	55.68%	−3.202
		LLAMA 3.1	**55.94%**	**−2.998**
		LLAMA 3.3	55.56%	−3.190
		MED42	55.26%	−3.068
	BIOASQ (summary)	MIXTRAL	55.57%	−3.849
		LLAMA 3.1	**57.91%**	**−3.422**
		LLAMA 3.3	57.46%	−3.516
		MED42	57.44%	−3.607

Table 29: Plausibility of cited passages (F_1-oriented BERTSCORE, BARTSCORE), zero-shot with/without RAG, and few-shot ($n = 5$) with RAG (MEDCPT, $k = 5$). One retrieved passage per few-shot example. Measures applicable only to datasets that provide gold passages, with RAG enabled.

Setting	Dataset	LLM	Passage-level scores F_1 (↑)	MAP (↑)	LLM-judge (↑)
Zero-shot with RAG	PUBMEDQA	MIXTRAL	42.21%	70.04%	75.25%
		LLAMA 3.1	**47.17%**	**74.58%**	**82.12%**
		LLAMA 3.3	36.01%	71.32%	72.58%
		MED42	42.56%	73.93%	78.69%
	BIOASQ (yes/no)	MIXTRAL	18.91%	33.92%	64.56%
		LLAMA 3.1	**19.63%**	37.20%	**74.03%**
		LLAMA 3.3	18.85%	35.46%	65.3%
		MED42	19.49%	**39.95%**	66.01%
	BIOASQ (summary)	MIXTRAL	19.62%	30.04%	78.07%
		LLAMA 3.1	21.31%	**32.01%**	**87.52%**
		LLAMA 3.3	**21.55%**	31.22%	82.65%
		MED42	13.54%	20.30%	69.45%
	MEDMCQA	MIXTRAL	–	–	**39.08%**
		LLAMA 3.1	–	–	33.92%
		LLAMA 3.3	–	–	34.06%
		MED42	–	–	36.76%
	HEALTHFC	MIXTRAL	–	–	72.91%
		LLAMA 3.1	–	–	**77.01%**
		LLAMA 3.3	–	–	76.21%
		MED42	–	–	76.86%
Few-shot with RAG	PUBMEDQA	MIXTRAL	34.20%	76.52%	60.68%
		LLAMA 3.1	**40.26%**	**78.98%**	**85.89%**
		LLAMA 3.3	32.78%	77.14%	80.45%
		MED42	39.08%	78.96%	67.98%
	BIOASQ (yes/no)	MIXTRAL	20.03%	38.34%	61.39%
		LLAMA 3.1	**22.93%**	**41.71%**	**84.64%**
		LLAMA 3.3	22.24%	40.78%	84.43%
		MED42	12.66%	40.12%	64.69%
	BIOASQ (summary)	MIXTRAL	14.87%	35.56%	53.87%
		LLAMA 3.1	**26.60%**	37.45%	81.75%
		LLAMA 3.3	26.38%	36.49%	**82.23%**
		MED42	9.24%	**47.90%**	67.93%
	MEDMCQA	MIXTRAL	–	–	**35.04%**
		LLAMA 3.1	–	–	29.27%
		LLAMA 3.3	–	–	31.49%
		MED42	–	–	25.08%

Table 30: Plausibility of cited passages (passage-level F_1, MAP, LLM-judge scores), zero-shot with/without RAG, and few-shot ($n = 5$) with RAG (CONTRIEVER, $k = 5$). One retrieved passage per few-shot example. Passage-level measures applicable only to datasets that provide gold passages, with RAG enabled.

Setting	Dataset	LLM	F_1-oriented scores BERTSCORE (↑)	BARTSCORE (↑)
Zero-shot with RAG	PUBMEDQA	MIXTRAL	56.71%	−2.229
		LLAMA 3.1	**57.56%**	**−1.999**
		LLAMA 3.3	56.79%	−2.172
		MED42	56.88%	−2.135
	BIOASQ (yes/no)	MIXTRAL	56.22%	−3.253
		LLAMA 3.1	56.37%	**−3.226**
		LLAMA 3.3	**56.57%**	−3.285
		MED42	56.25%	−3.249
	BIOASQ (summary)	MIXTRAL	58.92%	−2.919
		LLAMA 3.1	58.96%	**−2.855**
		LLAMA 3.3	58.56%	−2.941
		MED42	**59.06%**	−2.882
Few-shot with RAG	PUBMEDQA	MIXTRAL	56.77%	−2.287
		LLAMA 3.1	**57.33%**	**−2.045**
		LLAMA 3.3	56.78%	−2.167
		MED42	56.74%	−2.224
	BIOASQ (yes/no)	MIXTRAL	56.15%	−3.270
		LLAMA 3.1	56.24%	**−3.198**
		LLAMA 3.3	**56.32%**	−3.267
		MED42	55.54%	−3.294
	BIOASQ (summary)	MIXTRAL	56.95%	−3.340
		LLAMA 3.1	**58.72%**	**−2.847**
		LLAMA 3.3	58.56%	−2.952
		MED42	58.25%	−2.899

Table 31: Plausibility of cited passages (F_1-oriented BERTSCORE, BARTSCORE), zero-shot with/without RAG, and few-shot ($n = 5$) with RAG (CONTRIEVER, $k = 5$). One retrieved passage per few-shot example. Measures applicable only to datasets that provide gold passages, with RAG enabled.

Regarding faithfulness of self-explanations and cited passages, for dense retrievers we reach the same conclusions as when using BM25, we show results for MEDCPT (Table 32) and CONTRIEVER (Table 33). The FG_{self} and FG_{ans} scores show that there is good alignment between the cited passages and the retrieved passages ranked by c_{self}, yet poor alignment between cited passages and retrieved passages ranked by c_{ans} (same in Table 6).

Dataset	LLM	ZERO-SHOT with RAG		
		PCC-SHAP (↓)	FG_{self} (↑)	FG_{ans} (↑)
PUBMEDQA	LLAMA 3.1	0.876	0.725	**0.998**
	LLAMA 3.3	0.914	**0.825**	0.997
	MED42	**0.758**	0.776	0.978
BIOASQ-QA (yes/no)	LLAMA 3.1	**0.690**	0.842	0.968
	LLAMA 3.3	0.922	**0.898**	**0.992**
	MED42	0.718	0.788	0.882
BIOASQ-QA (summary)	LLAMA 3.1	–	0.833	–
	LLAMA 3.3	–	**0.953**	–
	MED42	–	0.706	–
MEDMCQA	LLAMA 3.1	0.849	0.655	0.837
	LLAMA 3.3	0.879	**0.844**	**0.952**
	MED42	**0.765**	0.622	0.827
HEALTHFC	LLAMA 3.1	0.930	0.805	0.958
	LLAMA 3.3	0.860	**0.908**	**0.998**
	MED42	**0.833**	0.763	0.923

Table 32: Faithfulness of self-explanations (PCC-SHAP) and cited passages (FG_{self}, FG_{ans}). Zero-shot RAG, using MEDCPT, $k = 5$. Metrics applicable only with RAG enabled. Results with dense retrievers are similar (§G). The summary questions of BIOASQ do not have exact answers (§B.3); PCC-SHAP and FG_{ans} cannot be computed. No scores for MIXTRAL are provided for the same reasons as in Table 6.

Dataset	LLM	ZERO-SHOT with RAG		
		PCC-SHAP (↓)	FG_{self} (↑)	FG_{ans} (↑)
PUBMEDQA	LLAMA 3.1	0.955	0.652	**0.997**
	LLAMA 3.3	0.949	**0.780**	0.995
	MED42	**0.742**	0.732	0.973
BIOASQ-QA (yes/no)	LLAMA 3.1	0.879	0.805	**0.989**
	LLAMA 3.3	0.974	**0.857**	0.988
	MED42	**0.660**	0.750	0.936
BIOASQ-QA (summary)	LLAMA 3.1	–	0.932	–
	LLAMA 3.3	–	**0.990**	–
	MED42	–	0.723	–
MEDMCQA	LLAMA 3.1	0.846	0.665	0.906
	LLAMA 3.3	0.916	**0.797**	**0.957**
	MED42	**0.810**	0.637	0.862
HEALTHFC	LLAMA 3.1	0.961	0.780	0.969
	LLAMA 3.3	**0.911**	**0.876**	**0.995**
	MED42	0.915	0.753	0.934

Table 33: Faithfulness of self-explanations (PCC-SHAP) and cited passages (FG_{self}, FG_{ans}). Zero-shot RAG, using CONTRIEVER, $k = 5$. Metrics applicable only with RAG enabled. Results with dense retrievers are similar (§G). The summary questions of BIOASQ do not have exact answers (§B.3); PCC-SHAP and FG_{ans} cannot be computed. No scores for MIXTRAL are provided for the same reasons as in Table 6.

The results are also consistent for pertinence of gold retrieved passages when using MEDCPT (Table 34) or CONTRIEVER (Table 35). In PUBMEDQA, RAG-LLMs use gold passages to generate self-explanations (high CA_{self}), which is consistent with the high MAP scores observed when assessing the plausibility of cited passages (Tables 5, 25). On the contrary, their answers are less based on gold passages (low CA_{ans}), which is again consistent with the low PCC-SHAP scores. In the other datasets of BIOBENCH, the scores are lower.

Dataset	LLM	ZERO-SHOT with RAG	
		CA_{ans} (↑)	CA_{self} (↑)
PUBMEDQA	LLAMA 3.1	0.482	0.868
	LLAMA 3.3	0.451	**0.885**
	MED42	**0.553**	0.841
BIOASQ (yes/no)	LLAMA 3.1	0.297	0.392
	LLAMA 3.3	0.273	0.394
	MED42	**0.362**	**0.405**
BIOASQ (summary)	LLAMA 3.1	–	0.162
	LLAMA 3.3	–	**0.163**
	MED42	–	0.151

Table 34: Pertinence of retrieved passages. Zero-shot RAG, using MEDCPT, $k = 5$. Applicable only to datasets with gold passages, with RAG enabled. The summary questions of BIOASQ do not have exact answers (§B.3). No scores for MIXTRAL are provided for the same reasons as in Table 6.

Dataset	LLM	ZERO-SHOT with RAG	
		CA_{ans} (↑)	CA_{self} (↑)
PUBMEDQA	LLAMA 3.1	0.375	0.742
	LLAMA 3.3	0.411	**0.753**
	MED42	**0.475**	0.722
BIOASQ (yes/no)	LLAMA 3.1	0.297	0.392
	LLAMA 3.3	0.273	0.394
	MED42	**0.362**	**0.405**
BIOASQ (summary)	LLAMA 3.1	–	0.162
	LLAMA 3.3	–	**0.163**
	MED42	–	0.151

Table 35: Pertinence of retrieved passages. Zero-shot RAG, using CONTRIEVER, $k = 5$. Applicable only to datasets with gold passages, with RAG enabled. The summary questions of BIOASQ do not have exact answers (§B.3). No scores for MIXTRAL are provided for the same reasons as in Table 6.

We provide additional results[12] for all datasets of BIOBENCH, all combinations of retrievers and RAG-LLMs, including zero-shot and few-shot inference for many different configurations. We also provide a public leaderboard[13] and instructions for other researchers to share their results from additional RAG-LLMs with the community.

[12] https://1drv.ms/b/c/8f1b54d455091bb9/IQBIGo2c10pZRazE8-9Rl-WTAT7BKWUjwPs64OA4qjHMkos?e=OsgNnw

[13] The leaderboard URL will be released at the camera-ready.