

Overview of TREC 2024 Biomedical Generative Retrieval (BioGen) Track

Deepak Gupta¹, Dina Demner-Fushman¹, William Hersh², Steven Bedrick², and
Kirk Roberts³

¹National Library of Medicine, NIH

²Oregon Health & Science University

³University of Texas Houston

1 Overview

With the advancement of large language models (LLMs), the biomedical domain has seen significant progress and improvement in multiple tasks such as biomedical question answering, lay language summarization of the biomedical literature, clinical note summarization, etc. However, hallucinations or confabulations remain one of the key challenges when using LLMs in the biomedical and other domains. Inaccuracies may be particularly harmful in high-risk situations, such as medical question answering [1], making clinical decisions, or appraising biomedical research. Studies on evaluation of the LLMs abilities to ground generated statements in verifiable sources have shown that models perform significantly worse on lay-user generated questions [2], and often fail to reference relevant sources[3]. This can be problematic when those seeking information want evidence from studies to back up the claims from LLMs[4]. Unsupported statements are a major barrier to using LLMs in any applications that may affect health. Methods for grounding generated statements in reliable sources along with practical evaluation approaches are needed to overcome this barrier. Towards this, in our pilot task organized at TREC 2024, we introduced the task of reference attribution as a means to mitigate the generation of false statements by LLMs answering biomedical questions.

In TREC 2024, we received a total of 30 runs from 5 teams. The majority of the runs utilized a two-stage retrieval augmented generation (RAG) approach to generate the answers with references. In the first stage, a retriever was used for relevant literature, and in the second stage, the LLMs were used to generate the answers with appropriate references.

2 Task

Reference Attribution: Given a biomedical topic (question) and a stable version of PubMed documents, the task was to generate answers using Large Language Models (LLMs) or any other approaches. Each sentence¹ must be supported by up to three attributions (cited references), but no more than 30 documents per answer. Each document should be referenced in the answers as PMIDs in square brackets as shown in Figure 1. The document sentences that support the answer statements are shown to illustrate the manual evaluation process.

3 Topics

The first 40 BioGen 2024 topics were developed using information requests submitted by self-identified non-clinicians to the National Library of Medicine. Additionally, 25 topics were developed based on the collection² that contains questions based on a documents from Mayo Clinic³,

¹assertions/statements were approximated as sentences in the 2024 task

²<https://github.com/kevinwu23/SourceCheckup>

³<https://www.mayoclinic.org/diseases-conditions>

Figure 1: Sample reference answer

<p>Topic: <i>iron and ferritin levels in COVID-19</i></p> <p>Question: <i>Why is transferrin and iron low in covid patients but ferritin high?</i></p> <p>Narrative: <i>The patient is interested in the link between iron and infection, the role iron plays in infection and the implications for the COVID-19 course.</i></p> <p>Sample Answer [adapted to patient-level health literacy]: <i>During infections, a battle for iron takes place between the human body and the invading viruses [34389110]. The immune system cells need iron to defend the body against the infection [34389110]. The virus needs iron to reproduce [31585922]. If iron balance is disrupted by the infection, ferritin levels are high [34883281], which signals the disease is severe and may have unfavorable outcomes [34048587, 32681497]. Ferritin is maintaining the bodys iron level [18835072]. Some researchers believe that high levels of ferritin not only show the body struggles with infection, but that it might add to the severity of disease [34924800]. To help covid patients, the doctors may lower the ferritin levels that are too high using drugs that capture iron [32681497].</i></p> <p>References [to be returned by the system] and supporting statements manually extracted from the documents during the evaluation:</p> <ul style="list-style-type: none">• 34389110 A S1: During infections, a battle for iron takes place between the human host and the invading pathogens. A S2: Once primed by the contact with antigen presenting cells, lymphocytes need iron to sustain the metabolic burst required for mounting an effective cellular and humoral response.• 31585922 A S3: Viruses depend on iron in order to efficiently replicate within living host cells.• 34883281 A S4: Ferritin was initially described to accompany various acute infections, both viral and bacterial, indicating an acute response to inflammation.• 34048587 A S4: Elevated serum ferritin and IL-6 levels associated with increased mortality and with reduced mortality at ferritin levels <100 ng mL-1.• 32681497 A S4: Numerous studies have demonstrated the immunomodulatory effects of ferritin and its association with mortality and sustained inflammatory process.• 18835072 A S5: Ferritin, a major iron storage protein, is essential to iron homeostasis and is involved in a wide range of physiologic and pathologic processes.• 34924800 A S6: The inflammation cascade and poor prognosis of COVID-19 may be attributed to high ferritin levels.• 32681497 A S7: Iron chelation represents a pillar in the treatment of iron overload. In addition, it was proven to have an anti-viral and anti-fibrotic activity. Herein, we analyse the pathogenic role of ferritin and iron during SARS-CoV-2 infection and propose iron depletion therapy as a novel therapeutic approach in the COVID-19 pandemic.

UpToDate⁴, and Reddit r/AskDocs⁵. Example topics are shown in Table 1.

4 Data

The BioGen task used the latest annual baseline snapshot of Medline/PubMed, which goes approximately through the end of 2023. We provided a pre-processed set of 20,727,695 PMIDs representing the abstracts in the 2023 snapshot. The participants were asked to cite and use the PMIDs while generating the answers available in this collection.

⁴<https://www.uptodate.com/contents/table-of-contents/patient-education>

⁵<http://www.reddit.com/r/askdocs>

Topic	<i>natural treatments for sleep apnea</i>
Question	<i>Are there ways to prevent sleep apnea or treat it naturally?</i>
Narrative	<i>The patient is looking for natural remedies to prevent and treat sleep apnea</i>
Topic	<i>Drug treatment for COPD stage 4</i>
Question	<i>What drug or combination of drugs is most popular for treating stage 4 copd?</i>
Narrative	<i>This patient is looking for the most effective and popular medications for advanced Chronic obstructive pulmonary disease (COPD).</i>
Topic	<i>fbn1 mutation</i>
Question	<i>what is fbn1 mutation?</i>
Narrative	<i>A young woman who is planning to start a family has heard that one of her cousins had a mutation in the gene for fibrillin-1 (FBN1). She would like to know more about these mutations and the health problems that might be caused by the mutation</i>

Table 1: Example topics from the TREC 2024 BioGen track.

5 Participating Teams and Methods

5.1 Participating Teams

We used the NIST-provided Evalbase platform⁶ to release the datasets, registration, and submissions of the participating teams. In total, 5 teams participated in the BioGen track and submitted 30 individual runs for the task.

5.2 Methods

We summarize the approaches used by the participants as follows:

- **ur-iw**: The team used multiple LLMs (`gemini-1.5-flash-001`, `gpt4o-mini` etc.) for query expansion and utilized Elasticsearch to retrieve the relevant documents. After that, they extracted and reranked the snippets based on query relevance. The retrieved snippets were used to generate the answer using `gemini-1.5-flash-001` and `gpt4o-mini` models.
- **ii_research**: The team adopted BM25-based retrieval to generate a silver-standard dataset containing the query and relevant documents. The generated dataset was used to train a T5-based seq2seq ranking model. The trained model was used to rank the relevant documents against a question that was used to generate the answer using the `GPT-4o` model.
- **h2oloo**: In the first stage, the team used BM25 with Rocchio query expansion to rank the top 1000 documents against a query. After that, they used a multi-stage reranker to retrieve the top 20 relevant documents/snippets that were used to generate an answer using `Llama3.1` and `GPT-4o` models.
- **webis**: The team used retrieval followed by reranking to retrieve relevant documents; thereafter `GPT-4o` was used to generate the answers.
- **ielab**: The team followed a three-step approach where the relevant documents were retrieved and re-ranked based on the query and then LLMs were used to generate the answer. In the final stage, LLMs attributed each of the generated sentences to the list of retrieved documents.

6 Assessment

For the BioGen task, we focused on **(a)** reference attribution and **(b)** the quality and factuality of the text generated by LLMs to answer clinical questions asked by clinicians to **(i)** satisfy their own information needs or **(ii)** answer health-related questions asked by their patients. For patients, we envisioned that the answers would be reviewed by clinicians and subsequently explained in

⁶<https://ir.nist.gov/evalbase>

plain language. This evaluation aimed to verify the references attributed and the quality of LLM-generated answers. For the former, we evaluated how well the answers to clinical questions are supported by evidence provided by the models in the form of references. Here are the details of the two-part evaluation for the submissions (also shown in Figure 2):

- **Part 1: Evaluating Answer Alignment with Questions and Answer Quality and Completeness:** In the first step of the evaluation, we evaluated whether the generated text, taken as a whole, directly answered the question. We then evaluated the relevance of the assertions in the answer sentences to the question. Each assertion in the generated answer was labeled with one of the following four labels:
 - **Required:** The assertion XXX is necessary to have in the generated answer for completeness of the answers.
 - **Unnecessary:** The assertion XXX is not required to have included in the generated answer. An assertion may be unnecessary for several reasons: it provides general information on the topic; it comments on the lack of information in provided documents; it recommends to see a doctor, while the task stated the patient has already contacted the provider, or the provider is asking the question.
 - **Borderline:** If an assertion is relevant, possibly even good to know, but not required, the assertion may be marked borderline. For example, if the question is about most commonly used treatments, information about treatments in the early stages of clinical trials is not necessary.
 - **Inappropriate:** The assertion may harm the patient, e.g., if according to the answer, physical therapy reduces the pain level, but the patient experiences more pain due to hip mobilization, the patient may start doubting they are receiving adequate treatment.
- **Part 2: Evaluating Answer Alignment with Evidence Support:**

In the second step, for each generated answer sentence, we assessed the referenced document(s) to determine each document’s relation to the generated assertion, if any. We labeled each cited document with one of the four possible relations between the answer sentence and the document: ‘*Supports*’, ‘*Contradicts*’, ‘*Neutral*’, and ‘*Not Relevant*’.

- **Supports:** There is at least one sentence in the referenced document that supports/agrees with the statement, e.g.: “*opioids were the mainstay of perioperative pain control*”. In addition, no other sentence in the document contradicts the statement.
- **Contradicts:** There is at least one sentence in the referenced document that disagrees with the assertion or states its opposite, e.g.: “*Increasing pain levels after the first week postoperatively, for 3 days, are most likely to be caused by the change to more extensive mobilization and physiotherapy in the rehabilitation unit.*” (The answer in this case stated that the pain decreases steadily after the surgery.)
- **Neutral:** The referenced document is topically relevant, but lacks any information to validate or invalidate the assertion.
- **Not relevant:** The referenced document is not relevant to the sentence.

7 Evaluation

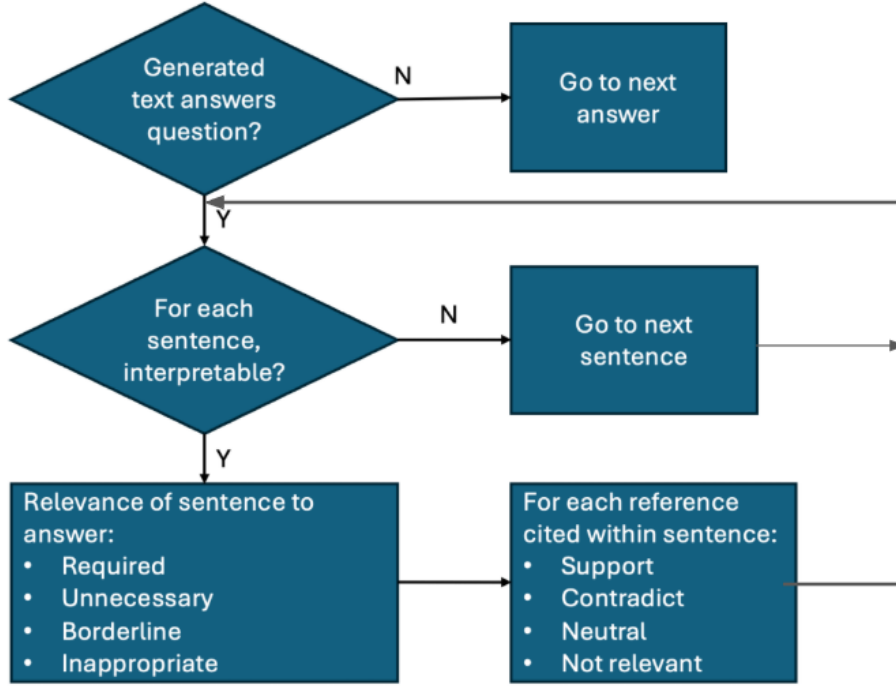
We carried out an evaluation of approaches at several levels and along several axes.

7.1 Answer quality

- **Answer Accuracy** (measured at the run level) – measures how many of the answers to the total of 65 questions were deemed acceptable (judged as answering the question at least partially) for each run.

$$\text{Accuracy}_{run} = \frac{\text{Number of Acceptable Answers}}{\text{Total Number of Questions (Topics)}} \quad (1)$$

Figure 2: Evaluation steps



- **Answer Completeness (Recall)** (answer level) – measures how many of the answer aspects (pooled across all submitted runs from all participating teams) are covered in one answer to a question. For the initial evaluation, we group/cluster the sentences using the embeddings from SentenceTransformer⁷ and SimCSE⁸ models. In a strict evaluation, only sentences that were judged required and supported by evidence will be considered. For a lenient evaluation, all sentences judged required will be considered. For a relaxed evaluation, the borderline sentences will be considered in addition to the required sentences. The number of aspects for the automated grouping is set to 10 using K-means clustering.

$$\text{Completeness}_{\text{answer}} = \frac{\text{Number of Distinct Clusters Containing Sentences from Answer A}}{\text{Number of Clusters}} \quad (2)$$

- **Answer Precision** (answer level) – measures how many of the assertions in the answer were judged required or acceptable. The strict, lenient, and relaxed evaluations also apply.

$$\text{Precision}_{\text{answer}} = \frac{\text{Number of Generated Required Sentences}}{\text{Total Number of Generated Sentences}} \quad (3)$$

- **Redundancy Score** (answer level): penalizes a system for generating *unnecessary* answer sentences. It measures the informativeness of the generated answers.

$$\text{Redundancy Score} = \frac{\text{Number of Generated Unnecessary Sentences}}{\text{Total Number of Generated Sentences}} \quad (4)$$

- **Irrelevant Score** (answer level): penalizes a system for generating *inappropriate/potentially harmful* answer sentences.

$$\text{Harmfulness Score} = \frac{\text{Number of Generated Harmful Sentences}}{\text{Total Number of Generated Sentences}} \quad (5)$$

⁷ sentence-transformers/all-mpnet-base-v2

⁸ princeton-nlp/sup-simcse-roberta-large

7.2 Citation Quality

Answer statements may be supported or contradicted by the documents provided as references. Sentences may not include any references or include references that are only topically relevant or not relevant. The following metric capture the quality of references:

- **Citation Coverage:** measures how well the required and borderline generated answer sentences are backed by the appropriate (judged as supports) citations.

$$\text{Citation Coverage} = \frac{\text{Number of Systems Generated Answer Sentences with One or More Supportive Citation}}{\text{Total Number of Generated Answer Sentences}} \quad (6)$$

- **Citation Support Rate:** assesses how well the system-predicted citations are aligned with the human-judged support citations.

$$\text{Citation Support Rate} = \frac{\text{Number of } \textit{Supports} \text{ Citations}}{\text{Total Number of Citations}} \quad (7)$$

- **Citation Contradict Rate:** penalizes the answers that are providing documents assessed as *Contradicting* the statement. Note that in a fact-verification task, this measure may show how well a system is finding contradictory evidence.

$$\text{Citation Contradict Rate} = \frac{\text{Number of } \textit{Contradict} \text{ Citations}}{\text{Total Number of Citations}} \quad (8)$$

7.3 Document relevancy

Pooling all documents judged relevant for a given topic, we can compute standard recall and precision. Note that relevant documents include documents judged as supporting, contradicting or being neutral.

- Recall:

$$\text{Recall} = \frac{\text{Number of relevant retrieved documents}}{\text{all relevant documents}} \quad (9)$$

- Precision:

$$\text{Precision} = \frac{\text{Number of relevant retrieved documents}}{\text{Number of references provided}} \quad (10)$$

We computed each aforementioned metric for each question and then averaged over all the questions in the test collection to obtain the final scores.

8 Results and Discussion

We evaluated the performance of the submitted runs on multiple levels for answer quality (Tables 2, 3, 4), citation quality (Table 5) and document relevance (Table 6). For the answer accuracy (deemed acceptable for a given question), we found most of the runs achieved more than 90% accuracy. For precision of the answer quality, the run `norarr.llm_only_2.llama3-8b_fixed` achieved a maximum score of 90.68%, while the run `webis-2` (where the answer was generated by the `Mistral-7B-Instruct-v0.3` model) recorded the lowest precision of the 52.82. For redundancy, we found the run `rarr_atronly.llm_only_2.llama3-70b_fixed` recorded the lowest redundancy score of 3.51. The run used `Llama-3-70B-Instruct` model to generate the answer. The `listgalore_gpt-4o_arenuggetsallyouneed` run obtained the highest recall in multiple settings (S+R, R, R+B). The best citation coverage of 92.21% was obtained by `norarr.llm_only_2.llama3-70b_fixed` run, in which `Llama-3-70B-Instruct` was asked to attribute each of the generated answer sentences explicitly. The best citation support rate (CSR) of 79.97% was achieved by `listgalore_gpt-4o_ragnarokv5biogen_top20`. The best recall and precision scores of 24.12 and 90.04 for document relevancy were obtained by the `listgalore_gpt-4o_arenuggetsallyouneed` and `listgalore_l31-70b_ragnarokv5biogen_top20` runs.

Team Name	Run Name	Acceptable Answers	Accuracy
h2oloo	listgalore_gpt-4o_ragnarokv5biogen_top20	65	100
	listgalore_l31-70b_ragnarokv5biogen_top20	65	100
	rl31-70b_l31-70b_ragnarokv5biogen_top20	65	100
	rl31-70b_gpt-4o_ragnarokv5biogen_top20	65	100
	rl31-70b_gpt-4o_ragnarokv5biogennc_top20	65	100
	listgalore_gpt-4o_arenuggetsallyouneed.json	65	100
ielab	norarr.llm_only_2.llama3-70b_fixed	63	96.92
	rarr_attronly.llm_only_2.llama3-70b_fixed	61	93.85
	rarr_qgenfix.llm_only_2.llama3-70b_fixed	62	95.38
	rarr_attrfix.llm_only_2.llama3-70b_fixed	64	98.46
	rarr_qgen.llm_only_2.llama3-70b_fixed	62	95.38
	rarr_attrfix_custprompt.llm_only_3_v2.llama3-70b_fixed	63	96.92
	rarr_qgenfix_custprompt.llm_only_3_v2.llama3-70b_fixed	64	98.46
	rarr_qgenfix.llm_only_2.llama3-8b_fixed	62	95.38
	norarr.llm_only_2.boolena.llama3-70b_fixed	58	89.23
norarr.llm_only_2.llama3-8b_fixed	64	98.46	
ii_research	iiresearch_trec_bio2024_t5base_run	65	100
ur-iw	ten-shot-gpt4o-mini	65	100
	ten-shot-gemini-flash	62	95.38
	ten-shot-gpt4o-mini-wiki	61	93.85
	ten-shot-gemini-flash-wiki	61	93.85
	zero-shot-gpt4o-mini	64	98.46
	zero-shot-gemini-flash	56	86.15
webis	webis-1	53	81.54
	webis-2	43	66.15
	webis-3	39	60
	webis-5	43	66.15
	webis-gpt-4	54	83.08
	webis-gpt-6	60	92.31
	webis-gpt-1	65	100

Table 2: Performance comparison of the submitted runs for the answer quality in terms of *Accuracy* metric.

Team Name	Run Name	Precision	Redundancy	Harmfulness
h2oloo	listgalore_gpt-4o_ragnarokv5biogen_top20	80.5	14.07	0
	listgalore_l31-70b_ragnarokv5biogen_top20	83.53	12.8	0.38
	rl31-70b_l31-70b_ragnarokv5biogen_top20	82.08	14.37	0.26
	rl31-70b_gpt-4o_ragnarokv5biogen_top20	82.87	12.32	0
	rl31-70b_gpt-4o_ragnarokv5biogennc_top20	80.9	16.31	0
	listgalore_gpt-4o_arenuggetsallyouneed.json	82.15	12.97	0.93
ielab	norarr.llm_only_2.llama3-70b_fixed	89	4.03	0
	rarr_attronly.llm_only_2.llama3-70b_fixed	84.31	3.51	0
	rarr_qgenfix.llm_only_2.llama3-70b_fixed	83.28	8.74	0.38
	rarr_attrfix.llm_only_2.llama3-70b_fixed	87.31	6.95	0.82
	rarr_qgen.llm_only_2.llama3-70b_fixed	87.51	5.18	0
	rarr_attrfix_custprompt.llm_only_3_v2.llama3-70b_fixed	87.54	4.49	0.77
	rarr_qgenfix_custprompt.llm_only_3_v2.llama3-70b_fixed	90.54	4.28	0
	rarr_qgenfix.llm_only_2.llama3-8b_fixed	84.7	6.6	1
	norarr.llm_only_2.boolena.llama3-70b_fixed	80.28	5.21	0.38
norarr.llm_only_2.llama3-8b_fixed	90.68	4.99	0.31	
ii_research	iiresearch_trec_bio2024_t5base_run	69.44	23.41	1.54
ur-iw	ten-shot-gpt4o-mini	83.93	12.9	0
	ten-shot-gemini-flash	75.88	15.21	0.48
	ten-shot-gpt4o-mini-wiki	75.62	15.37	0
	ten-shot-gemini-flash-wiki	74.07	17.13	0
	zero-shot-gpt4o-mini	81.98	12.59	0.22
	zero-shot-gemini-flash	68.4	13.09	0
webis	webis-1	65.54	11.9	0
	webis-2	52.82	12.05	0
	webis-3	52.79	6.51	0
	webis-5	55.59	8.64	0
	webis-gpt-4	64.23	16.54	0
	webis-gpt-6	62.56	26.41	0
	webis-gpt-1	81.67	14.74	0

Table 3: Evaluation of submitted runs for answer quality, focusing on *Precision*, *Redundancy*, and *Harmfulness* metrics.

Team Name	Run Name	Recall (S+R)	Recall (R)	Recall (R+B)
h2oloo	listgalore_gpt-4o_ragnarokv5biogen_top20	34.31	38.15	40.15
	listgalore_l31-70b_ragnarokv5biogen_top20	34	38.46	41.08
	rl31-70b_l31-70b_ragnarokv5biogen_top20	34.15	38.15	38
	rl31-70b_gpt-4o_ragnarokv5biogen_top20	32.92	39.38	41.69
	rl31-70b_gpt-4o_ragnarokv5biogennc_top20	0	40.31	41.23
	listgalore_gpt-4o_arenuggetsallyouneed.json	43.08	46.31	46.15
ielab	norarr.llm_only_2.llama3-70b_fixed	26	26.62	27.69
	rarr_attrfix_custprompt.llm_only_3_v2.llama3-70b_fixed	25.23	26	27.69
	rarr_qgenfix_custprompt.llm_only_3_v2.llama3-70b_fixed	23.38	26.31	27.38
	rarr_atronly.llm_only_2.llama3-70b_fixed	24.15	24.62	27.23
	rarr_attrfix.llm_only_2.llama3-70b_fixed	25.69	26	27.69
	rarr_qgenfix.llm_only_2.llama3-8b_fixed	23.08	26.31	27.23
	rarr_qgenfix.llm_only_2.llama3-70b_fixed	23.38	26	26.46
	rarr_qgen.llm_only_2.llama3-70b_fixed	22.46	24	25.08
	norarr.llm_only_2.boolena.llama3-70b_fixed	21.69	22.77	23.54
	norarr.llm_only_2.llama3-8b_fixed	24.31	27.08	27.69
ii_research	iiresearch_trec_bio2024_t5base_run	1.38	25.08	26.15
ur-iw	ten-shot-gpt4o-mini	28.77	38.15	39.54
	ten-shot-gemini-flash	27.54	33.54	32.31
	ten-shot-gemini-flash-wiki	24.77	32.77	32.92
	zero-shot-gpt4o-mini	26.77	37.85	39.54
	zero-shot-gemini-flash	25.08	32.15	33.54
	ten-shot-gpt4o-mini-wiki	25.08	34.92	35.85
webis	webis-1	7.38	9.69	11.08
	webis-5	3.23	8.15	8.62
	webis-gpt-6	7.08	16.31	17.08
	webis-gpt-1	0	20.15	20.15
	webis-2	4.77	7.69	8.31
	webis-3	5.85	7.85	8.31
	webis-gpt-4	7.08	14	14.77

Table 4: Comparison of submitted runs based on answer quality using the Recall metric. The abbreviations are as follows: **(S+R)**– only answer sentences that were judged required and supported by evidence were considered to cluster, **(R)**– answer sentences that were judged required were considered to cluster, **(R+B)**– answer sentences that were judged either required or borderline were considered to cluster.

Team Name	Run Name	Citation Coverage	Citation Support Rate	Citation Contradict Rate
h2oloo	listgalore_gpt-4o_ragnarokv5biogen_top20	82.79	79.97	1.76
	listgalore_l31-70b_ragnarokv5biogen_top20	78.47	77.02	2.43
	rl31-70b_l31-70b_ragnarokv5biogen_top20	81.14	77.41	2.88
	rl31-70b_gpt-4o_ragnarokv5biogen_top20	80.32	78	2.49
	rl31-70b_gpt-4o_ragnarokv5biogennc_top20	0	0	0
	listgalore_gpt-4o_arenuggetsallyouneed.json	88.58	62.54	1.93
ielab	norarr.llm_only_2.llama3-70b_fixed	92.21	73.62	1.89
	rarr_atronly.llm_only_2.llama3-70b_fixed	86.54	70.84	4.69
	rarr_qgenfix.llm_only_2.llama3-70b_fixed	85.77	65.01	1.93
	rarr_attrfix.llm_only_2.llama3-70b_fixed	89.05	73.22	6.13
	rarr_qgen.llm_only_2.llama3-70b_fixed	83.1	67.55	0.97
	rarr_attrfix_custprompt.llm_only_3_v2.llama3-70b_fixed	88.56	71.98	3.53
	rarr_qgenfix_custprompt.llm_only_3_v2.llama3-70b_fixed	87.54	67.45	3.16
	rarr_qgenfix.llm_only_2.llama3-8b_fixed	81.74	68	2.12
	norarr.llm_only_2.boolena.llama3-70b_fixed	81.36	60	1.55
	norarr.llm_only_2.llama3-8b_fixed	88.18	55.16	2.45
ii_research	iiresearch_trec_bio2024_t5base_run	5	11.54	0
ur-iw	ten-shot-gpt4o-mini	66.35	68.63	1.88
	ten-shot-gemini-flash	72.94	64.02	0.26
	ten-shot-gpt4o-mini-wiki	60.61	65.77	0.19
	ten-shot-gemini-flash-wiki	65.89	56.59	1.26
	zero-shot-gpt4o-mini	64.13	60.56	1.55
	zero-shot-gemini-flash	63.06	58.13	1.35
webis	webis-1	52.51	41.56	3.82
	webis-2	34.1	41.79	4.62
	webis-3	42.97	45	5.38
	webis-5	29.36	34.62	0.77
	webis-gpt-4	42.18	68.72	2.31
	webis-gpt-6	32.31	37.51	3.64
	webis-gpt-1	0	0	0

Table 5: Comparison of submitted runs based on citation quality using multiple metrics.

Team Name	Run Name	Recall	Precision
h2oloo	listgalore_gpt-4o_ragnarokv5biogen_top20	12.91	88.03
	listgalore_l31-70b_ragnarokv5biogen_top20	14.31	90.04
	rl31-70b_l31-70b_ragnarokv5biogen_top20	14.23	89.21
	rl31-70b_gpt-4o_ragnarokv5biogen_top20	12.63	88.64
	rl31-70b_gpt-4o_ragnarokv5biogennc_top20	0	0
	listgalore_gpt-4o_arenuggetsallyouneed.json	24.12	77.08
ielab	norarr.llm_only_2.llama3-70b_fixed	9.36	83.64
	rarr_attronly.llm_only_2.llama3-70b_fixed	10.37	81.75
	rarr_qgenfix.llm_only_2.llama3-70b_fixed	9.16	76.93
	rarr_attrfix.llm_only_2.llama3-70b_fixed	11.29	86.59
	rarr_qgen.llm_only_2.llama3-70b_fixed	8.63	77.98
	rarr_attrfix_custprompt.llm_only_3_v2.llama3-70b_fixed	10.56	81.34
	rarr_qgenfix_custprompt.llm_only_3_v2.llama3-70b_fixed	9.26	76.47
	rarr_qgenfix.llm_only_2.llama3-8b_fixed	7.25	80.79
	norarr.llm_only_2.boolena.llama3-70b_fixed	8.52	74.22
norarr.llm_only_2.llama3-8b_fixed	9.01	72.52	
ii_research	iiresearch_trec_bio2024_t5base_run	0.6	22.82
ur-iw	ten-shot-gpt4o-mini	9.06	73.62
	ten-shot-gemini-flash	8.89	69.23
	ten-shot-gpt4o-mini-wiki	8.4	69.72
	ten-shot-gemini-flash-wiki	8.73	60.22
	zero-shot-gpt4o-mini	8.29	70.64
	zero-shot-gemini-flash	6.51	61.22
webis	webis-1	3.03	50.77
	webis-2	1.15	52.31
	webis-3	1.16	46.92
	webis-5	0.96	40.51
	webis-gpt-4	2.28	77.69
	webis-gpt-6	2.93	54.62
	webis-gpt-1	0	0

Table 6: Evaluation of submitted runs for document relevancy focusing on *Precision* and *Recall* metrics.

9 Conclusion

This overview of the TREC 2024 BioGen track discussed the tasks, datasets, evaluation metrics, participating systems and their performance. We evaluated the performance of the submitted runs at multiple levels (answer, citation and documents) using the traditional metrics. Most of the teams used the two-step approach where they first retrieved the documents with lexical search method (BM25), followed by reranker to obtain the top-k relevant documents/snippets. In the second stage, LLMs were used to generate the answer citing appropriate documents. We hope that introducing the task has created ground truth datasets for fostering research toward designing systems that generate answers to health-related questions grounded with appropriate citations, which, in turn, provide a trusted and reliable source to support the generated assertions in the answer.

Acknowledgments

This research was supported in part by the Division of Intramural Research (DIR) of the National Library of Medicine (NLM), National Institutes of Health. Bill Hersh, Kate Fultz Hollis, and Kirk Roberts were supported by grant R01LM011934 from the National Library of Medicine. The authors thank Kate Fultz Hollis, OHSU and Srishti Kapur, Centaur Labs for managing the manual evaluation process at OHSU and Centaur labs, respectively. This work utilized the computational resources of the NIH HPC Biowulf cluster (<https://hpc.nih.gov>).

References

- [1] Wenting Zhao, Zhongfen Deng, Shweta Yadav, and Philip S Yu. Heterogeneous knowledge grounding for medical question answering with retrieval augmented large language model. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1590–1594, 2024.
- [2] Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi Riantawan, Daniel E Ho, and James Zou. How well do llms cite relevant medical references? an evaluation framework and analyses. *arXiv e-prints*, pages arXiv–2402, 2024.
- [3] Bojana Bašaragin, Adela Ljajić, Darija Medvecki, Lorenzo Cassano, Miloš Košprdić, and Nikola Milošević. How do you know that? teaching generative language models to reference answers to biomedical questions. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii, editors, *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 536–547, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [4] William Hersh. Search still matters: information retrieval in the era of generative AI. *Journal of the American Medical Informatics Association*, 31(9):2159–2161, 01 2024.