Evaluating LLM-as-a-Judge for Biomedical Question Answering in BioASQ

Abstract—**This paper examines whether large language models (LLM)-as-a-judge methods can be used as an alternative to the expert in evaluating the ideal answers for the BioASQ Task 13B. This study examines an Evidence Prime-style biomedical question answer (QA) pipeline, compares ROUGE scores with official BioASQ ROUGE scores, and evaluates multiple LLM judge configurations against official manual scores for recall, precision, repetition, and readability. The results show strong agreement between generated and official ROUGE scores, supporting the evaluation pipeline. LLM judges aligned most strongly with manual scores for recall, while precision, repetition, and readability showed weaker or more variable agreement. Prompt design, including the use of score descriptors and reference answers, substantially affected judge behaviours. Overall, LLM-as-a-judge appears useful as a scalable supplement to expert biomedical QA evaluation, but not as a replacement for human assessment.**

Keywords—Large language models, LLM-as-a-judge, biomedical question answering, Bio-ASQ, automated evaluation, manual scoring, ROUGE

I. Introduction

Biomedical QA has attracted substantial interest because of its potential to improve access to complex medical knowledge. An established benchmark in this area is the BioASQ challenge, in which participating systems are given biomedical questions and relevant content and must generate answers to those questions, which are then scored using a variety of metrics based on yes/no, factoid, list and summary answer types [1]. Each question also contains an ideal answer. Ideal answers are multi-sentence responses, and BioASQ reports both automated ROUGE-based scores and expert assessment scores for this answer type. This makes BioASQ especially useful as an evaluation setting for this study because it explicitly recognizes that answer quality cannot be captured by lexical overlap alone and provides motivation to develop more automated methods.

As biomedical QA systems continue to improve, evaluating answer quality remains a central challenge. In particular, assessing long-form generated answers is difficult because answer quality depends not only on factual overlap with a reference, but also on qualities such as completeness, clarity, conciseness, and redundancy [1]. BioASQ has noted that stronger ROUGE performance does not necessarily correspond to stronger manual evaluation scores [2]. This creates an important evaluation problem. While manual scoring provides a richer assessment of answer quality, it is costly, time-consuming, and difficult to scale.

Recent works have suggested that LLM-as-a-judge methods may offer a more flexible alternative to traditional automated metrics [3]. Rather than relying only on word overlap, LLM judges can potentially assess broader qualities such as relevance, completeness, readability, and repetition. However, their usefulness for evaluating biomedical QA answers, particularly in the BioASQ setting, remains underexplored. Questions remain regarding judge prompt design, LLM model choice, and the use of LLMs specifically trained using biomedical text. Within the biomedical domain, it is also unclear how sensitive LLM judges are to the use of score descriptors, which provide brief guidance on how score levels should be interpreted, or reference answers.

This paper examines the use of the LLM-as-a-judge approach to assess how well judge scores correlate with expert assessment scores. It investigates smaller open-weight models as LLM judges, how score descriptors and reference answers affect the judge model's correlation with expert-aggregated scores, and whether small open models trained on biomedical text offer any advantage. These questions are assessed over four evaluation dimensions: readability, recall, precision, and repetition.

II. Related Work

A. BioASQ

Prior work has treated BioASQ's ideal answers as a form of biomedical query-focused summarization, in which systems generate paragraph-length responses from retrieved biomedical snippets [4], [5]. Studies in this area have explored extractive summarization, ontology-based retrieval, neural ranking methods and transformer-based approaches for producing ideal answers from BioASQ snippets [6], [7]. This prior work is important because it frames ideal-answer generation more than merely returning the correct fact; the generated response must include relevant biomedical content, avoid unsupported information, minimize repetition, and remain readable.

More recently, BioASQ has also been used to study automatic evaluation methods for biomedical summaries [8]. Hijazi et al. (2024) examined whether LLM-based evaluation methods could approximate human expert scores for BioASQ human expert evaluators. Their work is closely related to this

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

study because it directly investigates LLM-based evaluation in the BioASQ ideal-answer setting.

However, this study differs from Hijazi et al. in several key respects. Unlike prior work that compared generated answers against human-per-question evaluations, this study examines correlation at the aggregated system and batch level. This reflects a common external research setting in which detailed expert assessments are not available, but official leaderboard scores are published. As a result, the aim of this study is to explore whether LLM judges can reflect broad expert trends across BioASQ systems and batches.

This study also examines smaller, more deployable judge models than those commonly associated with proprietary LLM evaluation, including Qwen-based judges. In particular, it tests whether judge alignment improves when the prompt includes reference answers and explicit score descriptors for the expert assessment scores. This allows the study to evaluate not only if LLM judges align with BioASQ aggregate expert outcomes, but also whether smaller judge models can be made more effective through prompt design.

B. *LLM-as-a-judge*

LLM-based evaluation is a broader category than the specific LLM-as-a-Judge setup as examined in this study. For example, prior work has used LLMs through prompted scoring and related evaluation methods [8]. In contrast, our study focuses on direct LLM-as-a-judge scoring, in which the model assigns scores to generated BioASQ ideal answers based on the manual evaluation dimensions.

LLM-as-a-judge has emerged as a promising approach for automated evaluation in natural language generation tasks. Instead of relying only on surface-level text overlap, these methods use large language models to assess qualities such as relevance, completeness, coherence, and redundancy. Earlier work showed that strong proprietary models, such as GPT-4, could align reasonably well with human preferences in open-ended conversational evaluation, while also revealing limitations, including position bias, verbosity bias, and self-enhancement bias [3]. That line of research helps explain why LLM-as-a-judge became attractive as an evaluation method. At the same time, much of the early evidence comes from very large frontier models in relatively general settings, rather than from domain-specific tasks such as biomedical QA.

That distinction matters for this study. In a practical biomedical QA setting, it is not always realistic to assume access to the strongest available judge models because of cost, access, or deployment constraints. For that reason, it is still worthwhile to study smaller or more practical judges, even if they are not expected to match the performance of the largest models. In this sense, the question is not only whether LLM-as-a-judge can work under ideal conditions, but also how well it performs under more realistic constraints and how sensitive it is to prompt design choices.

III. Study Objectives and Methodological Rationale

This study evaluates the extent to which LLM judges align with official BioASQ Task 13B expert assessment scores on ideal-answers across the four reported evaluation dimensions: information recall, information precision, information repetition, and readability. In addition, it examines how judge alignment is affected by prompt design, including the use of score descriptors and reference answers.

Prompt design is treated as a methodological variable in this study, as LLM scores may depend on how the evaluation task is presented to the model. Specifically, the study examines the effect of including or omitting score descriptors and reference answers in the judge's prompt.

This focus is supported by Yamauchi, Yano, and Oyamada, who examine how prompt design affects LLM correlation to human expert scores [9]. Their study examines the role of reference answers and score descriptors in defining the meaning of each score level in the evaluation metric. Their findings show that the explicit evaluation criteria are important for reliable judging, and removing such criteria can reduce alignment with human judges.

Their work provides the methodological basis for comparing judge prompts with and without score descriptors, as well as with and without reference answers. The purpose of this comparison is to determine whether these prompt designs affect the alignment of the judge scores with human expert assessment scores.

IV. Materials and Competition System

BioASQ evaluates both retrieval and answer generation. Task 13A focuses on retrieving relevant biomedical snippets, while Task 13B focuses on generating answers using the provided question data and supporting evidence. In Task 13B, competitor systems processed four batches of 85 questions, for a total of 340 questions. For each question, the systems generated ideal answers, which were manually scored by biomedical experts. Because this study examines the quality of generated answers, it focuses on the Task 13B ideal-answer component. The retrieved snippets are relevant only to the extent that they are included in the Task 13B gold batch files and are used for answer generation.

This study uses BioASQ Task 13B ideal-answer data from four golden batches [2]. For the official comparison, the analysis relies on BioASQ's batch-level expert assessment scores for information recall, information precision, information repetition, and readability, as well as the official ROUGE scores. Since the official expert assessment scores are reported at the batch/competitor-entry level rather than at the individual-question level, the judge-human comparison is conducted across matched system-entry observations across batches.

Regarding the scores themselves, information recall measures whether the answer includes the necessary biomedical content. Information precision measures whether the included content is relevant and avoids unsupported or irrelevant material. Information repetition measures whether the answer avoids unnecessary redundancy. Readability measures clarity and fluency. Higher scores reflect better answer quality, with a scale of 1 to 5 for each score.

These expert assessment scores provide the human-evaluation baseline for the study. By comparing aggregated LLM judge scores with the official BioASQ manual scores, this

study examines whether LLM judges merely produce plausible ratings or align with the official human assessment of ideal-answer quality.

V. Methods

Evidence Prime [10], a competition team whose work was submitted to BioASQ Task 13B, was selected as a comparison system because its published methodology details its answer-generation process, making it suitable for a controlled judge-alignment study. The section then describes the judge models, prompt conditions, manual-score comparison procedure, and the ROUGE analysis used to validate the pipeline and examine whether LLM-as-a-judge scoring can approximate expert human evaluation.

A. Evidence Prime

Evidence Prime's entries were designed to compare smaller open-weight LLMs with large proprietary models for biomedical question answering. Because their paper describes the methods used to produce the EP submissions, including snippet ranking, prompting, and model and ensemble selection, the EP systems provide useful reference points for this study. The EP-1, EP-2, EP-3, EP-4, and EP-5 entries, therefore, provide a basis for assessing whether LLM-as-a-judge scores align with the official BioASQ Task 13B expert assessment scores.

In each BioASQ batch, Evidence Prime submitted multiple system entries labelled EP-1, EP-2, EP-3, EP-4 and EP-5. The EP labels are batch-specific system labels rather than fixed model identities across all batches. Therefore, the EP labels should be interpreted as the submitted system configurations for each batch, not as a single, unchanged method repeated across all four batches.

In creating answers from each batch of questions, the Evidence Prime answer generation pipeline involves three steps in prompt formation and answer generation/selection: system and user prompts, snippet selection, and model selection. Evidence Prime has a system and user prompt for each answer generated that incorporates a set system and user template and snippets from the gold answer batches [10]. For each question, the ten most relevant snippets are selected using sentence-transformer embeddings and cosine similarity. The selected snippets are then ordered from most to least similar to the question and provided to the model using the Evidence Prime ideal-answer prompt template.

The fundamental difference between Evidence Primes submissions per batch is the LLMs used for answer generation. Table 1 is from the Evidence Prince paper that shows each batch/submission's models [10]. For submissions that included multiple LLM models, an ensemble selection process was used. This process followed the Evidence Prime method to identify the ideal answers. Each candidate's answer was scored against the corresponding question using a reranking model, and the candidate with the highest question-answer similarity score was selected as the final prediction. The ensemble process, therefore, did not combine, average, or rewrite candidate answers. It selected one of the generated answers as the final system output.

TABLE 1.

LLM MODELS PER BATCH SUBMISSION

Batch	EP-1	EP-2	EP-3	EP-4	EP-5
1	P4	H8	Q14	O4	C35
2	G12, Q14, P4, O4, C35	C35	P4	P4	Q14
3	M24, G12, G27, Q14, P4	O4, O41, C35	O41	P4	—
4	O41, O4, C35, C37	G12, Q14, MP14, P4	Q14, MP14, P4, O41, O4, C35, C37	Q14, P4, O41, O4, C37	O41

P4 = Phi-4; H8 = HuatuoGPT-o1 8B; Q14 = Qwen2.5-14B; G12 = Gemma-3-12B; G27 = Gemma-3-27B; M24 = Mistral-Small-3.1-24B; O4 = GPT-4o; O41 = GPT-4.1; C35 = Claude 3.5 Sonnet; C37 = Claude 3.7 Sonnet; MP14 = Meditron3-Phi4-14B. Batch 3 did not include an EP-5 submission.

The list contains a few noteworthy items. Batch 3 does not include an EP-5 submission. In Batch 2, EP-3 and EP-4 use the same base model, but EP-4 adds a prompt-generation method using DSPy (Declarative Self-improving Python) for factoid questions. Because this modification is reported as applying only to factoid questions, it does not affect the ideal-answer generation process examined in this study. The Batch 2 EP-4 entry is therefore included in the official Evidence Prime system definitions for completeness. Across four batches, the Evidence Prime submissions produce 19 matched batch-entry observations.

This study reconstructs the Evidence Prime prompts, snippet-ranking method, model assignments, and ensemble-selection procedure described [10]. This step was important because the purpose of the study was not simply to evaluate arbitrarily generated answers. Instead, the goal was to evaluate whether LLM-as-a-Judge methods could assess ideal answers produced under a methodology based on a competitive BioASQ Task 13B system. By executing the Evidence Prime generation process, the judge scores could be compared against official BioASQ expert assessment scores in a setting that reflects the original task structure.

B. Judges

The LLM judge models examined in this study are GPT-4.1 [11], Meditron3/Phi-4-14B [12], and Qwen3.5-32B [13]. Each judge was evaluated under four prompt conditions created by varying two prompt-design factors: whether score descriptors were included and whether a reference answer was provided.

The score descriptors provided explicit guidance for the four BioASQ expert assessment dimensions: information recall, information precision, information repetition, and readability. The reference-answer condition provided the judge with an example of a high-quality answer for the same question. These four prompt conditions allowed the study to examine whether rubric detail and gold-answer guidance improved alignment

between LLM judge scores and official BioASQ expert assessment scores using Pearson and Spearman correlation. The design of these prompt conditions was informed by prior work on the reliability of LLM-as-a-judge evaluation [9], which shows that judges' performance can be affected by prompt and evaluation design choices.

C. ROUGE Analysis

The judge scores were computed from Evidence Prime pipeline outputs and compared with official Evidence Prime batch-entry manual scores. Therefore, the ROUGE analysis was used as a validation check to ensure that the local Evidence Prime pipeline outputs were sufficiently aligned with the official submissions for aggregate comparison.

Strong agreement would suggest that differences between LLM-judge scores and expert assessment scores were less likely to be due to data-processing or evaluation errors. Instead, such differences would more likely reflect the substantive limits of LLM-judge alignment with human expert evaluation.

VI. Results

The results are organized around the central question of whether LLM-as-a-judge scores can approximate the official BioASQ expert manual evaluation for ideal answers. The selection begins with a ROUGE check to confirm that the Evidence Prime pipeline is broadly consistent with the official BioASQ scoring process. It then examines the correlation between LLM-judge scores and the official manual scores, followed by the analysis of how prompt-design choices affect judge alignment.

A. ROUGE results

The Evidence Prime evaluation pipeline was compared against the official BioASQ Task 13B ROUGE results for ideal answers. Across the 19 matched entry/batch observations, locally computed ROUGE scores showed strong agreement with the official BioASQ ROUGE scores according to Table 2. Pearson correlations ranged from 0.791 to 0.886, while Spearman correlations ranged from 0.735 to 0.850 across ROUGE-2 recall, ROUGE-2 F1, ROUGE-SU4 recall, and ROUGE-SU4 F1.

TABLE 2

ROUGE PEARSON SPEARMAN CORRELATIONS

Metric	Pearson	Spearman
ROUGE-2 Recall	0.886	0.850
ROUGE-2 F1	0.822	0.759
ROUGE-SU4 Recall	0.874	0.815
ROUGE-SU4 F1	0.791	0.735

Although the scores were not expected to be numerically identical to the official BioASQ scores, the strong correlations observed across ROUGE metrics suggest that the Evidence Prime evaluation pipeline was duplicated with sufficient similarity for aggregate comparative analysis. This supports its use for subsequent comparisons between LLM judge scores and BioASQ expert assessment scores.

B. LLM judge alignment with official expert assessment scores

Table 3 reports Pearson correlations between LLM judge scores and official BioASQ manual scores across the four BioASQ manual dimensions, while Table 4 reports the corresponding Spearman correlations. Alignment between LLM judge scores and official expert scores varied substantially by metric, according to Table 3. Recall showed the strongest and most consistent agreement, with the best judge configurations producing moderate-to-strong correlations with official scores. Precision also showed meaningful alignment for several configurations, although the results were generally weaker than recall.

TABLE 3

PEARSON CORRELATION PER LLM JUDGE CONFIGURATION

Judge	Rec.	Prec.	Rep.	Read.
Q32-DR	0.741	0.549	0.585	0.646
G41-DR	0.698	0.517	0.671	0.590
G41-R	0.694	0.506	0.658	0.572
Q32-R	0.663	0.446	0.612	0.519
M14-DR	0.549	0.347	0.552	0.492
G41-D	0.433	0.476	0.666	0.591
Q32-D	0.416	0.355	0.570	0.601
G41	0.396	0.322	0.708	0.608
M14-R	0.328	-0.005	0.542	0.330
M14	0.175	-0.039	0.562	0.151
MI4-D	0.048	-0.122	0.505	0.509
Q32	-0.017	-0.048	0.632	0.488

Note. Rec. = recall; Prec. = precision; Rep. = repetition; Read. = readability. Q32 = Qwen3.5-32B; G41 = GPT-4.1; M14 = Meditron3/Phi-4-14B; D = score descriptors; R = reference answer. No suffix indicates that neither score descriptors nor reference answers were used.

TABLE 4

SPEARMAN CORRELATION PER LLM JUDGE CONFIGURATION

Judge	Rec.	Prec.	Rep.	Read.
Q32-DR	0.730	0.637	0.463	0.380
G41-DR	0.668	0.605	0.614	0.259
G41-R	0.760	0.574	0.604	0.023
Q32-R	0.610	0.652	0.585	0.145
M14-DR	0.628	0.502	0.245	-0.139
G41-D	0.614	0.546	0.617	0.132
Q32-D	0.467	0.407	0.347	0.165
G41	0.500	0.378	0.625	0.150
M14-R	0.547	0.171	0.405	0.160
M14	0.389	-0.029	0.419	-0.053
MI4-D	0.183	-0.263	0.064	-0.96

Q32	0.174	0.148	0.563	0.043

Note. Rec. = recall; Prec. = precision; Rep. = repetition; Read. = readability. Q32 = Qwen3.5-32B; G41 = GPT-4.1; M14 = Meditron3/Phi-4-14B; D = score descriptors; R = reference answer. No suffix indicates that neither score descriptors nor reference answers were used.

Repetition produced moderate correlations across multiple judges, while readability was the least reliable dimension, particularly when rank-order agreement was considered. Overall, these results suggest that LLM-as-a-judge evaluation shows promise for BioASQ ideal-answer assessment, but its reliability depends strongly on the specific manual evaluation dimension being measured.

One notable result is that the biomedical-oriented Meditron3/Phi-4-14B judge did not outperform the more general-purpose judges. Under its strongest prompt condition, with both score descriptors and a reference answer, Meditron3/Phi-4-14B showed moderate alignment for recall but weaker alignment across the remaining dimensions, particularly readability. This suggests that biomedical domain specialization alone is not sufficient to produce reliable LLM-as-a-judge performance. In this setting, judge reliability appeared to depend more on the specific model and prompt configuration than on whether the model was explicitly adapted for biomedical text.

C. Prompt design effects

Prompt design had a substantial, non-uniform effect on judges' alignment with the BioASQ manual scores. The strongest overall correlations were obtained when reference answers were included, particularly for Qwen-32B, which included both score descriptors and reference answers and was the best-performing configuration overall. In contrast, as shown in Table 3, Qwen-32B, without score descriptors or reference answers, was the weakest-performing configuration, exhibiting near-zero or negative correlations for recall and precision. GPT-4.1 also showed consistently strong recall and repetition alignment under reference-based conditions, while removing score descriptors sometimes improved individual metrics, such as GPT-4.1 recall Spearman correlation.

These effects were not consistent across dimensions or judges. Withholding the reference answer generally reduced alignment for recall and precision, with the largest declines observed for Qwen-32B and Meditron/Phi-4 configurations. Repetition was less dependent on the reference answer and remained moderately correlated across several prompt settings, while readability correlations were weaker and more variable, especially for Spearman rank correlation. Overall, the results indicate that prompt design should be treated as a central part of LLM-as-a-judge evaluation.

VII. Discussion

The results of this study suggest that LLM-as-a-judge evaluation can provide a useful approximation of expert manual scoring for BioASQ ideal answers. The strongest alignment with official manual scores was observed for information recall, while precision and repetition showed moderate alignment, and readability was less reliable, particularly in terms of rank-order agreement. This pattern suggests that LLM judges may be better suited to evaluating content coverage than more subjective dimensions of answer quality. In the BioASQ setting, this is an important distinction because the quality of an ideal answer depends not only on its fluency but also on whether it includes the relevant biomedical information needed to answer the question.

The results also show that prompt design is a factor in judge reliability. The inclusion of reference answers generally improved alignment for content-oriented dimensions such as recall and precision, although the effect of score descriptors was less uniform across judges and metrics. This indicates that the LLM-as-a-judge evaluation should not be treated individually. Instead, the judge model, rubric wording, score descriptors, and reference-answer availability should all be considered part of the evaluation design. For biomedical QA, where small differences in content may affect answer quality, this sensitivity to prompt design is especially important.

The ROUGE analysis provides useful support for the validity of the experimental local pipeline. The locally generated Evidence Prime pipeline generated ROUGE scores that showed a strong correlation with the official BioASQ ROUGE scores, suggesting that the Evidence Prime-style answer generation process was sufficiently accurate for aggregate comparison. However, the judge-score results also reinforce the limitation of ROUGE as an evaluation method for long-form biomedical answers. ROUGE captures lexical overlap with reference answers, but it does not necessarily capture broader qualities assessed in manual evaluation, such as completeness, precision, readability, and non-redundancy.

Overall, these findings support the use of LLM-as-a-judge methods as a scalable supplement to expert evaluation in biomedical QA, especially when reference answers and clear scoring criteria are available. However, the variability across dimensions and prompt conditions indicates that LLM judges should not be treated as direct replacements for expert assessors. Their value is strongest when used to support comparative analysis, identify broad quality patterns, and reduce the evaluation burden, while final claims about the quality of biomedical answers should remain grounded in expert evaluation whenever possible.

Future work should extend this analysis across additional BioASQ competitions, additional judge models, and repeated judging runs to measure judge stability. Further work should also examine whether judge ensembles improve reliability and whether answer-level human annotations can be collected for a subset of questions. Such annotations would enable testing whether per-answer judge scores align with expert judgments rather than only with aggregate system-level scores.

VIII. Conclusion

This study examined whether LLM-as-a-judge methods can approximate BioASQ manual evaluation for ideal biomedical question-answering responses. The results suggest that LLM judges can provide useful signals, especially for information recall, where alignment with official manual scores was strongest. Precision and repetition also showed some meaningful agreement, while readability was less reliable. The study also found that prompt design matters. Reference answers and score descriptors changed judge behaviour, showing that LLM-as-a-judge evaluation depends not only on the model used,

but also on how the judging task is framed.Overall, the findings suggest that LLM-as-a-judge is a promising supplement to expert evaluation for biomedical QA, but not yet a replacement.

Future work should extend the evaluation to additional BioASQ competitions, evaluate models, and run repeated judging runs, with particular attention to whether judge ensembles can improve the reliability of LLMs as judges. In addition, if a BioASQ entry could be designed and submitted to a future BioASQ competition, and the corresponding per-answer expert judge scores could be accessed, the correlation analysis could be more accurate and detailed.

This study has several limitations. First, the comparison between LLM judge scores and official BioASQ expert assessment scores was based on a relatively small set of matched entry/batch observations. As a result, the reported Pearson and Spearman correlations should be interpreted cautiously. They provide useful preliminary evidence about judge alignment, but they should not be treated as stable estimates of general LLM judge reliability across all biomedical QA settings.

Second, the official BioASQ expert assessment scores used in this study are available at the aggregate batch-entry level rather than at the individual-answer level. This means that the main judge-human comparison could evaluate whether LLM judges reflected system-level or competitor-level patterns, but it could not directly test whether judge scores aligned with expert judgments for each individual generated answer.

REFERENCES

[1] G. Tsatsaronis *et al.*, "BioASQ: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering," in *2012 AAAI Fall Symposium Series*, 2012. [Online]. Available: sites/default/files/PublicDocuments/2012-Tsatsaronis-BioASQ.pdf

[2] BioASQ, "BioASQ Task 13b Phase B Results." Accessed: Apr. 24, 2026. [Online]. Available: https://participants-area.bioasq.org/results/13b/phaseB/

[3] L. Zheng *et al.*, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Curran Associates, Inc., 2023, pp. 46595–46623. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf

[4] D. Mollá, U. Khanna, D. Galat, V. Nguyen, and M. Rybinski, "Query-focused extractive summarisation for finding ideal answers to biomedical and COVID-19 questions," *arXiv preprint arXiv:2108.12189*, 2021.

[5] D. Mollá, "Macquarie University at BioASQ 5b – Query-based Summarisation Techniques for Selecting the Ideal Answers," in *Proceedings of the 16th BioNLP Workshop*, K. B. Cohen, D. Demner-Fushman, S. Ananiadou, and J. Tsujii, Eds., Vancouver, Canada,: Association for Computational Linguistics, Aug. 2017, pp. 67–75. doi: 10.18653/v1/W17-2308.

[6] A. Naresh Kumar *et al.*, "Ontology-Based Retrieval & Neural Approaches for BioASQ Ideal Answer Generation," in *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, I. A. Kakadiaris, G. Paliouras, and A. Krithara, Eds., Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 79–89. doi: 10.18653/v1/W18-5310.

[7] G. Wiese, D. Weissenborn, and M. Neves, "Neural Question Answering at BioASQ 5B," in *Proceedings of the 16th BioNLP Workshop*, K. B. Cohen, D. Demner-Fushman, S. Ananiadou, and J. Tsujii, Eds., Vancouver, Canada,: Association for Computational Linguistics, Aug. 2017, pp. 76–79. doi: 10.18653/v1/W17-2309.

[8] H. Hijazi, D. Molla, V. Nguyen, and S. Karimi, "Using Large Language Models to Evaluate Biomedical Query-Focused Summarisation," in *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, and J. Tsujii, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 236–242. doi: 10.18653/v1/2024.bionlp-1.18.

[9] Y. Yamauchi, T. Yano, and M. Oyamada, "An Empirical Study of LLM-as-a-Judge: How Design Choices Impact Evaluation Reliability," Jun. 2025, [Online]. Available: http://arxiv.org/abs/2506.13639

[10] D. Stachura, J. Konieczna, and A. Nowak, "Are Smaller Open-Weight LLMs Closing the Gap to Proprietary Models for Biomedical Question Answering?," 2025. [Online]. Available: https://arxiv.org/abs/2509.18843

[11] "OpenAI GPT-4.1." Accessed: Apr. 28, 2026. [Online]. Available: https://platform.openai.com/docs/models/gpt-4.1

[12] "OpenMeditron Meditron3-Phi4-14B." Accessed: Apr. 28, 2026. [Online]. Available: https://huggingface.co/OpenMeditron/Meditron3-Phi4-14B

[13] "Qwen Qwen3-32B." Accessed: Apr. 28, 2026. [Online]. Available: https://huggingface.co/Qwen/Qwen3-32B