Towards Unified Factuality Evaluation for Biomedical QA and Summarization: Aligning Metrics with Clinical Use-Cases

Mahule Roy[1] **Subhas Roy**[2]
[1]University of Oxford [2]TATA Consumer Products Limited
Correspondence: mroy25@bwh.harvard.edu

Abstract

Large language models achieve strong performance on biomedical question answering and summarization benchmarks, yet traditional evaluation metrics often fail to detect clinically significant factual errors. We introduce a unified evaluation framework that combines reference-based measures with evidence-grounded factuality verification to assess biomedical text generation. Evaluating four open-source models across three benchmarks (BioASQ, PubMedQA, MedLFQA), we find that 13.4–24.7% of generated claims are contradicted and 23–41% are unsupported, despite high lexical overlap scores. Our proposed Fact-Aligned Score (FAS) correlates strongly with claim-level verifiability ($\rho = 0.68$), substantially outperforming ROUGE-L ($\rho = 0.41$). We release an open-source toolkit with model outputs and analysis scripts to support reproducible factuality evaluation and safer deployment of biomedical LLMs.

1 Introduction

Large language models have shown remarkable capabilities in generating biomedical text for question answering, summarization, and clinical documentation. These models promise to assist clinicians, researchers, and patients by providing rapid access to structured and unstructured biomedical knowledge. However, current evaluation practices predominantly rely on lexical overlap metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and F1, which fail to capture clinically significant factual errors. A model may produce fluent and reference-aligned text while introducing subtle contradictions, unsupported claims, or outdated recommendations—errors that could have tangible consequences in healthcare settings. This gap creates an urgent need for evaluation frameworks that assess not only linguistic quality but also factual reliability and evidence alignment. In this work, we propose a unified evaluation framework that integrates traditional reference-based metrics with evidence-grounded factuality verification. Our framework systematically extracts atomic claims using spaCy (Honnibal et al., 2020), retrieves supporting biomedical literature from PubMed (Cock et al., 2009; Jin et al., 2019; Liu et al., 2005), and applies domain-adapted natural language inference (NLI) via BioBERT-NLI trained on MedNLI (Shivade, 2017; Gu et al., 2021) to verify each claim. By combining these components, we enable a more granular and clinically meaningful assessment of LLM outputs. This approach allows researchers and practitioners to identify factual inconsistencies, quantify reliability, and compare models not just on fluency, but on evidence-supported correctness, ultimately advancing the safe deployment of LLMs in biomedical contexts.

2 Related Work

Benchmark datasets such as BioASQ Task B (10th edition) (Nentidis et al., 2023), PubMedQA (Jin et al., 2019), and MedLFQA (Jeong et al., 2024) provide structured evaluation for biomedical QA across factoid, list, and long-form questions. BioASQ includes 492 test questions with 1–3 expert references, PubMedQA contains yes/no/maybe questions derived from PubMed abstracts, and MedLFQA provides 200 long-form QA pairs with clinician-verified claim-level annotations. While these resources are central to biomedical NLP evaluation, conventional metrics (e.g., ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019)) primarily capture lexical overlap and often overestimate performance by failing to detect unsupported or contradicted claims. Prior efforts have explored evidence-based and NLI-driven evaluation, including biomedical NLI models trained on MedNLI (Shivade, 2017). General-domain factuality metrics such as FactCC and QAFactEval

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 15–19

focus on reference consistency within news or open-domain summarization, rather than external biomedical evidence alignment. In contrast, our framework integrates reference-based metrics with claim-level verification against retrieved PubMed abstracts, enabling detection of unsupported and contradicted scientific assertions even when reference answers are incomplete. By combining systematic multi-model evaluation with an open-source, reproducible toolkit, our approach advances clinically aligned factuality assessment for biomedical text generation.

3 Unified Evaluation Framework

3.1 Reference-based Accuracy

Reference-based accuracy measures the lexical and semantic alignment between generated and reference answers using ROUGE-1/2/L (F1), BLEU-4 (smoothing method 1), and BERTScore (DeBERTa-large-mnli, batch size 64, no baseline rescaling) via HuggingFace `evaluate`. For multi-reference datasets such as BioASQ, we adopt a maximum-score strategy following standard practice in BioASQ evaluations. These metrics provide a baseline measure of answer quality while capturing both n-gram overlap and contextual semantic similarity, forming the reference foundation for our composite factuality assessment.

3.2 Evidence-Grounded Factuality and Safety Assessment

Generated answers are decomposed into atomic claims using spaCy (Honnibal et al., 2020) and queried against PubMed (2014–2024) via Entrez API, retrieving the top 5 abstracts per claim. Each claim–abstract pair is evaluated with BioBERT-NLI (Shivade, 2017) (dmis-lab/biobert-large-nli-v1.1) using a 0.7 threshold to classify claims as supported, contradicted, or unsupported. Answer-level aggregation computes Support@5, Contra@5, and unsupported proportions, averaged per answer. Additionally, rule-based safety checks using RxNorm (Liu et al., 2005) and a curated FDA contraindication database (U.S. Food and Drug Administration, 2023) flag missing dosages or contraindicated medications. These safety flags are reported separately from the Fact-Aligned Score (FAS), a composite metric defined as $\text{FAS} = 0.3 \cdot \text{BERTScore} + 0.4 \cdot \text{Support@5} + 0.3 \cdot (1 - \text{Contra@5})$, which balances reference alignment, evidence support, and contradiction avoidance to provide a clinically meaningful measure of output reliability.

4 Experimental Setup

We evaluate three open-source generative models—Llama-2-13B-chat (Touvron et al., 2023), Llama-2-7B-chat (Touvron et al., 2023), and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)—all of which are pre-trained and instruction-tuned but not further fine-tuned on the benchmark datasets. We also include PubMedBERT-base (Gu et al., 2021) as an encoder-only baseline. Evaluation is performed on three biomedical QA benchmarks: BioASQ Task B (10th edition; 492 test questions: 157 factoid, 138 list, 197 summary; average 1.4 references per question; mean reference length 87.3 tokens, SD 42.1) (Nentidis et al., 2023), PubMedQA (500 instances evenly split across yes/no/maybe; mean 42.6 tokens, SD 18.3) (Jin et al., 2019), and MedLFQA v1.0 (200-instance evaluation subset; long-form clinical QA pairs; claim-level factuality labels derived from the dataset's Must Have (MH) and Nice to Have (NH) annotations; mean 156.2 tokens, SD 67.8) (Jeong et al., 2024). Models were accessed via HuggingFace Transformers (v4.36.0) using the following identifiers: meta-llama/Llama-2-13b-chat-hf, meta-llama/Llama-2-7b-chat-hf, mistralai/Mistral-7B-Instruct-v0.2, and microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract (all accessed June 2024). Generation was performed with deterministic decoding (max_new_tokens=512, temperature=0, top_p=1, do_sample=False) in FP16 precision; PubMedBERT, being encoder-only, uses the source abstract as the answer for PubMedQA. Experiments were conducted on NVIDIA A100-40GB GPUs (156 total GPU-hours). Traditional metrics were computed using HuggingFace `evaluate` (v0.4.0): ROUGE-1/2/L (Lin, 2004) with Porter stemming and Penn Treebank tokenization (Bird et al., 2009), BLEU-4 (Papineni et al., 2002) with Chen and Cherry smoothing method 1, and BERTScore (Zhang et al., 2019) using `microsoft/deberta-large-mnli` (batch size 64). Factuality evaluation employed a modular pipeline with spaCy-based claim extraction (Honnibal et al., 2020), PubMed retrieval (2014–2024; top five abstracts per claim via Bio.Entrez), and NLI inference (batch size 16; 512-token truncation). Statistical significance was assessed using answer-level bootstrap resampling (10,000 iterations) (Efron and Tibshirani, 1986),

reporting two-tailed 95% confidence intervals; performance differences were considered significant when intervals excluded zero. Safety checks for clinical hazards leveraged the FDA contraindication database (U.S. Food and Drug Administration, 2023) and RxNorm (Liu et al., 2005).

For generation, we used the following prompt templates:

QA tasks (BioASQ factoid/list, PubMedQA):

```
Answer concisely and accurately based on
medical knowledge.
Question: {question}
Answer:
```

Summarization tasks (BioASQ summary, MedLFQA):

```
Provide a comprehensive summary based on
biomedical literature. Include findings,
mechanisms, and clinical implications.
Question: {question}
Summary:
```

5 Results

5.1 Traditional Metric Performance

All generative models evaluated (Llama-2-13B-chat, Llama-2-7B-chat, Mistral-7B-Instruct-v0.2) are pre-trained and instruction-tuned; no further fine-tuning was applied on the benchmark datasets. PubMedBERT is included as an encoder-only baseline. Table 1 reports reference-based performance across datasets. Llama-2-13B-chat achieves the highest ROUGE-L scores on all three benchmarks, with statistically significant improvements over Llama-2-7B on BioASQ and PubMedQA, though not on MedLFQA due to higher variance in long-form responses. Performance scales with model size, while Mistral-7B performs comparably to Llama-2-7B with no significant differences. PubMedBERT reveals a limitation of lexical metrics: on PubMedQA, it achieves ROUGE-L 0.412 despite not generating answers but simply reproducing the source abstract. This indicates that reference-based metrics reward lexical overlap even when outputs are not faithful answers to the question, motivating the need for evidence-grounded evaluation.

5.2 Evidence-Grounded Factuality

Table 2 presents Support@5 and Contra@5 scores. Although larger models demonstrate higher support rates, contradiction rates remain non-trivial (13.4%–24.7%), even for the strongest model. This confirms that high ROUGE does not imply factual reliability. Notably, PubMedBERT's relatively high ROUGE on PubMedQA (0.412) corresponds to lower support rates on BioASQ, illustrating dissociation between lexical similarity and claim-level verification.

Table 1: ROUGE-L scores (95% CI).

Model	BioASQ	PubMedQA	MedLFQA
Llama-2-13B	0.387±0.013	0.501±0.012	0.362±0.016
Llama-2-7B	0.358±0.013	0.473±0.012	0.334±0.016
Mistral-7B	0.365±0.013	0.481±0.012	0.341±0.016
PubMedBERT	0.187±0.013	0.412±0.012	0.143±0.016

Table 2: Factuality metrics (Support@5 and Contra@5) across BioASQ and PubMedQA. MedLFQA factuality is evaluated via the Fact-Aligned Score (FAS; Table 3). 95% CI shown.

Model	BioASQ S@5	BioASQ C@5	PubMedQA S@5
Llama-2-13B	71.2% [67.9-74.5]	13.4% [11.2-15.6]	78.4% [75.7-81.1]
Llama-2-7B	63.5% [60.2-66.8]	17.1% [14.9-19.3]	71.2% [68.5-73.9]
Mistral-7B	65.8% [62.5-69.1]	16.3% [14.1-18.5]	72.5% [69.8-75.2]
PubMedBERT	42.3% [39.0-45.6]	24.7% [22.5-26.9]	63.7% [61.0-66.4]

5.3 Per-Question-Type Analysis and Correlation with Gold Factuality

Factuality on BioASQ varies by question type: for Llama-2-13B, factoid questions (N=157) achieve the highest Support@5 (76.8%) and ROUGE-L (0.412) but also the highest contradiction rate (16.2%); list questions (N=138) show intermediate support (73.4%) and contradictions (14.1%), while summary questions (N=197) have lower support (65.2%) and fewer contradictions (10.8%). Differences between factoid and summary questions are statistically significant (Support@5 $\Delta = 11.6\%$, 95% CI [7.2, 15.9]; Contra@5 $\Delta = 5.4\%$, 95% CI [2.1, 8.7]). On MedLFQA, the Fact-Aligned Score (FAS) correlates most strongly with gold claim-level annotations ($\rho = 0.68$), outperforming ROUGE-L ($\rho = 0.41$) and BLEU-4 ($\rho = 0.36$); Support@5 alone also shows strong correlation ($\rho = 0.61$). Claim-level labels were derived from expert "Must Have" (MH) and "Nice to Have" (NH) statements, mapping MH to supported claims and treating unsupported or contradictory statements as factuality violations.

5.4 External Validation Considerations

Our primary validation relies on MedLFQA, from which we derive expert-informed claim-level fac-

tuality labels. While this enables direct quantitative correlation analysis, we acknowledge that additional human evaluation on BioASQ and PubMedQA would further strengthen external validity. To partially assess generalization, we analyzed consistency patterns across datasets. The relative ranking of models under FAS remained stable across BioASQ, PubMedQA, and MedLFQA, suggesting that the metric captures systematic reliability differences rather than dataset-specific artifacts. Table 3 shows the consistent ranking of models across all three datasets under our proposed FAS metric, with Llama-2-13B consistently outperforming other models and PubMedBERT serving as a lower-bound baseline. All evaluated models are pre-trained and instruction-tuned; no dataset-specific fine-tuning was applied. PubMedBERT is included as an encoder-only baseline.

Table 3: FAS scores (95% CI) across datasets. Higher is better. For MedLFQA, claim-level factuality labels were derived heuristically from expert-annotated "Must Have" (MH) and "Nice to Have" (NH) statements, mapping MH to supported claims and treating unsupported or contradictory statements as violations.

Model	BioASQ	PubMedQA	MedLFQA
Llama-2-13B	**0.612±0.013**	**0.703±0.012**	**0.584±0.016**
Mistral-7B	0.578±0.013	0.672±0.012	0.553±0.016
Llama-2-7B	0.571±0.013	0.665±0.012	0.547±0.016
PubMedBERT	0.412±0.013	0.589±0.012	0.398±0.016

5.5 Ablation Studies

We conduct ablation experiments on MedLFQA to assess the contribution of individual components of the FAS pipeline. Table 4 shows Spearman correlation with gold factuality under different configurations. Removing literature retrieval ("w/o retrieval") tests the importance of external evidence, resulting in a decrease from $\rho = 0.68$ to $\rho = 0.52$ ($\Delta = -0.16$). Replacing BioBERT-NLI with a general-domain MNLI model ("w/ general MNLI") evaluates the effect of domain adaptation, lowering correlation to $\rho = 0.45$ ($\Delta = -0.23$). Limiting retrieval to a single abstract per claim ("Single abstract") examines the impact of multi-document grounding, reducing correlation to $\rho = 0.56$ ($\Delta = -0.12$). These results suggest that both multi-document retrieval and domain-specific NLI significantly contribute to factuality alignment.

The NLI confidence threshold determines the minimum probability at which a claim–abstract pair is classified as supported or contradicted.

Table 4: Ablation on MedLFQA: Spearman ρ measures rank correlation between the Fact-Aligned Score (FAS) and gold claim-level factuality annotations. 95% CI shown; Δ indicates change relative to the full model.

Configuration	ρ	95% CI	Δ
Full model (FAS)	0.68	[0.62-0.74]	—
w/o retrieval	0.52	[0.45-0.59]	-0.16
w/ general MNLI	0.45	[0.38-0.52]	-0.23
Single abstract	0.56	[0.49-0.63]	-0.12

Claims with model confidence below the threshold are labeled as unsupported. To assess sensitivity, we varied this threshold from 0.5 to 0.9. Table 5 shows that Spearman correlation with gold factuality remains stable between 0.6–0.8, with optimal performance at 0.7, though differences within this range are not statistically significant. Based on these results, we selected 0.7 as the default threshold for our main analysis.

Table 5: Threshold sensitivity analysis on MedLFQA. Correlation with gold factuality at different NLI confidence thresholds.

Threshold	Spearman ρ [95% CI]
0.5	0.63 [0.56-0.70]
0.6	0.66 [0.59-0.73]
0.7	0.68 [0.62-0.74]
0.8	0.65 [0.58-0.72]
0.9	0.59 [0.52-0.66]

6 Conclusion

Open-source biomedical language models make 13–25% unsupported claims, with 4–7% flagged for missing dosages. Lexical metrics correlate weakly with factuality ($\rho = 0.36$–0.49), while the Fact-Aligned Score (FAS) aligns better ($\rho = 0.68$). Our framework—combining references, claim extraction, retrieval, and NLI—assesses biomedical text generation. Limitations include extraction errors, retrieval gaps, NLI mismatch, heuristic settings, English-only evidence (2014–2024), and multi-reference metric inflation.

References

Nentidis, A., Katsimpras, G., Krithara, A., Lima López, S., Farré-Maduell, E., Gasco, L., ... & Paliouras, G. (2023, September). Overview of bioasq 2023: The eleventh bioasq challenge on large-scale biomedical

semantic indexing and question answering. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 227-250). Cham: Springer Nature Switzerland.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W., & Lu, X. (2019, November). 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) (pp. 2567-2577).

Jeong, M., Hwang, H., Yoon, C., Lee, T., & Kang, J. (2024). Olaph: Improving factuality in biomedical long-form question answering. arXiv preprint arXiv:2405.12701.

Shivade, C. (2017). Mednli-a natural language inference dataset for the clinical domain.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Jiang, Y., Li, X., Zhu, G., Li, H., Deng, J., Han, K., ... & Zhang, R. (2023). 6G non-terrestrial networks enabled low-altitude economy: Opportunities and challenges. arXiv preprint arXiv:2311.09047.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), 1-23.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science, 54-75.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python.

Liu, S., Ma, W., Moore, R., Ganesan, V., & Nelson, S. (2005). RxNorm: prescription for electronic drug information exchange. IT professional, 7(5), 17-23.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

FDA, U. (2023). FDA adverse event reporting system (FAERS) public dashboard.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... & De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422.