Healthcare QA Evaluation Metrics Comparison

Traditional n-gram metrics like BLEU, ROUGE and even METEOR often **fail to capture meaning and domain-specific nuances** in QA. They rely on exact word overlap and ignore synonyms or factual accuracy [1] [2]. By contrast, RAGAS Answer Correctness and DeepEval's **GEval** use semantic and factual checks. RAGAS Answer Correctness explicitly combines semantic similarity with factual accuracy [3], while GEval can be tailored to any custom criterion and was designed as a superior alternative to BLEU/ROUGE for nuanced tasks [4] [5]. The table below summarizes each metric's ability (True/False) to capture key QA evaluation needs:

Aspect	GEval	RAGAS (Answer Correctness)	BLEU	ROUGE	METEOR
Human-Level Evaluation	True	True	False	False	False
Context Understanding	True	False	False	False	False
Semantic Similarity (Paraphrase Tolerance)	True	True	False	False	True
Short Answer Handling (Yes/No)	True	False	False	False	False
Date Format Flexibility	True	True	False	False	False
Factual Correctness	True	True	False	False	False
Completeness (No Key Info Missing)	True	True	False	True	True
Medical Terminology Sensitivity	True	True	False	False	False
Multiple Valid Answer Forms	True	True	False	False	True

- **Human-Level Evaluation:** BLEU/ROUGE/METEOR correlate poorly with human judgment on open-ended or domain-specific QA [1] [2], whereas GEval and RAGAS use LLM-based semantic reasoning (True).
- **Context Understanding:** Only GEval can explicitly include context in its criteria (True); n-gram metrics ignore context [6]. RAGAS's *Answer Correctness* focuses on answer vs ground truth (context handled elsewhere in RAGAS).
- **Semantic Similarity:** METEOR tolerates synonyms (True) [7]. RAGAS and GEval explicitly measure semantic similarity (True) [3] [5], but BLEU/ROUGE need exact wording (False).
- **Short Answer Handling:** BLEU/ROUGE poorly handle trivial answers (brevity penalty issues [8]), so False. GEval can be instructed to check correctness of yes/no responses (True).
- **Date Format Flexibility:** RAGAS's Answer Correctness uses semantic matching, so "Jan 15, 1967" vs "January 15, 1967" scores high (True) [3]. BLEU/ROUGE/METEOR require exact match (False). GEval can be taught these are equivalent (True).

- **Factual Correctness:** BLEU/ROUGE/METEOR do not check facts (False). RAGAS explicitly includes factual overlap [3], and GEval can be given factual criteria (True).
- **Completeness:** ROUGE/METEOR consider recall of key content (True), but BLEU does not (False). GEval can require no missing info, and RAGAS's correctness will drop if facts are missing (both marked True).
- **Medical Terminology Sensitivity:** Traditional metrics fail on domain synonyms (False) [1]. RAGAS and GEval can account for medical synonyms/terms via semantic scoring (True).
- **Multiple Valid Answers:** BLEU/ROUGE expect one form (False), METEOR is better with synonyms (True). RAGAS and GEval handle multiple correct phrasings by design (True).

Overall, **GEval** (and to a lesser extent RAGAS Answer Correctness) succeed where BLEU/ROUGE/METEOR fail, especially on semantic and factual aspects [3] [4]. This makes GEval particularly suitable for healthcare QA, where precise meaning, terminology, and human-aligned judgment are crucial.

Sources: Comparison informed by RAGAS metric definitions [3], traditional metric analyses [7] [8], and DeepEval/GEval documentation [5] [4].

---

[1] arxiv.org
https://arxiv.org/pdf/2308.07201

[2] [6] [7] Which traditional language generation metrics are applicable for evaluating RAG-generated answers, and what aspect of quality does each (BLEU, ROUGE, METEOR) capture?
https://milvus.io/ai-quick-reference/which-traditional-language-generation-metrics-are-applicable-for-evaluating-raggenerated-answers-and-what-aspect-of-quality-does-each-bleu-rouge-meteor-capture

[3] Answer correctness - Ragas
https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/answer_correctness/

[4] The G-Eval Guide to LLM Evaluation: Simply Explained - Confident AI
https://www.confident-ai.com/blog/g-eval-the-definitive-guide

[5] RAG Evaluation | DeepEval - The Open-Source LLM Evaluation Framework
https://docs.confident-ai.com/guides/guides-rag-evaluation

[8] LLM evaluation metrics — BLEU, ROGUE and METEOR explained | by Avinash | Medium
https://avinashselvam.medium.com/llm-evaluation-metrics-bleu-rogue-and-meteor-explained-a5d2b129e87f