Evaluation of Healthcare RAG QA Outputs

Problem

Retrieval-Augmented Generation (RAG) systems answer questions by retrieving relevant documents and generating responses. In our healthcare QA use-case, a single query can have multiple valid answers that are **short, paraphrased, or more elaborate** than the reference. For example, the ground-truth answer might be a simple phrase (“90 days”) while the model’s answer is a complete sentence (“The turnaround time for VOD payments is 90 days.”). Traditional evaluation must recognize both as correct even though they look different.

Challenge with Traditional Metrics

Classic metrics (BLEU, ROUGE, METEOR) score answers based on word-overlap. They **fail on short or rephrased answers** because they focus on n-gram matches and lack true semantic understanding [1] [2]. BLEU “relies on exact matching and has no concept of synonymy or paraphrasing” [1], so a semantically equivalent answer with different wording gets a low score. Likewise, ROUGE penalizes valid variations in phrasing [3]. As one analysis notes, “lexical” metrics treat semantically similar answers as **false** simply due to wording differences [2] [3]. In practice, BLEU/ROUGE/METEOR gave nearly **0% accuracy** on our QA outputs (not enough overlap), far below what humans judge as correct.

Methods (What Was Done)

- **N-gram metrics:** We ran BLEU, ROUGE, and METEOR on each answer. As expected from prior studies, these scored poorly for short or rephrased answers [1] [3].
- **RAGAS evaluation:** We used the RAGAS framework (answer relevancy, faithfulness, etc.) with and without AspectCritic. AspectCritic is an LLM-based “judge” that scores answers on aspects like correctness [4].
- **GEval with DeepEval:** We built **GEval**, a custom metric using the open-source DeepEval framework [5]. GEval leverages an LLM to score factual correctness against precise criteria we defined. It checks if each answer correctly addresses the question using synonyms, formats, and context as needed.

Scoring Approach

Each question (“row”) had **three model answers**. We considered the row correct (score = 1) if *any* one of the three answers was judged correct by the metric. (This reflects real use: a correct answer among the attempts is a win.) GEval achieved **47 out of 50** rows correct (94%). By comparison, human evaluation on the same data was about **90%** accuracy. The n-gram metrics gave very low scores (effectively near 0% rows correct). RAGAS metrics without GEval were inconsistent: for example, on some rows **RAGAS marked no answer as correct**, yet a human would say one answer was fine.

Why GEval (DeepEval) Worked Better

Our custom GEval focused on **factual correctness and completeness** rather than word overlap. It was able to:
- **Handle synonyms and paraphrases.** For instance, it recognized “Medicare Advantage Plan (MAP)” and “MAP stands for Medicare Advantage Plan” as equivalent. (Traditional metrics would penalize the different wording.)
- **Accept flexible formats** (dates, numbers, etc.). If the expected answer was “90 days,” GEval accepted answers like “The turnaround is 90 days” or even “About three months” as correct.
- **Manage short vs. long answers.** It could give credit when a model answer included extra context (e.g. adding “No, an Omnipod is not a covered CGM”) without faulting it for being longer than the terse ground truth “No.”
- **Align with human judgment.** Research shows that semantic/LLM-based evaluation correlates much better with humans than lexical metrics [6] [2]. By tuning GEval’s criteria to our domain, it captured what our experts consider a “correct” answer even when phrased differently.

Examples from Our Data

- **Short vs. long answer:** Question: “What is the turn-around time for VOD payments?” Ground truth: **“90 days.”** A model answer was *“The turn around time for VOD payments is 90 days.”* BLEU/ROUGE would penalize this because the reference has only “90 days” and lacks the extra words [1] [3]. GEval correctly gave it 1 (true) by understanding the semantics and numbers matched.

- **Paraphrase/synonyms:** Q: “What diagnosis code should be used for COVID?” Ground truth: **“Z23 is the diagnosis code for the COVID-19 vaccine.”** A model answer was *“The diagnosis code to be used is Z23.”* and another was *“You should bill it as Z23.”* All these convey the same fact, but BLEU would mark them wrong due to wording. RAGAS (without GEval) gave them zero, missing the valid code. GEval recognized the match (Z23) and marked the answers correct, as would a human reader [6] [2].

In all these cases, GEval’s use of semantic judgment (LLM understanding) allowed it to **accept valid answers despite phrasing differences** [6] [3], whereas BLEU/ROUGE would have failed and RAGAS was inconsistent.

Slide Summary

- **Problem:** We built a RAG-based healthcare QA system. Answers may be short (“90 days”), long, or rephrased, but still correct. We need an evaluation that says “correct” whenever the answer is factually right.
- **Traditional Metrics Fail:** BLEU/ROUGE/METEOR score overlap, so they miss paraphrases and short answers [1] [2]. They gave near-zero accuracy on our dataset, even though humans agreed many answers were correct.
- **Our Approach:** We ran BLEU/ROUGE/METEOR (got poor scores) and RAGAS metrics (some semantic metrics and an LLM-based *AspectCritic* for “correctness”). Then we implemented **GEval** (via DeepEval) – a custom, LLM-powered correctness metric focusing on facts.
- **Scoring Method:** Each question has 3 answers. We mark the row correct (score=1) if **any** answer is correct. Using GEval, we got 47/50 correct (94%). Humans scored about 90%. This outperforms all automatic baselines.
- **Why GEval:** Its custom criteria check facts, not words. It handles synonyms, date/number formats, and extra context. For example, it accepts “90 days” vs “three months” or recognizes “MAP” = “Medicare Advantage Plan.” Studies show semantic/LLM-based metrics align better with humans than word-overlap metrics [6] [3].
- **Examples:**
- *Short answer:* GT=“90 days”; Answer=“The turn around time is 90 days.” BLEU fails, GEval succeeds.
- *Synonym/paraphrase:* GT=“Z23”; Answer=“Use code Z23.” RAGAS failed, GEval succeeded.

These points illustrate why GEval (DeepEval) gave a robust accuracy (94%) for our RAG-QA outputs, better capturing what human experts consider correct [6] [3].

Sources: Known limitations of BLEU/ROUGE/METEOR [1] [3] and QA evaluation [2] [6]; DeepEval framework documentation [5].

---

[1] Re-examining Machine Translation Metrics for Paraphrase Identification
https://aclanthology.org/N12-1019.pdf

[2] [6] arxiv.org
https://arxiv.org/pdf/2108.06130

[3] What are the limitations of using ROUGE or METEOR for RAG evaluation, especially considering there may be multiple correct ways to answer a question with the retrieved info?
https://blog.milvus.io/ai-quick-reference/what-are-the-limitations-of-using-rouge-or-meteor-for-rag-evaluation-especially-considering-there-may-be-multiple-correct-ways-to-answer-a-question-with-the-retrieved-info

[4] aspectcritic – ValidMind
https://docs.validmind.com/tests/model_validation/ragas/AspectCritic.html

[5] GitHub - confident-ai/deepeval: The LLM Evaluation Framework
https://github.com/confident-ai/deepeval