Evaluating RAG Outputs with DeepEval's G-Eval Metric

Retrieval-Augmented Generation (RAG) systems combine information retrieval with LLMs to answer questions, often from specialized domains like healthcare. Evaluating RAG-generated answers in healthcare is challenging: answers can be short (yes/no), date-specific, or semantically correct but phrased differently. Traditional NLP metrics often fail in these cases. The DeepEval **G-Eval (GEval)** metric offers an LLM-based alternative that closely aligns with human judgment and overcomes many limitations of BLEU, ROUGE, METEOR, and RAGAS metrics.

Introduction to G-Eval and DeepEval

DeepEval is an open-source evaluation framework that uses LLMs as "judges" of model outputs. Its most flexible metric, **G-Eval (Generalized Eval)**, prompts an LLM (e.g. GPT-4) with a user-defined criterion (in everyday language) and asks it to score the output accordingly [1]. For example, one can instantiate a G-Eval metric with a criterion like *"Determine whether the actual output is factually correct based on the expected output."* [2]. G-Eval then typically uses *chain-of-thought (CoT)* prompting: the LLM first generates evaluation steps from the criterion and then fills out a form-style prompt to assign a final score [3] [2]. In practice, this means G-Eval **reasons about the answer** much as a human would, rather than relying on token overlap.

G-Eval is **task-agnostic** and can be tailored to any evaluation criterion. For RAG answer correctness, one would supply the input, generated output, and expected answer to G-Eval; it then uses the criterion to judge factual correctness. In benchmarking, G-Eval has shown *near-human performance*: it outperformed traditional metrics and other LLM-based scorers, achieving the highest correlation with human judgments on tasks like summarization [4]. In fact, it has been reported that DeepEval's early non-LLM metrics failed to register meaningful score changes even when answers were blatantly altered, and the switch to an LLM-as-judge (G-Eval) solved this issue [5].

How G-Eval Scores Are Computed

G-Eval's **technical implementation** hinges on prompt engineering. When you define a GEval metric, you provide a natural-language criterion (and optionally detailed evaluation steps). DeepEval automatically constructs a prompt containing: 1. The **evaluation criterion** (e.g. correctness, relevance).
2. Any **evaluation steps** (generated or given) that break down the task.
3. The **LLM test case details** (the question, context, generated answer, and expected answer).

The LLM processes this prompt and outputs a structured response – effectively "filling in the blanks" of an evaluation form [3]. For example, it might output a score on a 0–1 scale or answer yes/no to specific sub-questions, which DeepEval then converts into a final numeric score. By using CoT and multiple sub-questions, G-Eval reduces noise and bias in the LLM's judgment [6] [7]. Each GEval score ranges from 0 to 1, and can be interpreted as the degree to which the output meets the criterion.

Importantly, G-Eval's reliance on a powerful LLM gives it access to **semantic understanding and factual knowledge**. It can recognize paraphrases, synonyms, and real-world facts, and it can follow complex logic. Empirically, G-Eval "applied targeted techniques to reduce [LLM] limitations — ultimately resulting in a metric framework that performed on par with human judgment" [6] [4]. (In summarization tests, for instance, it achieved a Spearman correlation of 0.514 with human scores, the highest among all evaluated metrics [4].)

In summary, **GEval uses an LLM-based *chain-of-thought* evaluator**: it turns evaluation criteria into step-by-step reasoning, asks the LLM to answer those steps, and aggregates the results into a final score. This allows it to capture nuances of meaning, context relevance, and factual accuracy that surface-level metrics miss.

Limitations of Traditional Metrics (BLEU, ROUGE, METEOR)

Traditional text-overlap metrics each compare a generated answer to a reference answer by counting matching n-grams or stems. For example, **BLEU** measures n-gram *precision*, while **ROUGE** (particularly ROUGE-L) measures n-gram *recall*. **METEOR** improves on BLEU/ROUGE by considering word stems and synonyms (using resources like WordNet), and by aligning words between reference and candidate. These metrics work reasonably well for tasks like machine translation or summarization, but **they have well-known flaws**:

- **Surface overlap only:** BLEU and ROUGE count exact token matches. As Confident AI notes, they "rely on surface-level token overlap, ignoring meaning and fluency" [8]. In other words, if a model rephrases a fact correctly but uses different words, these metrics can score it very low. METEOR's synonym matching helps somewhat, but it still uses a fixed lexicon and can miss many domain-specific terms.
- **Punish valid variations:** These metrics essentially assume one "correct" reference text. Any valid answer that differs (different phrasing, additional context) can be unfairly penalized [9]. For example, BLEU would give a low score if the ground truth is "Yes" but the model says "Yes, the patient does have ...," even though the answer is correct [8]. ROUGE likewise would miss key facts if phrased differently. METEOR handles synonyms, but its coverage is limited to common words – it might not recognize a medical abbreviation or jargon as equivalent.
- **Short answers / single tokens:** These metrics struggle when answers are very short (a word or two). BLEU in particular applies a brevity penalty: a one-word answer can get a 0 score unless it exactly matches. Thus a correct "Yes" answer gets no credit for longer, more detailed outputs.
- **Non-unique answers:** Real-world QA often has multiple valid answers. Standard metrics assume the reference is the gold standard. In practice, an answer like "CO_2 emissions" vs "fossil fuel burning" may both be correct answers to "What causes climate change?", but BLEU/ROUGE might penalize one for not matching the other [9].
- **No fact-checking:** Crucially, BLEU/ROUGE/METEOR do **not** verify factual accuracy or logical consistency. They can give a high score to an answer with the right words in wrong order or even supporting a falsehood, as long as the n-grams overlap. For example, "The capital of France is Berlin" would score highly against a reference "Berlin is the capital of Germany" simply because the same words appear, despite the answer being factually wrong [10].

The Milvus blog on RAG evaluation summarizes these issues well: BLEU focuses on n-gram precision, ROUGE on recall, and METEOR adds limited synonym support [11]. But all three "rely on surface-level text

overlap and struggle with semantic equivalence" [9]. They fail to capture **factual correctness** or coherence – aspects that are *crucial* in healthcare answers [10].

Example failure cases:
- *Binary (Yes/No) answers:* If the true answer is **"Yes"** and the model answers **"Yes, the patient is allergic to penicillin."**, BLEU/ROUGE get low scores because the model added extra words; METEOR might match "Yes" but penalize the longer sentence. G-Eval, however, can recognize that both mean "correct answer: yes" and score it positively.
- *Date-specific answers:* Suppose the correct date is **"2023-05-01"** but the model outputs **"May 1, 2023."** This is perfectly correct, but token-based metrics give almost zero overlap. G-Eval can understand date equivalence.
- *Single-word ground truth:* If the reference answer is a single term (e.g. **"Insulin"**), but the model replies with a full sentence ("The patient's blood sugar requires insulin"), BLEU/ROUGE may penalize the extra content. G-Eval's criterion of correctness can accept the elaboration.

In short, BLEU, ROUGE, and METEOR are brittle for QA tasks. They may be suitable for measuring text similarity in certain tasks, but **not for nuanced, factual evaluation** of RAG outputs.

RAGAS Metrics (Context & Answer Evaluation)

The RAGAS framework offers specialized metrics for RAG pipelines. Key ones include:

- **Context Precision:** Measures whether all ground-truth relevant documents/chunks are retrieved at the top of the list [12]. In other words, if the RAG system had an annotated set of relevant context for each question, context precision checks if these are ranked highly.
- **Context Recall:** Measures the extent to which the retrieved context covers the answer (ground truth) [13]. For example, if the answer depends on a certain fact in the documents, context recall checks that the RAG's retrieved documents include it.
- **Answer Correctness:** A higher-level metric that evaluates the generated answer against the reference answer. RAGAS computes **Answer Correctness** by combining a factual F1-score and a semantic similarity [14]. It identifies true/false positives and negatives of facts in the answer versus the ground truth (like an F1 score on extracted statements), and also computes an embedding or LLM-based semantic similarity. These are then weighted and averaged into one score [14].

Limitations of RAGAS metrics: While more sophisticated than BLEU, RAGAS metrics also have blind spots for our scenario. Context precision/recall focus on the *retrieval step*. They assume one has an annotated "gold" set of supporting documents, which many real datasets lack. If your healthcare QA dataset does not label the exact retrieval context, these metrics cannot be computed. Even if they can, they say nothing about answer phrasing or format. They simply tell you if the model "looked up" the right info.

Answer Correctness improves on lexical metrics by explicitly evaluating facts, but it still depends on how facts are defined. For example, if the ground truth is "Yes" and the answer is an explanation, there is no clear set of *statements* to match. The factual F1 calculation might treat the extra explanation as "false positives," hurting the score. And if the answer uses synonyms or related terminology not present in the reference facts, the semantic similarity component may catch it, but this is heuristic and can be sensitive to wording. Thus, an elaborated correct answer could be under-scored [14].

In summary, RAGAS metrics *partition* evaluation into retrieval vs. answer scoring. They help diagnose which part of the pipeline failed (e.g. low context recall = retrieval issue). But for an end-to-end healthcare QA evaluation, especially when answers are short or binary, these metrics may not fully capture correctness. They also remain reference-based (e.g. Answer Correctness still compares to one "ground truth" answer), so they inherit some of the same limitations of BLEU/ROUGE on answer text.

GEval vs. Traditional Metrics: A Comparison

Metric	Basis	Failure Mode (Healthcare QA)
BLEU / ROUGE	n-gram overlap (precision / recall)	Penalizes valid paraphrases or added context. E.g. "Yes, …" vs. "Yes" gets low score [8]. Ignores meaning, only counts tokens [8] [9].
METEOR	n-gram + synonyms (WordNet stems)	Better than BLEU for synonyms, but misses domain terms (no WordNet for medical terms) [9]. Still punishes differently phrased but correct answers.
RAGAS: Context Precision / Recall	Retrieval relevance	Not applicable if no gold "context" labels. Even if available, these ignore answer phrasing and only measure how well retrieval finds the source info.
RAGAS: Answer Correctness	Factual F1 + semantic sim (LLM-based)	Improves over n-gram metrics, but can penalize extra correct detail or uncommon synonyms [14]. A single-word ground truth ("No") vs an explanatory answer may yield a lower score than deserved.
GEval (DeepEval)	LLM judgment with CoT reasoning	Flexible, captures nuances. Can evaluate correctness even for yes/no, dates, elaborations. Shown to align closely with human scores [1] [4].

The table above illustrates that **GEval's LLM-based approach** is fundamentally different. Instead of exact word matching, it measures *semantic and factual equivalence* according to the chosen criterion. It is designed to handle the "blind spots" of surface metrics [8] [9]. For instance, GEval can be instructed that *any elaboration of a correct yes/no answer should be treated as fully correct*, or that different date formats are acceptable. It effectively "knows" through its LLM knowledge base that two differently phrased answers are equivalent.

Why GEval is Preferred for Healthcare RAG

Healthcare QA demands **high reliability and factual correctness**. A patient's care might hinge on a single word ("positive" vs. "negative"), a date, or a clear yes/no answer. In such a domain:

- **Binary Questions:** Traditional metrics see "Yes" vs "Yes, symptoms are present" as very different. GEval, on the other hand, can follow a criterion such as *"Is the answer consistent with the fact that the patient tested positive?"* and will score both responses as correct if they mean the same. It can reason: "The additional explanation still indicates 'Yes'," giving full credit.

- **Date-Specific Answers:** BLEU would give zero score to any reformatted date. GEval can be told to check *"the timeline is correct"* and will understand date equivalence (e.g. treating "May 1, 2023" as the same point in time as "2023-05-01").
- **Elaborated vs. Single-Word Answers:** If the reference is a keyword (e.g. "Metformin"), a model might answer "The patient should take metformin daily." BLEU/ROUGE penalize the extra words. GEval can be prompted to focus on the key information ("Is the recommended medication correct?") and will recognize the elaboration as valid.

Moreover, GEval is **aligned with human judgment**. Because it reasons like a human, it naturally handles medical knowledge and context. As one G-Eval study found, it *"outperformed all other evaluators"* on a factuality/hallucination benchmark [15]. In practice, Confident AI reports that before switching to G-Eval, **DeepEval's non-LLM metrics gave nearly constant scores even when answers were clearly wrong** [5]. Only by using an LLM judge did their scores begin to reflect actual correctness.

In a mixed audience context, it is important to emphasize: **G-Eval isn't a black box without oversight**. It is fully configurable – evaluators define the criterion (e.g. "factual correctness for a medical question") and can review the LLM's reasoning if desired. This transparency, combined with higher accuracy, makes it appealing to both technical and decision-making stakeholders.

Finally, adopting GEval in healthcare RAG evaluation ensures that **critical errors are caught**. Traditional metrics could rate a dangerously incorrect answer as high, simply because it shares words with the reference [10]. GEval, by contrast, can be given a criterion about factual consistency ("Check whether the answer contradicts known medical facts or the patient record") and thus better reflect the true quality of the answer.

Conclusion

Traditional NLP metrics like BLEU, ROUGE, and METEOR are poor fits for RAG-based healthcare QA. They emphasize surface text overlap and assume fixed references, failing on short answers, paraphrases, and factual correctness [8] [9]. RAGAS metrics improve by measuring retrieval and factual overlap, but they still rely on references or require annotated contexts [14] [12]. DeepEval's **G-Eval metric** overcomes these limitations by harnessing an LLM's understanding. It evaluates answers against customized criteria using chain-of-thought prompting [2] [3], yielding scores that correlate much better with human judgment [4].

For healthcare RAG systems, where answers may be binary, date-specific, or concise, GEval provides a **more reliable, meaning-aware evaluation**. It can recognize correctness under various phrasing and flag factual errors that BLEU/ROUGE would miss. We therefore recommend using GEval (via DeepEval) as a primary metric for answer correctness in healthcare QA, possibly alongside targeted RAG-specific metrics, to obtain a comprehensive and trustworthy evaluation of system performance.

Sources: Definitions and usage of the G-Eval metric are taken from DeepEval's documentation [1] [2]. Comparative analysis draws on DeepEval/Confident AI blogs and RAGAS documentation [16] [9] [14], which discuss the shortcomings of BLEU/ROUGE/METEOR and the design of RAGAS metrics. G-Eval's performance and design (chain-of-thought scoring) are based on recent evaluation studies [3] [4]. The table and examples above illustrate typical failure modes of traditional metrics and how G-Eval addresses them.

---

1 2 **G-Eval | DeepEval - The Open-Source LLM Evaluation Framework**
https://docs.confident-ai.com/docs/metrics-llm-evals

3 5 8 **Top LLM Evaluators for Testing LLM Systems at Scale - Confident AI**
16 https://www.confident-ai.com/blog/top-llm-evaluators-for-testing-llms-at-scale

4 6 7 **The G-Eval Guide to LLM Evaluation: Simply Explained - Confident AI**
15 https://www.confident-ai.com/blog/g-eval-the-definitive-guide

9 **Which natural language generation metrics (e.g., BLEU, ROUGE, METEOR) can be used to**
10 **compare a RAG system's answers to reference answers, and what are the limitations of these**
11 **metrics in this context?**
https://blog.milvus.io/ai-quick-reference/which-natural-language-generation-metrics-eg-bleu-rouge-meteor-can-be-used-to-compare-a-rag-systems-answers-to-reference-answers-and-what-are-the-limitations-of-these-metrics-in-this-context

12 **Context Precision | Ragas**
https://docs.ragas.io/en/v0.1.21/concepts/metrics/context_precision.html

13 **Context Recall | Ragas**
https://docs.ragas.io/en/v0.1.21/concepts/metrics/context_recall.html

14 **Answer Correctness | Ragas**
https://docs.ragas.io/en/v0.1.21/concepts/metrics/answer_correctness.html

Reliable Factual Scoring with Custom GEval (DeepEval)

In healthcare QA, ensuring that a retrieval-augmented generation (RAG) answer is **factually correct and complete** is vital. Traditional reference-based scores like BLEU, ROUGE or METEOR focus on surface text overlap and often miss factual errors [1]. For example, BLEU/ROUGE can give a high score to an answer that borrows words from the reference even if it's wrong ("The capital of France is Berlin") [2] [1]. These metrics also assume a single "correct" wording and fail to penalize missing details or reward semantic equivalence [3] [2]. In contrast, a custom GEval metric uses an LLM-based judge with explicit instructions to check **factual consistency and completeness**. By phrasing the criterion as *"Determine whether the actual output is factually correct based on the expected output"* [4], we train the evaluator to behave like a domain expert checking each answer.

Custom GEval Criterion and Steps

The GEval framework (DeepEval's G-Eval) lets us write evaluation rules in natural language [4]. In our healthcare QA setting, we define the **Correctness** metric as follows:

- **Criterion:** "Determine whether the actual output is factually correct based on the expected output." [4]
- **Evaluation Steps:** We guide the LLM through specific checks:
- **Check contradictions:** Does the actual answer contradict any facts in the expected answer?
- **Penalize omissions:** Heavily penalize any missing detail or incomplete facts.
- **Allow ambiguity:** Vague phrasing or harmless differing opinions are acceptable.

These steps (copied from the DeepEval example [4]) ensure the evaluator focuses on truth. Step 1 ensures **factual consistency** – any direct factual contradiction to the known answer causes a low score. Step 2 enforces **completeness** – important medical details cannot be skipped without heavy penalty (an incomplete answer is as dangerous as an incorrect one). Step 3 builds in practical flexibility: the evaluator will *not* penalize cautious or ambiguous language (e.g. saying "research is mixed on this point" or "some studies suggest..."), nor will it penalize a benign difference of interpretation. This way the score reflects truthfulness and coverage, not writing style.

Capturing Factuality, Not Fluency

By using an LLM judge with **chain-of-thought reasoning**, GEval mimics how a human expert evaluates an answer [5]. DeepEval's G-Eval has been shown to greatly outperform BLEU/ROUGE: for example, with GPT-4 it achieved a Spearman correlation of 0.514 with human judgments on summarization, beating all previous methods [6] [5]. In practice, our metric's scoring is interpretable and consistent: the model actually *explains* if a contradiction or omission was found. This is unlike BLEU/ROUGE, which blindly count n-grams [1] [2], or RAGAS (a composite score in the ragas library) which merely averages generic metrics. GEval's focus on *"is this answer true and complete?"* aligns with how clinicians or experts review QA responses.

In short, this setup builds confidence because it rewards exactly the qualities we care about. A high GEval score means the answer matches the expected facts and includes all key information (complete and consistent). A low score means it failed on something fundamental (a hallucination or missing fact). Harmless differences in wording or nuance don't unduly hurt the score. This closely matches real-world judgment: experts are lenient about phrasing but strict about truth.

Trust and Decision-Making

For technical teams, GEval+DeepEval provides a **human-like, reproducible evaluation**. Instead of tweaking BLEU or METEOR parameters, you simply state what "correct" means, and the LLM-as-judge applies those rules. The result is a numeric score accompanied by reasoning (e.g. *"Contradiction found on fact X"*). Empirically, using G-Eval with GPT-4-like models yields far more reliable assessments than baseline scorers [6] [1].

For decision-makers, this means we can **trust the evaluation scores**. In a healthcare product, an answer that passes the GEval check has been vetted for factual accuracy and completeness. Teams can set a threshold (e.g. only accept answers scoring ≥ 0.5) knowing that the metric penalizes dangerous omissions or false claims. In contrast, relying on BLEU/ROUGE could let a wrong answer slip through if it sounds similar to the reference [3] [2]. By aligning the metric with domain correctness, GEval+DeepEval ensures that model evaluations reflect real-world quality and human expert standards. Ultimately, this leads to **better selection of RAG answers**: trusted, accurate medical answers are kept, while dubious ones are flagged – increasing overall confidence in the system's outputs.

Sources: Research shows traditional metrics like BLEU/ROUGE correlate poorly with human-judged correctness [1] [2], whereas G-Eval (an LLM-based metric) aligns much better with human judgments by using chain-of-thought evaluation [5] [6]. The above custom GEval criterion and steps are drawn from DeepEval's documentation [4], which illustrates precisely how to incorporate factual consistency and completeness into scoring.

---

[1] [5] [6] **aclanthology.org**
https://aclanthology.org/2023.emnlp-main.153.pdf

[2] [3] **Which natural language generation metrics (e.g., BLEU, ROUGE, METEOR) can be used to compare a RAG system's answers to reference answers, and what are the limitations of these metrics in this context?**
https://blog.milvus.io/ai-quick-reference/which-natural-language-generation-metrics-eg-bleu-rouge-meteor-can-be-used-to-compare-a-rag-systems-answers-to-reference-answers-and-what-are-the-limitations-of-these-metrics-in-this-context

[4] **G-Eval | DeepEval - The Open-Source LLM Evaluation Framework**
https://docs.confident-ai.com/docs/metrics-llm-evals