variA/Bly

Deterministic Evaluation Powered by Experimentation

Audit-ready. Deterministic. Cost-predictable at any scale.

WWW.VARIABLY.DEV

PubMedQA Benchmark - Technical Buyer Brief

Methodology, head-to-head results, reproduction

October 25, 2026

Audience: AI/ML engineers, platform and MLOps leads, technical evaluators running a scorer bake-off. **Bottom line:** on 2,000 clinical Q&A evaluations from PubMedQA, variA/Bly's grounding scorer leads the field on **AUC (0.943)** and is **~14× cleaner on false positives (5.5% vs 78.2%)** than DeepEval's hallucination metric — winning both the ranking-quality metric the AI-judge tools advertise and the false-positive rate they don't publish. Methodology, head-to-head results, and reproduction below.

0. Glossary

Evaluation terms

Term	Meaning
Grounding	Whether a response's claims are actually supported by the provided source material.
Hallucination	A claim the source does not support — fabrication, misattribution, or numeric/temporal drift.
Positive / negative case	Positive = a genuinely grounded response (truth = grounded). Negative = a response containing an unsupported claim (truth = not grounded).
Confusion matrix	The 2×2 of predicted vs actual: true/false × positive/negative.
Precision	Of the responses a scorer calls "grounded," the fraction that truly are. TP / (TP + FP).
Recall	Of the truly grounded responses, the fraction the scorer accepts. TP / (TP + FN).
FPR (false-positive rate)	Of the truly *ungrounded* responses, the fraction the scorer wrongly approves. FP / (FP + TN). The key safety metric.
Hallucination catch rate	1 – FPR. The fraction of hallucinations the scorer correctly flags.
AUC	Area under the ROC curve; threshold-independent ranking quality in [0, 1].

Datasets and tools

Term	Meaning
PubMedQA	Public biomedical QA dataset (Jin et al., 2019): expert-annotated questions with a PubMed abstract as source.
DeepEval	An open-source LLM-as-judge evaluation library. Two grounding-relevant metrics: `FaithfulnessMetric` and `HallucinationMetric`.
RAGAS	An open-source LLM-as-judge RAG-evaluation library (faithfulness metric), included as a second reference point.

---

1. The problem we're solving

A clinical AI assistant answers a question from a study abstract:

> **Q:** Does this therapy reduce recurrence? **Source (abstract):** "...the therapy was associated with a lower rate of recurrence over 24 months (p = 0.03)..." **AI response:** "Yes — it reduces recurrence, with a **40% reduction in high-risk patients**."

The first clause is grounded. The **"40% reduction in high-risk patients"** is not in the abstract — it is specific, plausible, and fabricated. A grounding scorer's entire job is to catch that second clause. The failure mode that matters is a scorer that reads the grounded first clause and approves the whole response — a **false positive**.

In clinical, financial, and legal AI, the two error types are asymmetric:

- **False positive** (scorer says grounded; it wasn't): the dangerous one — patient harm, a fine, a lawsuit.
- **False negative** (scorer flags for review; a human confirms it was fine): cheap — a few cents of triage.

A scorer optimized for high recall at the cost of a high false-positive rate offers little protection against the error that actually hurts.

---

2. What the benchmark measured

Construction. From 1,000 PubMedQA samples we built two test cases each:

- **Positive:** a response faithfully grounded in the abstract; truth = grounded.
- **Negative:** the same response style with a single subtle, clinically plausible but unsupported claim injected; truth = not grounded.

= **2,000 evaluations**, balanced 1,000 / 1,000. The response style is held constant across the pair so the comparison isolates **scorer behaviour**, not response quality. This is deliberately adversarial: the negative case is designed to look grounded.

Decision rule. Each scorer emits a grounding score; a response is "called grounded" at the operating threshold (0.5, the standard operational convention applied uniformly to every scorer).

Fairness rules. - Same inputs, same labels, same threshold for every scorer. - DeepEval and RAGAS use an LLM judge (the library default); variA/Bly uses its own hosted grounding scorer. No scorer is given information the others don't have. - Metrics reported over the cases each scorer actually returned (null / judge-failure outputs are excluded per-scorer and noted).

What variA/Bly does (black box). variA/Bly evaluates grounding at the granularity of individual claims and holds a response to a strict standard for what counts as supported. The scoring method is proprietary; what this benchmark measures is its externally observable behaviour on labeled data, exactly as a buyer would.

---

3. Understanding the metrics

The 2×2 confusion matrix is the foundation:

	Actual: grounded	Actual: NOT grounded
Predicted grounded	True Positive (TP)	**False Positive (FP)** ⚠
Predicted NOT grounded	False Negative (FN)	True Negative (TN)

- **Precision = TP / (TP + FP)** — trust in a "grounded" verdict.
- **Recall = TP / (TP + FN)** — coverage of real groundings.
- **FPR = FP / (FP + TN)** — the rate of dangerous approvals.
- **AUC** — threshold-independent ranking quality.

Why FPR is the number. Recall is trivially maxed by approving everything — but that drives FPR toward 100%. The honest question for a high-stakes scorer is: *when it says "grounded," how often is it wrong, and how many hallucinations does it wave through?* That is FPR and hallucination-catch, and it is the metric AI-judge tools tend not to publish.

---

4. Results — head-to-head (n = 2,000)

4.1 The headline table @ threshold 0.5

Metric	variA/Bly	DeepEval-Hall	DeepEval-Faith	RAGAS
AUC	**0.943**	0.895	0.774	0.884
Precision	**93.2%**	55.7%	50.8%	56.3%
Recall	75.4%	98.4%	98.8%	97.3%
FPR	**5.5%**	78.2%	95.5%	75.4%
Hallucination catch (1 – FPR)	**94.5%**	21.8%	4.5%	24.6%

variA/Bly leads AUC **and** FPR simultaneously — it is not trading ranking quality for precision, it is ahead on both. The AI-judge metrics post near-ceiling recall (97–99%) precisely because they approve almost everything, which is why their FPR is 75–95%.

4.2 Confusion matrices (balanced 1,000 / 1,000)

variA/Bly — precision 93.2%, FPR 5.5%:

	Actual grounded	Actual NOT grounded
Predicted grounded	754	**55**
Predicted NOT grounded	246	945

DeepEval-Hall — precision 55.7%, FPR 78.2%:

	Actual grounded	Actual NOT grounded
Predicted grounded	984	**782**
Predicted NOT grounded	16	218

DeepEval accepts 16 more real groundings than variA/Bly (984 vs 754) — but at the cost of **782 hallucinations approved vs 55**. On this set that is **~14× more dangerous approvals** for a ~30% relative gain in recall — the wrong side of the trade for any high-stakes workflow.

4.3 What it costs at production volume

At 1,000,000 evaluations/month, 50/50 mix:

Outcome (per month)	DeepEval-Hall	variA/Bly
Hallucinations wrongly approved ⚠	~391,000	**~27,500**
Hallucinations correctly caught	~109,000	**~472,500**

~363,000 more hallucinations surfaced for review per month at the same volume.

---

5. Reproduction

Public benchmark repo: github.com/varia-bly/variably-benchmark (Apache 2.0). Two layers:

Verify the math from committed raw scores (no account, zero cost):

```
git clone https://github.com/varia-bly/variably-benchmark
cd variably-benchmark/pubmedqa
python3 compare.py     # recomputes AUC / precision / recall / FPR from results/
```

Re-run the scorers from scratch:

Scorer	Requirement	Per-eval cost
DeepEval	OpenAI API key	LLM-judge surcharge
RAGAS	OpenAI API key	~$0.030 (judge)
variA/Bly	Free key from variably.dev	**$0.010** (SEU, all-in)

The AI-judge sides are self-contained with an OpenAI key. The variA/Bly side runs on Variably's hosted infrastructure and needs a Variably API key. Per-benchmark commands are in `pubmedqa/README.md`.

6. Summary

On a public, labeled, adversarial clinical benchmark:

- **Highest AUC in the set (0.943)** — best ranking quality, ahead of DeepEval's advertised metric (0.895).
- **5.5% false-positive rate** vs DeepEval's 78.2% — **~14× cleaner** on the failure mode that carries regulatory and patient-safety cost.
- **94.5% of hallucinations caught** vs 21.8%.
- **Per-claim audit trail** and **about two-thirds lower cost** than the AI-judge surcharge alone.

variA/Bly is the grounding scorer to deploy where a confidently wrong "grounded" verdict is the error you cannot afford. The numbers are public and reproducible.

The benchmark evaluation was run using **variA/Bly's deterministic evaluation platform**, which independently scores 49 metrics across 6 categories (grounding, faithfulness, attribution, coherence, safety, semantic) - with per-claim audit trails, multi-agent workflow coherence, and byte-for-byte reproducible scores across your AI workflow.

If you're building RAG systems and want to see what your evaluations actually look like, reach out:

✉ **info@variably.dev** in **linkedin.com/company/variably** ⌘ **www.variably.dev**