variA/Bly

Deterministic Evaluation Powered by Experimentation

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

WWW.VARIABLY.DEV

PubMedQA Benchmark Whitepaper

variA/Bly grounding vs DeepEval, head-to-head

October 25, 2026

variA/Bly grounding evaluation vs DeepEval, head-to-head on a public clinical Q&A benchmark.

A whitepaper on the head-to-head we ran on PubMedQA — 2,000 clinical question-answer evaluations, same data, same labels, multiple scorers, reproducible scripts. The headline this time: variA/Bly wins on the metric the LLM-as-judge tools advertise (AUC) **and** on the one they don't publish (false-positive rate).

For the full engineering write-up, see the technical brief.

---

1. The headline

We benchmarked variA/Bly's grounding scorer against **DeepEval** — a widely used LLM-as-judge evaluation library — on **PubMedQA**, a public biomedical Q&A dataset. 1,000 samples × 2 test cases each (one grounded, one subtly hallucinated) = **2,000 head-to-head evaluations** on clinical content.

Metric	variA/Bly	DeepEval (Hallucination)	DeepEval (Faithfulness)	RAGAS
AUC (ranking quality)	**0.943**	0.895	0.774	0.884
Precision (right when we say "grounded")	**93.2%**	55.7%	50.8%	56.3%
Recall (grounded answers accepted)	75.4%	98.4%	98.8%	97.3%
False-positive rate on hallucinations (FPR)	**5.5%**	78.2%	95.5%	75.4%
Hallucination catch rate (1 – FPR)	**94.5%**	21.8%	4.5%	24.6%
Per-evaluation price	**$0.010** (SEU, all-in)	LLM-judge surcharge per call	LLM-judge surcharge per call	~$0.030 (judge alone)
SDK integration latency	**<1 ms** (async scoring)	seconds (LLM judge in-line)	seconds	~2.5 sec

Metric	variA/Bly	DeepEval (Hallucination)	DeepEval (Faithfulness)	RAGAS
Per-claim audit trail	**Yes**	No	No	No

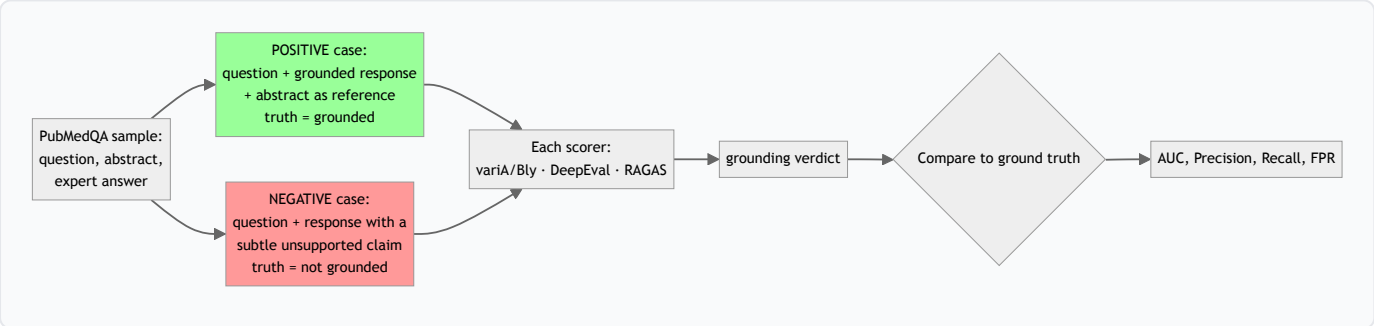
> **The number nobody else publishes:** on this clinical set, DeepEval's hallucination metric marks **78% of hallucinated answers as grounded** when they aren't. variA/Bly's false-positive rate on the same answers is **5.5%** — it catches **94.5% of hallucinations vs DeepEval's 22%**. And it does this while scoring the **highest AUC in the set (0.943)** — so this isn't a precision-for-ranking trade. variA/Bly is simply the better scorer on both axes.

---

2. What this benchmark measured

PubMedQA is a public dataset of expert-annotated biomedical questions, each paired with a PubMed abstract as the source and a yes/no/maybe answer. It is a hard, specialized-domain test — exactly the kind of high-stakes content where a wrong "this is grounded" verdict carries real cost.

For every sample we created two test cases — one positive (a response faithfully grounded in the abstract) and one negative (the same style of response with a subtle, clinically plausible but unsupported claim) — and ran every scorer on both:



The same realistic response style was used across cases — only the grounding of the claim changes. This isolates **scorer behaviour** from response quality, and it is a deliberately adversarial test: the hallucinated claim is subtle, the kind a busy clinician might not catch.

---

3. Why the false-positive rate is the number that matters

In a regulated or clinical AI workflow, the two failure modes have very different costs:

- A **false positive** (scorer said grounded; a reviewer finds it wasn't) is the dangerous one. In clinical or financial contexts it can mean patient harm, a fine, a remediation order, or a lawsuit.
- A **false negative** (scorer flagged for review; a human confirms it was fine) is just routing work — a few cents of human triage.

An evaluator that approves **78% of hallucinations** — DeepEval's hallucination metric on this set — offers almost no protection against the failure mode that actually hurts you. High recall on a near-100% "everything is grounded" verdict is not the same as being *right*.

For a 1,000,000-evaluations-per-month workflow with a 50/50 mix of real groundings and hallucinations:

Outcome (per month)	DeepEval (Hallucination)	variA/Bly
Hallucinations wrongly approved as grounded ⚠ (breach risk)	~391,000	**~27,500**
Hallucinations correctly caught for review	~109,000	**~472,500**

variA/Bly surfaces ~363,000 more hallucinations per month at the same evaluation volume — the ones a human should look at before they reach a patient or a customer.

> **Audit note.** Every variA/Bly grounding decision comes with a structured per-claim record — the claim, the verdict, the failure reason, and the exact source passage the scorer was looking at. LLM-as-judge tools return one aggregate number, and the same inputs can produce different scores on different runs because of judge drift.

---

4. Cost predictability at scale

DeepEval's metrics call a paid LLM judge on every evaluation — a per-call surcharge, before the customer pays to host and operate the harness. variA/Bly is priced as a monthly subscription with **SEU pricing $0.010 per evaluation on Starter and $0.009 on Business, down to $0.005 at enterprise volume** (a typical evaluation is 1 SEU) — all-in, no separate hosting cost.

Workload	LLM-as-judge cost *(alone)*	variA/Bly all-in (SEU)	You save
100,000 evaluations / month	~$3,000 / mo	**~$1,000 / mo** *($0.010/SEU)*	**67%**
1,000,000 / month	~$30,000 / mo	**~$10,000 / mo** *($0.010/SEU)*	**67%**
10,000,000 / month	~$300,000 / mo	**~$50,000 / mo** *($0.005/SEU, enterprise volume)*	**83%**

variA/Bly is about two-thirds less than the LLM-judge surcharge alone — before the harness's own hosting and ops. The gap widens at scale, because production LLM-as-judge setups typically use stronger (pricier) judge models than the benchmark default to hold accuracy.

---

5. Verifying these numbers

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

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

```
git clone https://github.com/varia-bly/variably-benchmark
cd variably-benchmark/pubmedqa
python3 compare.py
```

Recomputes the AUC / precision / FPR table at the top of this page directly from the per-sample JSON outputs committed in `pubmedqa/results/`. No rescoring required.

Re-run the scorers from scratch:

Scorer	What you need	Per-eval cost
DeepEval	OpenAI API key	LLM-judge surcharge per call
RAGAS	OpenAI API key	~$0.030 (judge)
variA/Bly	A free key from variably.dev	**$0.010** (SEU, all-in)

The DeepEval and RAGAS sides are self-contained with an OpenAI key. The variA/Bly side needs a Variably API key because the scoring runs on Variably's hosted infrastructure. Commands are in the per-benchmark README.

---

6. The summary, one line

On 2,000 clinical Q&A evaluations, variA/Bly is the most accurate grounding scorer in the set — highest AUC (0.943), and it catches 94.5% of hallucinations where DeepEval catches 22% — with a per-claim audit trail and a predictable subscription price about two-thirds below the LLM-judge surcharge alone.

If your AI touches patients, money, or contracts, the cost of approving a hallucination is the cost that matters — and it's the one this benchmark is about.

We're expanding the public benchmark surface — **AggreFact** (summarisation) and **FEVER** (claim verification) are next. Follow github.com/varia-bly/variably-benchmark for updates.

---

7. Glossary

Term	Meaning
FPR (False Positive Rate)	The rate at which a scorer wrongly approves a hallucinated answer as grounded. **The headline differentiator for high-stakes buyers.** variA/Bly: 5.5%. DeepEval: 78.2%.
AUC	"Area Under the Curve" — a threshold-independent ranking-quality score in [0, 1]; the metric LLM-as-judge tools tend to advertise. variA/Bly leads it here at 0.943.
Recall	Of the genuinely grounded answers, the share the scorer accepts. High recall alone is easy — approve everything — which is why it must be read alongside FPR.
PubMedQA	A public biomedical question-answering dataset (Jin et al., 2019): expert-annotated questions with a PubMed abstract as the source. Used here because it is public, labeled, and clinically realistic.
SEU pricing	"Standard Evaluation Unit" — variA/Bly's all-in subscription model: $0.010 per SEU on Starter, $0.009 on Business, and from $0.005 at enterprise volume. A typical evaluation is 1 SEU.

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**