variA/Bly

Deterministic Evaluation Powered by Experimentation

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

WWW.VARIABLY.DEV

# PubMedQA Benchmark - Business Buyer Brief

Clinical accuracy, audit, and the math at scale

October 25, 2026

**Audience:** compliance leads, clinical-quality and risk officers, operations directors, heads of customer support, AI program owners. People who buy AI tools but don't necessarily build them. **Length:** 3-4 pages. No AI/ML background required. **Bottom line:** variA/Bly catches the AI failure mode that costs you a compliance breach — a confidently wrong answer. On a public clinical benchmark we ran head-to-head against DeepEval, a widely used AI-judge evaluation library, variA/Bly is **~14× cleaner** on that failure — and it also scores the **highest overall accuracy in the test**, while costing **about two-thirds less** per evaluation.

## 1. The headline

We ran 2,000 clinical question-answer evaluations from **PubMedQA** (a public biomedical dataset) — half grounded, half with a subtle, plausible-but-unsupported claim — through each scorer.

| Metric | variA/Bly | DeepEval | What it means in business terms |
|---|---|---|---|
| Precision when AI says "this is grounded" | **93.2%** | 55.7% | When variA/Bly approves an answer, it's right **~9 out of 10** times. DeepEval is barely better than a coin flip. |
| Wrong-yes rate on hallucinations (FPR) | **5.5%** | 78.2% | DeepEval approves ~78 of 100 hallucinated answers. variA/Bly approves ~6. **~14× cleaner on the failure that hurts you most.** |
| Hallucinations caught | **94.5%** | 21.8% | variA/Bly flags ~19 of 20 made-up answers for review. DeepEval flags ~1 in 5. |
| Overall accuracy (AUC) | **0.943** | 0.895 | variA/Bly is the **most accurate scorer in the test** — it doesn't trade ranking quality for safety, it leads on both. |
| Per-evaluation cost | **$0.010** | AI-judge surcharge per call | variA/Bly is **about two-thirds cheaper** than the AI-judge cost alone, before hosting. |
| SDK integration latency | **<1 ms** | seconds | variA/Bly evaluates *off* your customer's request path. AI-judge tools sit on it. |

| Metric | variA/Bly | DeepEval | What it means in business terms |
|---|---|---|---|
| Per-claim audit trail | **Yes** | No | variA/Bly returns a structured record per claim — the exact source it checked and why the claim passed or failed. AI-judge tools return one aggregate number that can drift between runs. |
| Pricing model | **Monthly subscription** | Per-call third-party API + your own ops | Predictable spend at any scale. |

A quick story to make it concrete.

A health system runs an AI assistant that answers clinician and patient questions from the medical literature. Someone asks: *"Does this therapy reduce the risk of recurrence?"* The assistant reads the study abstract and replies:

> *"Yes — the study found it reduces recurrence, with a 40% reduction in high-risk patients."*

The abstract supports the first half. The *"40% reduction in high-risk patients"* was **made up** — plausible, specific, and nowhere in the source. A clinician skims it, it reads authoritatively, and it informs a decision it shouldn't.

This is what AI tools call a **hallucination**: the AI stated something the source documents don't support.

**variA/Bly catches that fabricated claim ~19 times out of 20. DeepEval lets it through ~78% of the time.**

That single contrast — 5.5% vs 78.2% — is the variA/Bly thesis in one line: *be right when you say yes*, even on adversarial answers designed to look grounded. It's the property regulated and clinical buyers refuse to compromise on, and it's the one nobody else publishes.

---

## 2. Cost predictability at scale

DeepEval's metrics call a paid third-party AI model as their judge on every evaluation — a per-call surcharge — and you also pay separately 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 | AI-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%** |
| 100,000,000 / month | ~$3,000,000 / mo | **~$500,000 / mo** *($0.005/SEU, enterprise volume)* | **83%** |

**variA/Bly is about two-thirds less than the AI-judge cost alone — before adding what you pay separately to operate the harness itself.** Including hosting and ops, the all-in comparison widens further.

This matters for two reasons:

1. **AI workloads grow fast.** An assistant that starts at 100K evaluations/month often hits 1M within a year and 10M within two. With a per-call AI judge, that line item alone grows from $36K/yr to $360K/yr to $3.6M/yr on the

same product — *before* hosting, *before* the inevitable upgrade to a stronger (pricier) judge model. variA/Bly stays about two-thirds below that line as you scale.

2. **Budget approval is easier.** Procurement can plan for a tiered subscription with a known per-evaluation rate, instead of a per-call third-party surcharge that moves with a vendor's price list.

> **Audit note.** variA/Bly returns a structured, timestamped per-claim record — the claim, the verdict, the reason, and the exact source passage checked — so any decision can be reviewed later. AI-judge tools return a single aggregate number, and the same inputs can produce different scores on different runs because of judge drift.

---

## 3. Summary

| What you get | What it means in business terms |
|---|---|
| **93.2% precision** ("right when we say yes") | Fewer confidently-wrong answers reaching a patient, customer, or decision. |
| **5.5% false-positive rate** on hallucinations | When the source doesn't support the claim, we catch it ~14× more often than the alternative. |
| **0.943 AUC — highest in the test** | You don't trade accuracy for safety; variA/Bly leads on both. |
| **SEU pricing — $0.010/eval entry, from $0.005/eval at enterprise volume** | Predictable monthly bill, **about two-thirds less than the AI-judge cost alone**. |
| **<1 ms SDK integration latency** | Scoring is async — it doesn't sit on your customer's request path. |
| **Per-claim audit trail** | The compliance team has the records. The CTO has the explainability. |

**If you're operating AI in a regulated or clinical setting and an auditor's — or a clinician's — question keeps you up at night, this is the layer you want underneath your AI workflow.**

We'd be glad to share a deeper walkthrough on a 30-minute call, or to run a confidential evaluation on a sample of your live workflow data.

---

## 4. Glossary

Specialized terms only. Common AI terminology (RAG, hallucination, grounding, precision, recall) is defined inline above where it appears.

| Term | Meaning |
|---|---|
| **FPR** (False Positive Rate) | The rate at which a scorer wrongly approves a hallucinated answer as grounded. **The headline differentiator for compliance buyers.** variA/Bly: 5.5%. DeepEval: 78.2%. |
| **AUC** | "Area Under the Curve" — a threshold-independent overall-accuracy score in [0, 1]; the metric AI-judge tools tend to advertise. variA/Bly leads it here at 0.943. |
| **PubMedQA** | A public biomedical question-answering dataset (Jin et al., 2019): expert-annotated questions with a PubMed abstract as the source. Used because it's public, labeled, and clinically realistic. |

| Term | Meaning |
|---|---|
| **SEU pricing** | "Standard Evaluation Unit" pricing — 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**