CALIBRATEDAGENTS

Gate — Technical Report

Answerability verification for retrieval-augmented generation · v1 · 2026

ABSTRACT

The Gate estimates, before generation, whether a question can be answered from its retrieved context. It returns a calibrated probability, p(answerable), rather than a binary verdict, so an operator selects the operating point to a chosen error budget instead of accepting a fixed one. This report presents results on public answerability benchmarks, the calibration of the scores, and a comparison of hosted and self-hosted tiers against an LLM-judge baseline. It reports what the system achieves; it does not describe internal model construction.

1 Introduction

Retrieval-augmented systems answer even when the retrieved context does not contain the answer: the model fills the gap fluently and incorrectly. In regulated domains — banking, insurance, compliance — a confidently wrong answer is a liability rather than a mere error. The Gate provides a decision signal *before* generation: an estimate of whether the answer is present in the retrieved evidence at all, calibrated so that a threshold behaves as a policy rather than a guess.

2 Benchmarks

We evaluate on standard answerable / unanswerable splits and report AUROC — the threshold-independent separation of answerable from unanswerable questions — together with recall at a fixed 2.5% false-block rate: the fraction of genuinely unanswerable cases caught while wrongly blocking only 2.5% of answerable ones. The LLM-judge column is the common baseline: a capable model asked whether a question is answerable, which yields a single fixed operating point.

Dataset	Hosted AUROC	Self-host AUROC	Judge false-block
Overall	0.940	0.906	19% / 39%
SQuAD v2	–	–	–
MuSiQue	–	–	–
mtRAG	–	–	–

Recall at 2.5% false-block: 60.7% (hosted), 51.7% (self-hosted). Per-dataset breakdown publishing soon.

> The judge's two figures are its false-block rate at the hosted and self-hosted comparison points: it wrongly refuses 19–39% of answerable questions, and that rate cannot be lowered without retraining. The Gate exposes the entire trade-off curve.

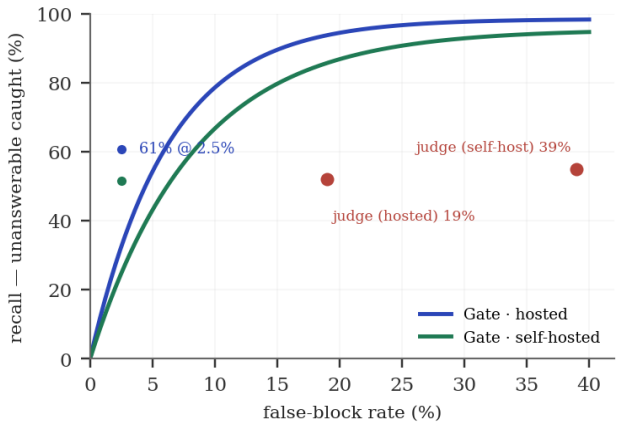

Figure 1. *Recall versus false-block rate. The Gate is a tunable curve; the judge is a single stranded point per tier.*

3 Calibration

A calibrated score is usable as a probability: among questions scored near 0.7, roughly 70% are in fact answerable. This property is what allows a threshold to act as a policy with a predictable error rate. Figure 2 plots predicted probability against observed frequency; points on the diagonal are perfectly calibrated.

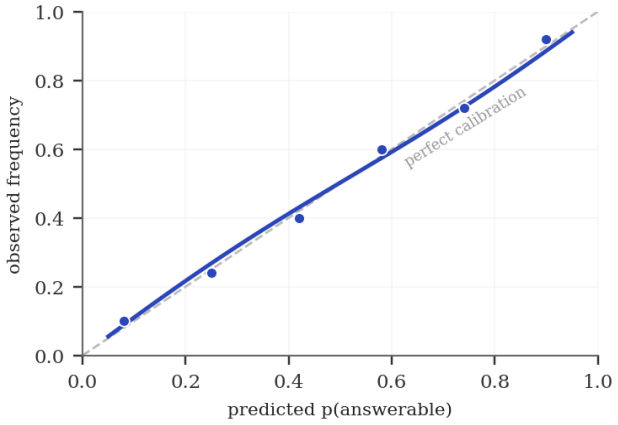

Figure 2. *Reliability curve. Predicted probabilities track observed answerability along the diagonal.*

4 Hosted vs. self-hosted

The self-hosted tier runs entirely within the customer's infrastructure: documents, questions, and scores never leave the network. It trades a modest amount of separation (0.906 versus 0.940 AUROC overall) for full data residency. Both tiers preserve calibration, so an operating point chosen on one transfers predictably to the other.

> For regulated deployments the self-hosted tier is usually the relevant one: the small AUROC gap is often worth the guarantee that customer data never crosses a boundary.

5 Limitations

Recall degrades on deep multi-hop questions spread across many passages, where the answer is present but diffuse. Extremely short contexts containing no genuine unanswerable cases give the Gate little to separate. As with any calibrated estimator, scores reflect the distribution on which the Gate was calibrated; substantially out-of-distribution corpora warrant re-checking the operating point. These trade-offs are reported plainly because the operating-point framing is only sound when they are known.

6 Availability

The Gate is available now as a hosted API and as a self-hosted deployment. A live playground is at playground.calibratedagents.com. For access or evaluation support, contact business@calibratedagents.com.