Evaluating Closed-Domain RAG Systems with Expert-Written MCQs

Hila Reicher
Tel Aviv University
Israel
hilareicher@mail.tau.ac.il

Stav Koren
Tel Aviv University
Israel
stavk@mail.tau.ac.il

Slava Novgorodov
Tel Aviv University
Israel
slavanov@post.tau.ac.il

Tova Milo
Tel Aviv University
Israel
milo@cs.tau.ac.il

Abstract

Evaluating Retrieval-Augmented Generation (RAG) systems in closed-domain settings — where answers must be grounded in a fixed, domain-specific corpus — remains an open problem. Existing approaches rely on synthetic question generation, manual annotation, or general-purpose benchmarks, each with structural limitations. We propose repurposing expert-written multiple-choice questions (MCQs) as a reusable evaluation signal for closed-domain RAG. MCQs are human-authored, curriculum-aligned assessment artifacts: they test concepts domain experts consider important, are written independently of the indexed corpus, and include expert-designed distractors that distinguish correct understanding from plausible misconceptions. Our methodology decomposes each correct answer into atomic factual claims, rewrites each option as a standalone declarative probe, and evaluates the RAG system through binary per-option judgments.

We instantiate the methodology on a production university tutoring system and release a publicly reproducible benchmark. The per-option evaluation exposes differences in generation accuracy, abstention behavior, and distractor credulity that are not captured by retrieval-side evaluation alone. These results show that expert-written MCQs provide a practical and diagnostically rich evaluation resource for closed-domain RAG systems.

CCS Concepts

• **Information systems** → **Retrieval models and ranking**; • **Computing methodologies** → *Natural language generation.*

Keywords

Retrieval-Augmented Generation, RAG evaluation, multiple-choice questions, nugget-based evaluation, closed-domain retrieval

ACM Reference Format:
Hila Reicher, Stav Koren, Slava Novgorodov, and Tova Milo. 2026. Evaluating Closed-Domain RAG Systems with Expert-Written MCQs. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3799682.3840817

1 Introduction

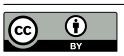
Retrieval-Augmented Generation (RAG) has emerged as the dominant paradigm for deploying large language models over private or domain-specific document corpora. A standard RAG pipeline operates in four stages [7]: (i) the corpus $\mathcal{D}$ is chunked into passages and each passage is embedded into a dense vector index; (ii) at inference time the user query is embedded and the top-k passages most semantically similar to the query are retrieved from the index; (iii) the query and retrieved passages are assembled into a prompt; (iv) a generator LLM produces a response conditioned on the retrieved context. Compared to relying solely on parametric knowledge encoded during pre-training, this retrieve-then-generate pattern grounds outputs in a specific authoritative source and reduces hallucination on long-tail facts [6]. The architecture has proven particularly valuable in *closed-domain* settings — where the retrieval corpus is fixed and bounded, as in enterprise knowledge bases, clinical documentation, legal repositories, and educational course materials — and answers must be grounded in this corpus rather than general world knowledge.

Despite their widespread adoption, evaluating RAG systems in closed-domain settings remains an open problem. The challenge arises from three sources. First, the evaluation must be domain-specific: a generic benchmark cannot capture whether a system correctly retrieves and reasons over its own corpus. Second, evaluation must distinguish failures at different pipeline components — retrieval, generation, and their interaction can fail independently and require different remediation. Third, obtaining high-quality ground-truth question-answer pairs for system evaluation is expensive: it typically requires manual annotation by domain experts or synthetic generation by LLMs, both of which introduce noise, bias, or significant cost.

A key observation of our work is that in many closed-domain RAG settings — educational institutions, professional certification bodies, medical licensing authorities, compliance training programs — a valuable and largely untapped evaluation resource already exists: expert-written multiple-choice questions (MCQs). MCQs are prevalent in these domains because they are machine-gradable and easier to verify than open-ended questions. As a running example throughout the paper (§3.3), we use the following MCQ from an introductory-psychology exam:

Q. How does Parkinson's disease affect movement?

A. Loss of serotonin neurons, producing mood symptoms without motor effects.
B. Degeneration of dopamine-producing neurons in the substantia nigra, reducing basal-ganglia signaling needed for smooth voluntary movement. ✓
C. Excess dopamine release in the cerebellum, causing uncontrolled rapid movement.
D. Damage to peripheral motor neurons, preventing muscles from receiving signals.



CIKM '26, Rome, Italy

ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3840817

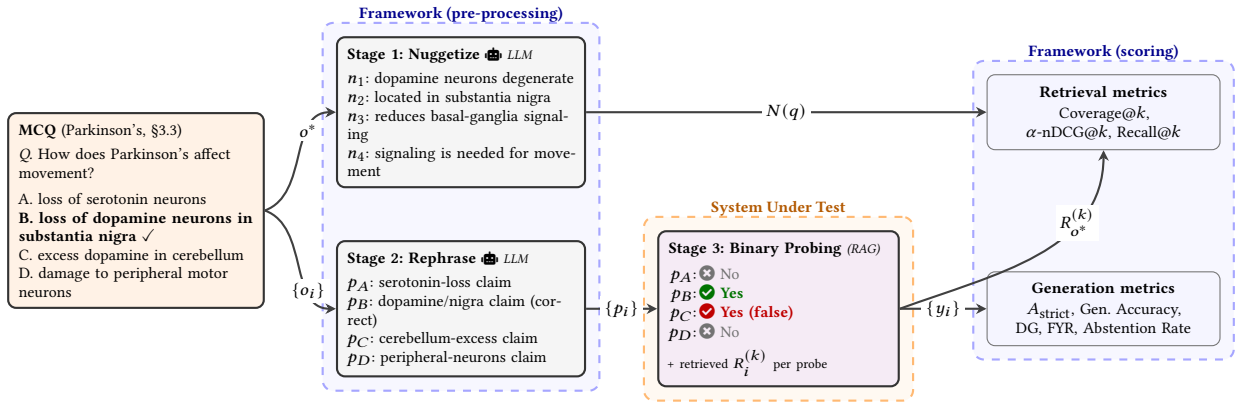


Figure 1: MCQ-to-metrics architecture, traced on the Parkinson's worked example (§3.3). Stage 1 decomposes the correct option into atomic nuggets; Stages 2–3 submit each option as a standalone probe to the RAG system under test and collect a Yes/No judgment. $N(q)$ and $R_{o^*}^{(k)}$ feed retrieval metrics; $\{y_i\}$, $N(q)$, and o^* feed generation metrics. Probe p_C produces a false-yes on the expert-designed distractor — the discrimination event DG and FYR surface.

For RAG evaluation, MCQs' value is not only that they provide a correct answer, but that the option structure separates the correct claim from several plausible-but-wrong alternatives. The example above illustrates why this structure can be turned into an evaluation signal: the correct option contains the facts the system should support, while the distractors encode confusable hypotheses the system should reject. We therefore first describe how the methodology converts this structure into probes, and then return to why this makes MCQs preferable to the main alternatives.

Figure 1 traces our pipeline on the question above. Given a set of MCQs and their correct answers, the pipeline (1) decomposes each correct option into atomic *nuggets* [13, 23] — minimal, self-contained factual claims (in our example: Parkinson's causes dopamine-neuron loss; dopamine neurons reside in substantia nigra; substantia-nigra dopamine loss reduces basal-ganglia signaling; basal-ganglia signaling enables voluntary movement); (2) rephrases each option as a standalone declarative probe (e.g., probe p_C asserts "Parkinson's involves excess dopamine in cerebellum"); (3) submits each probe independently to the RAG system, which returns a Yes/No judgment per probe and a retrieved set; (4) scores the outputs with standard nugget-based retrieval and answer-recall metrics, together with additional generation metrics per-option probing uniquely enables — notably the *Discrimination Gap* (DG), which captures the asymmetry between the system's judgments on correct and distractor options. On our running example, the system correctly rejects A and D and correctly accepts B but produces a *false-yes* on p_C: the expert-designed distractor (excess dopamine, wrong region) is accepted as corpus-supported — exactly the discrimination event the methodology is designed to surface. Framed differently: the contribution is not using MCQs to test what the system knows, but using the option structure to evaluate retrieval-grounded discrimination within the constraints of a fixed corpus.

This methodological framing clarifies why exam MCQs provide a different and stronger evaluation signal than existing evaluation sources. First, exam MCQs are *authored about the material rather than from it*. Synthetic generators derive questions from indexed passages, so they cannot ask about what the corpus does not contain, biasing the benchmark toward facts the retriever can already surface. Expert-written exams are produced independently of the corpus: instructors write questions to test what they consider essential, encoding an importance hierarchy the corpus may not reflect, and exposing corpus gaps as evaluation failures rather than masking them.

Second, exam MCQs come with *expert-designed distractors*: wrong options written by domain experts to represent plausible misconceptions and near-correct alternatives, calibrated to challenge respondents with partial understanding. Producing such distractors is itself a non-trivial expert craft [3], and the literature consistently reports that LLM-generated distractors underperform human-authored ones on plausibility, cognitive engagement, and difficulty balance [10, 11]. The MCQ banks already maintained by deployed assessment programs therefore represent expert investment our methodology reuses — and the per-option discriminative signal they carry is one our methodology measures directly through the *Discrimination Gap* metric introduced in §3.2.2. A natural alternative would be to convert each MCQ to an open-ended question by dropping its options and using the stem alone, then applying open-ended evaluation. This discards the option structure's distinctive value: only a per-option judgment tests whether the system rejects expert-designed distractors written precisely to be confusable with the correct option. The discriminative signal is in the options, not the stem.

Exam MCQs are also *pedagogical artifacts*: instruments designed to test whether a respondent can apply the material to a problem. In settings where users engage with the corpus to apply it, exam-derived probes and real user queries originate from the same underlying cognitive task, whereas synthetic questions, generated from corpus passages, do not reflect this kind of application-oriented usage. We verify this empirically on one deployment in §5.4.

To test whether these advantages translate into different evaluation signals in practice, we instantiate the methodology on a

production RAG-based university digital tutor system [14] and run experiments on three real undergraduate university courses: introductory psychology, calculus, and health-economics policy. On the same RAG system and corpora, exam-based and synthetic evaluation produce substantively different signals on both common evaluation metrics and the methodology's novel metrics (§5.3), and a practitioner relying on synthetic evaluation alone would draw materially different conclusions about how the system actually behaves. To our knowledge, no prior work has proposed a systematic methodology for leveraging existing MCQs as a RAG evaluation signal.

Contributions.

- An MCQ-based evaluation methodology for closed-domain RAG: a per-option declarative-probing pipeline that preserves the discriminative signal of expert-designed distractors, together with a metric set centered on the *Discrimination Gap* (DG). The methodology pairs these with standard nugget-based retrieval and answer-recall metrics, making expert-written exam MCQs a reusable evaluation artifact.
- A publicly reproducible benchmark on MIT OpenCourseWare's Introduction to Psychology (9.00SC), enabling direct comparison on the same corpus by future work.
- Empirical findings from a production digital tutor on three undergraduate courses: the metrics are *sensitive* to RAG configuration changes and *specific* to the component varied, and exam-based and synthetic evaluation diverge on both common and novel metrics in ways synthetic-only evaluation would hide.

2 Related Work

RAG and its evaluation. RAG [6, 7] grounds LLM responses in external knowledge and is now widely deployed in domain-specific settings. Existing IR benchmarks such as BEIR [18] evaluate retrieval in zero-shot settings across heterogeneous domains but were not designed for closed-domain RAG. CRAG [25], BRIGHT [17], and RAGTruth [9] address specific aspects of RAG evaluation (factual grounding, reasoning-intensive retrieval, hallucination detection), but none provides a general methodology for constructing benchmarks from existing domain resources.

Nugget-based benchmark construction. The closest work in spirit to ours is FreshStack [19], which builds IR/RAG benchmarks from Stack Overflow questions grounded in GitHub documentation. The shared design principle is repurposing evaluation artifacts that already exist for non-RAG-evaluation purposes — community Q&A in their case, expert-written exam MCQs in ours — rather than synthesizing new ones. Our methodology departs in two ways: (1) we source questions from MCQ exams rather than open-ended community Q&A, preserving the expert-designed distractor structure; (2) we extend the generation-side metric set, which prior work scores via nugget recall (A_{strict} [13]), with per-option discriminative metrics (DG, FYR, Abstention Rate) that the MCQ option structure uniquely enables.

MCQ benchmarks in NLP. MMLU [4], MedQA [5], and MedMCQA [12] use multiple-choice questions to evaluate LLM knowledge directly. These benchmarks are designed to test what the LLM knows, not how a RAG system retrieves and grounds information from a specific corpus.

Synthetic question generation. RAGAS [2] and ARES [15] construct evaluation sets by prompting an LLM to generate question–answer pairs from passages sampled from the same corpus the retriever retrieves from. This is appealing because it removes the manual annotation bottleneck, but two structural limitations compound: questions are biased toward facts easy to express in the corpus's own language (a property retrieval embeddings already reward), and the resulting (q, a) pairs lack expert-designed distractors and so cannot test discriminative reasoning over plausible misconceptions — exactly the property synthetic generation cannot reproduce [11]. Recent work has also questioned the trustworthiness of synthetic benchmarks: van Elburg et al. [22] find that synthetic-question rankings of RAG systems align with human-labeled benchmarks for retriever parameters but break down when comparing generator architectures. We show in our results (§5.3) that the two regimes can produce substantively different readings on both common evaluation metrics and the methodology's novel metrics on the same RAG system.

Option-aware retrieval. A separate line of work studies how to retrieve effectively for MCQ answering rather than how to use MCQs to evaluate retrieval. QA-GNN [26] forms per-choice queries over a knowledge graph; OADR [16] trains option-aware dense retrievers using (question, option) embeddings supervised to match the oracle (question + correct answer) embedding. Our option-rephrasing step is motivated by the same compound-query failure mode but treats decomposition as an evaluation-time query-construction choice rather than as part of the retrieval system, keeping the methodology non-invasive to the system under test.

3 Methodology

The methodology consists of an MCQ-to-probe pipeline (§3.1) that transforms each MCQ into per-option probes against the RAG system, and a metric set (§3.2) that scores the system's responses with standard nugget-based metrics together with additional metrics our methodology enables. Figure 1 gives an overview.

3.1 MCQ-to-Probe Pipeline

The pipeline takes as input a set Q of MCQs over a fixed corpus $\mathcal{D}$. Each question $q \in Q$ has a stem, n options $\{o_1, \ldots, o_n\}$, and a designated correct answer $o^*(q)$.[1] We assume the RAG system under evaluation was intended to be built over the same corpus; cases where the corpus is incomplete relative to what the exam tests are themselves a signal the methodology surfaces, not a precondition violation. The pipeline transforms each question through three stages — nuggetization, option rephrasing, and binary factual probing — producing n per-option *probes* that are submitted independently to the RAG system. The pipeline requires no modification or retraining of the retriever or generator; it only requires that retrieved sets be exposed for logging.

[1] For clarity we describe the single-correct-answer case throughout, as in our experiments; the methodology extends to multi-correct MCQs by treating each correct option as a Yes-target and averaging DG and FYR over the correct vs. distractor sets, respectively.

Stage 1: Nuggetization. The notion of a *nugget* originates in the TREC QA tracks [23]. A nugget is a minimal, self-contained factual statement that must hold for the correct answer to be considered supported. For each $q \in Q$, we decompose $o^*(q)$ into a finite set of atomic nuggets $N(q) = \{n_1, n_2, \ldots\}$ via an LLM prompted with the stem, all options, and the correct answer. The prompt instructs the model to produce statements sufficient to justify the correct answer and to resolve cross-option references ("none/all of the above", etc.). We extend the open-source Nuggetizer library of Pradeep et al. [13] to the MCQ setting; the library has been human-validated against gold nugget annotations in prior work [13, 19].

Stage 2: Option Rephrasing. We rephrase each option as an independent declarative probe p_i, producing n probes per question. The rephrasing converts the stem-plus-option construction into a standalone factual claim suitable for binary judgment in Stage 3, expanding cross-option references where present ("none/all of the above"). This step is motivated by a known failure mode in MCQ retrieval: presenting the system with the original stem-plus-options form produces noticeably worse retrieval as the embedding drifts toward the union of all option hypotheses. OADR [16] addresses a related concern by training option-aware dense retrievers using (question, option) embeddings supervised to mimic the oracle (question + correct answer) embedding; our per-option probing sidesteps the compound-query problem at the query-construction stage, preserving non-invasiveness with respect to the system under evaluation. The choice also mirrors a broader convergence in retrieval-for-RAG toward breaking complex queries into structured sub-queries: Thakur et al. [19] report that GPT-4o question decomposition outperforms other pooling techniques across five technical-corpus topics, supporting structured query decomposition as an evaluation-time tool rather than a retriever modification.

Stage 3: Binary Factual Probing. Each probe p_i is submitted to the RAG system as a binary-judgment query asking whether the probe's claim is supported by the retrieved context; the methodology requires only that the system commit to a Yes/No verdict per probe, or abstain, and that an explanation accompanying the judgment be available for nugget scoring. The system returns its judgment $y_i \in \{\text{Yes}, \text{No}, \perp\}$, where $\perp$ denotes abstention (the system declines to commit because the retrieved context is insufficient), together with a natural-language explanation and the retrieved set $R_i^{(k)}$. We write $R_{o^*}^{(k)}$ for the correct-option probe's retrieval and $R^{(k)}(q) = \bigcup_{i=1}^{n} R_i^{(k)}$ for the union across all probes. Following standard TREC RAG support-judgment protocols [20], we evaluate per-document/per-nugget support: the support function $\sigma : \mathcal{D} \times \mathcal{N} \to \{0, 1\}$, where $\mathcal{N}$ is the universe of nuggets, is an LLM entailment judgment [8, 21, 27] of whether document d supports nugget n; we overload notation to write $\sigma(R, n) = \max_{d \in R} \sigma(d, n)$ for any retrieved set $R \subseteq \mathcal{D}$. The retrieval metrics of §3.2 are computed on $R_{o^*}^{(k)}$, the canonical query's retrieval; the pool $R^{(k)}(q)$ serves as the judgment-pool device for Recall@k's relevant-document approximation.

3.2 Metrics

The pipeline of §3.1 produces, per question, per-probe retrieved sets, per-probe binary judgments and explanations, and the nugget set $N(q)$ derived from the correct answer. We score this output along two complementary metric families. The first comprises standard nugget-based retrieval and answer-recall metrics from the TREC RAG community [13] — Coverage@k, α-nDCG@k, Recall@k, and A_{strict} — providing comparability with prior nugget-based RAG evaluation work. The second family is *enabled by the per-option distractor structure* that the MCQ-to-probe pipeline preserves, and is the methodology's distinctive contribution: Generation Accuracy, the *Discrimination Gap* (DG), the False-Yes Rate on distractors (FYR), and the Abstention Rate. The two families align with two distinct cognitive tasks an MCQ poses to a student: knowing the answer and discriminating among expert-designed alternatives. We present the metric definitions in §3.2.1 (retrieval) and §3.2.2 (generation). Throughout this section, metrics are defined per-question; exam-level numbers reported in §5 are means over the evaluation set Q unless stated otherwise.

3.2.1 Retrieval metrics. The retrieval stage in a closed-domain RAG system is best read as *evidence collection for the generator* rather than as a ranked list for direct user consumption: given unlimited generation quality, could the correct answer in principle have been assembled from the retrieved context? The metrics below — standard nugget-based retrieval metrics from the TREC RAG community — are computed on the correct-option probe's retrieval $R_{o^*}^{(k)}$ and characterize how close that retrieval comes to a *minimal spanning document set*: the smallest collection of documents that jointly supports every nugget in $N(q)$. Distractor probes' retrievals are not evaluated here; their role in the methodology is to elicit per-option binary judgments scored by the generation metrics of §3.2.2.

Coverage@k. Fraction of nuggets supported by at least one document in the correct-option probe's retrieval:

$$\text{Coverage@}k(q) = \frac{1}{|N(q)|} \sum_{n_j \in N(q)} \sigma\left(R_{o^*}^{(k)}, n_j\right). \tag{1}$$

α-nDCG@k. Diversity-aware ranking metric [1]. Intuitively, α-nDCG rewards rankings that surface nugget-supporting documents diversely: at each rank, the gain from supporting an already-covered nugget is geometrically discounted by $(1 - \alpha)$, so retrieval that surfaces the same fact repeatedly is penalized relative to retrieval that surfaces distinct facts early. Let $R_{o^*}^{(k)} = (d_1, \ldots, d_k)$ denote the correct-option probe's ranked retrieval. For each rank r, let $c_j(r) = \sum_{\ell < r} \sigma(d_\ell, n_j)$ count prior coverage of nugget n_j by documents above rank r, and define the gain at rank r as $G(r) = \sum_{n_j \in N(q)} \sigma(d_r, n_j)\,(1 - \alpha)^{c_j(r)}$. Then

$$\alpha\text{-DCG@}k(q) = \sum_{r=1}^{k} \frac{G(r)}{\log_2(r + 1)}, \tag{2}$$

with α-nDCG@$k(q)$ normalizing by the diversity-greedy ideal α-DCG over the same document set at rank k. We use $\alpha = 0.5$, common in nugget-based evaluation.

Recall@k. Standard IR recall on the correct-option probe's retrieval, with relevance defined at the nugget-support level following standard nugget-based evaluation practice: a document d is judged relevant to q iff it supports at least one nugget in $N(q)$. Let $D_{\text{rel}}(q) =$

$\{d \in \mathcal{D} : \exists\, n_j \in N(q),\ \sigma(d, n_j) = 1\}$. Then

$$\text{Recall@}k(q) = \frac{|R_{o^*}^{(k)} \cap D_{\text{rel}}(q)|}{|D_{\text{rel}}(q)|}. \tag{3}$$

Computing $D_{\text{rel}}(q)$ exhaustively over the corpus would require $|\mathcal{D}| \cdot |N(q)|$ entailment calls per question, which is prohibitive; following standard TREC pooling practice [24], $D_{\text{rel}}(q)$ is approximated via a *judgment pool* formed from the union of retrieved sets across all probes and all evaluated configurations. Documents that no configuration ever retrieves are excluded from the pool; observed Recall@k is therefore an upper bound on true recall over the full corpus, with the gap shrinking as more diverse configurations contribute to the pool.

3.2.2 Generation metrics. The generation metrics measure, conditional on the retrieved context, whether the generator turns the retrieval into a correct, well-grounded judgment about each option. We use one nugget-recall metric (A_{strict}, defined below) and four that the per-option distractor structure of the MCQ-to-probe pipeline uniquely enables: Generation Accuracy, the *Discrimination Gap* (DG), the False-Yes Rate on distractors (FYR), and the Abstention Rate.

Generation Accuracy. Fraction of probes where the system's judgment matches its expected truth value, aggregated over all $n|Q|$ probes:

$$\text{Gen. Acc.}(Q) = \frac{1}{n|Q|} \sum_{q \in Q} \sum_{i=1}^{n} \mathbf{1}\big[y_i(q) = y_i^*(q)\big], \tag{4}$$

where $y_i(q) \in \{\text{Yes}, \text{No}, \perp\}$ is the system's judgment on probe p_i and the target $y_i^*(q) = \text{Yes}$ if option i is the correct answer $o^*(q)$, otherwise No. Since $y_i^*(q) \in \{\text{Yes}, \text{No}\}$, an abstention ($y_i = \perp$) never matches the target and is therefore counted as incorrect; abstention behavior is reported separately as the Abstention Rate below. Gen. Acc. admits trivial-strategy baselines that score above zero without doing any discriminative work: an always-Yes system scores $1/n$ (e.g., 25% with $n = 4$), and an always-No system scores $(n-1)/n$ (75%). The DG metric below collapses to zero on both, which is why DG complements Gen. Acc.

Discrimination Gap (DG). Direct measure of the asymmetry between the system's response to the correct option and to expert-designed distractors. Intuitively, DG asks whether the system distinguishes the correct option from the distractors: a perfect discriminator says Yes to the correct option and No to all distractors (DG = 1), while a system saying Yes (or No) to everything scores DG = 0. Per question:

$$\text{DG}(q) = \mathbf{1}[y_{o^*}(q) = \text{Yes}] - \frac{1}{n-1} \sum_{i \neq o^*} \mathbf{1}[y_i(q) = \text{Yes}], \tag{5}$$

with $\text{DG}(Q) = \mathbb{E}_q[\text{DG}(q)] \in [-1, 1]$. Under the indicator $\mathbf{1}[y_i = \text{Yes}]$, abstentions evaluate to 0 on both terms, so a system that uniformly abstains also scores DG = 0 — DG cannot distinguish a uniformly cautious system from a uniformly credulous one without the companion Abstention Rate. A perfect discriminator scores 1; a system that says Yes (or No) to every probe scores 0; a system more credulous on distractors than on the correct option scores below 0. Where Gen. Acc. aggregates correctness over both directions into one number, DG isolates the directional discriminative behavior the MCQ's distractor structure is designed to elicit, and is the metric whose name maps directly to the methodology's discriminative-power claim. Two systems can share the same Gen. Acc. yet have opposite-signed DG — e.g., a system that accepts the correct option and two of three distractors versus one that rejects the correct option and accepts one distractor both score Gen. Acc. = 50% ($n = 4$), but the first has DG = +1/3 (right-direction asymmetry) while the second has DG = −1/3 (anti-discriminating). The two metrics are not redundant.

False-Yes Rate on Distractors (FYR). Per-probe rate at which the system accepts a distractor claim as supported by the corpus:

$$\text{FYR}(Q) = \frac{1}{(n-1)|Q|} \sum_{q \in Q} \sum_{i \neq o^*} \mathbf{1}[y_i(q) = \text{Yes}]. \tag{6}$$

FYR localizes what DG aggregates: a given DG of, say, 0.5 is consistent with (correct-Yes = 1.0, distractor-Yes = 0.5) or with (correct-Yes = 0.7, distractor-Yes = 0.2), and these reflect operationally distinct behaviors (a credulous system that nevertheless accepts the correct option vs. a cautious system that selectively rejects distractors). FYR is the absolute distractor credulity rate, directly testing the property that expert-designed distractors are designed to probe.

Abstention Rate. Fraction of probes for which the system declines to commit to a Yes/No judgment:

$$\text{Abstain Rate}(Q) = \frac{1}{n|Q|} \sum_{q \in Q} \sum_{i=1}^{n} \mathbf{1}[y_i(q) = \perp]. \tag{7}$$

The generator returns $\perp$ when the retrieved context is insufficient to support a confident judgment; the rate at which this happens is a property of the retriever–generator pair, since the same generator may abstain under sparse retrieval and commit under richer retrieval. We track abstention because the synthetic-vs-exam comparison in §5.3 shows it is one of the metrics where the two evaluation regimes diverge most sharply, exposing system behavior that synthetic-only evaluation hides.

A_{strict} (Pradeep et al. [13], adapted). Fraction of nuggets supported by the system's response, following the TREC RAG 2024 nugget-recall convention [13]:

$$A_{\text{strict}}(q) = \frac{1}{|N(q)|} \sum_{n_j \in N(q)} \mathbf{1}\big[\text{answer}(q) \text{ supports } n_j\big], \tag{8}$$

where answer(q) is the system's natural-language explanation accompanying its judgment for the correct option's probe p_{o^*}. A_{strict} is evaluated only on the correct-option probe because nuggets are defined relative to the correct answer; distractor explanations are scored by the binary-judgment metrics (Gen. Acc., DG, FYR) rather than by nugget recall. The metric is structurally identical to Pradeep et al.'s A_{strict} (nugget recall against the system's text), but the artifact being judged differs: in their setting it is the system's free-form open-ended answer to the original question; in our MCQ-probing setting it is the shorter justification accompanying the binary Yes verdict on the correct-option probe. Because A_{strict} scores natural-language explanations, it is sensitive to the prompt eliciting them; we fix the prompt across configurations so that relative comparisons remain meaningful. Comparing A_{strict} to Coverage@k characterizes the retrieval-to-generation transmission on fact-stating content: a gap Coverage@$k - A_{\text{strict}} > 0$ indicates the retriever surfaced

support the generator failed to use; a gap in the other direction indicates the generator asserted nuggets that retrieval did not support — though on corpora where the exam requires derivation rather than fact recall, the latter direction may also reflect productive generator reasoning rather than transmission error.

3.3 Worked Example

We now compute the metrics end-to-end on the running Parkinson's MCQ of §1 (traced through the pipeline in Figure 1), making the formal definitions of §3.2 concrete. The MCQ decomposes into four nuggets (n_1: Parkinson's disease causes loss of dopamine-producing neurons; n_2: dopamine-producing neurons are located in the substantia nigra; n_3: loss of substantia-nigra dopamine neurons reduces basal-ganglia signaling; n_4: basal-ganglia signaling is needed for smooth voluntary movement). With $k = 3$, the per-probe retrievals from the course corpus are $R_A^{(3)} = (d_1, d_3, d_5)$, $R_B^{(3)} = (d_1, d_3, d_2)$, $R_C^{(3)} = (d_4, d_1, d_5)$, $R_D^{(3)} = (d_2, d_1, d_3)$, with pool $R^{(3)}(q) = \{d_1, \ldots, d_5\}$ and support values $\sigma(d, n_j)$ in Table 1. The system's verdict pattern is $(y_A, y_B, y_C, y_D) = (\text{No}, \text{Yes}, \text{Yes}, \text{No})$ against the target $(y_A^*, y_B^*, y_C^*, y_D^*) = (\text{No}, \text{Yes}, \text{No}, \text{No})$ — a correct acceptance of p_B, correct rejections of p_A and p_D, and a *false-yes* on the expert-designed distractor p_C.

Table 1: Support values $\sigma(d, n_j)$ over the pool $R^{(3)}(q)$ for the worked-example question.

Doc	n_1	n_2	n_3	n_4
d_1 (lecture: dopamine and Parkinson's pathology)	✓	–	–	–
d_2 (slide: basal ganglia dysfunction in Parkinson's)	–	–	✓	✓
d_3 (reading: dopamine in motor disease)	✓	–	–	–
d_4 (reading: cerebellum and motor coordination)	–	–	–	–
d_5 (slide: neurodegenerative disease overview)	–	–	–	–

Retrieval metrics. On p_B's retrieval $R_{o^*}^{(3)} = (d_1, d_3, d_2)$:
Coverage@3. From Table 1, n_1 is covered by $\{d_1, d_3\}$, n_3 and n_4 are covered by $\{d_2\}$, and n_2 is uncovered — a typical course-corpus gap where the anatomical anchor is implicit in the materials. Coverage@3 = 3/4.
α-nDCG@3. The metric additionally penalizes the ranking because d_3 repeats support for n_1 before d_2 contributes two new nuggets (n_3, n_4). With $\alpha = 0.5$, this gives α-nDCG@3 ≈ 0.80 — a signal Coverage@3 alone does not capture. Recall@3 is corpus-relative and reported course-level in §5.1.

Generation metrics. The false-yes on p_C reflects the distractor's surface keyword overlap (cerebellum + dopamine + motor) being accepted despite d_1's directionality cue, because n_2 (the substantia nigra anchor) is uncovered.
Gen. Acc. Matches across probes: A (No = No) ✓, B (Yes = Yes) ✓, C (Yes ≠ No) ✗, D (No = No) ✓. Gen. Acc. = 3/4.
DG.$\text{DG}(q) = \mathbf{1}[y_B = \text{Y}] - \frac{1}{3}(\mathbf{1}[y_A = \text{Y}] + \mathbf{1}[y_C = \text{Y}] + \mathbf{1}[y_D = \text{Y}]) = 1 - \frac{1}{3}(0 + 1 + 0) = 2/3$.
FYR. $\frac{1}{3}(0 + 1 + 0) = 1/3$ — one false-yes across three distractors.
A_{strict}. p_B's natural-language explanation asserts n_1, n_3, n_4 but not n_2, giving $A_{\text{strict}} = 3/4$, matching Coverage@3 (no transmission gap on this question).

Traditional MCQ-accuracy scoring would mark this question correct, hiding the p_C discrimination event entirely; the methodology reports it independently, and FYR aggregated course-wide would surface a systematic distractor-credulity pattern if one existed.

4 Experimental Setup

Target system. We evaluate the digital tutor [14], a production RAG-based digital tutor deployed at a large research university serving students across multiple faculties. The system retrieves from course-specific corpora of lecture transcripts, slides, and reading materials and generates responses using a large language model conditioned on the retrieved context.

Courses and corpora. We evaluate three courses spanning distinct knowledge domains:
Introduction to Psychology (MIT OCW 9.00SC), drawn from MIT OpenCourseWare. We use the entire 2011 final exam (42 questions) as the primary evaluation, plus the entire 2010 final exam (32 questions, run against the 2011 corpus) for the year-to-year stability check (§5.5). This course is our publicly reproducible benchmark. MIT OCW may appear in LLM pretraining; the deployed RAG system grounds its judgments in retrieved context by design, and the two non-public corpora below serve as additional contamination checks.
Calculus 1B (Anonymous University), an undergraduate calculus course covering sequences, series, limits, and real analysis. We use the entire final exam (23 questions). The course represents knowledge that is highly relational and conditional: correct answers depend on understanding the direction and scope of mathematical claims rather than isolated fact recall.
Cost-Effectiveness in Health and Medicine (Anonymous University), an undergraduate course in health economics and policy evaluation, with lecture materials and assigned readings. We use the entire final exam (20 questions). The course occupies an intermediate position between factual-declarative and relational knowledge, including both definitional content and policy reasoning.

Datasets. For each course we construct two evaluation datasets. The *exam-based* dataset draws MCQs directly from official course exams that were administered to enrolled students at MIT and the anonymous university; each evaluation set is the complete final exam for the course. We do not author or modify any items, only nuggetize each correct answer via the open-source Nuggetizer library [13] and rephrase each option into a probe per the pipeline of §3. The *synthetic* dataset is generated from the same corpus with RAGAS [2] configured with persona-conditioned question generation over a corpus-derived knowledge graph rather than a naive single-passage generator. On the Calculus corpus the RAGAS knowledge-graph construction failed because the corpus consists of lecture transcripts dense in mathematical symbols, derivations, and verbal transitions, from which knowledge-graph construction cannot extract clean factual triples; the pipeline fell back to passage-local generation. We discuss this in §5.3: the resulting weakness of the synthetic baseline on Calculus reflects a structural limitation that no synthetic-question generator can fully escape, since all of them derive questions from the corpus and therefore cannot produce questions that require derivation beyond what the corpus

directly states. Table 2 summarizes statistics. Most MCQs use $n = 4$ answer options; a small number of Calculus items have $n = 5$.

Table 2: Dataset statistics. $|\mathcal{D}|$ = number of corpus documents. †Used only for the year-to-year stability check against the 2011 corpus.

Course	$\lvert\mathcal{D}\rvert$	Exam $\lvert Q\rvert$	Avg. $\lvert N(q)\rvert$	Synth. $\lvert Q\rvert$
Psych. 2011 exam	287	42	1.7	50
Psych. 2010 exam†	287	32	1.9	—
Calculus 1B	94	23	1.4	69
Health Econ.	168	20	3.1	30

Configurations. The default (baseline) configuration uses the system's production hybrid retrieval (combining dense vector search with BM25 sparse retrieval) with cosine similarity on the dense side, $k = 7$ retrieved documents per probe, Voyage-3 embeddings (1024-dim), 1024-token chunks, and Claude Sonnet 4.5 as the generator. Entailment scoring uses GPT-4o under the same structured binary-judgment prompt as FreshStack [19]. For sensitivity analysis we vary one axis at a time relative to baseline: retrieval depth $k \in \{3, 7, 10, 20\}$; retrieval method (hybrid vs. dense-only vs. sparse-only); embedding dimension (Voyage-3 1024-dim baseline vs. Voyage-3 256-dim projection); chunk size (1024-token baseline vs. 256-token); generator LLM (Sonnet 4.5 vs. Haiku 3.5); and corpus extent (default vs. extended with the course's scanned textbooks). A random-retriever condition (uniform document sampling) provides a negative control.

5 Results

The results section reports the methodology's empirical behavior on our instantiated system. §5.1 profiles three courses sharing an identical retriever and generator, showing how the methodology's metrics expose structurally different system behaviors across corpora. §5.2 validates that the metrics are *sensitive* to RAG configuration changes and *specific* to the component varied. §5.3 contrasts exam-based with synthetic evaluation on the same system, showing that the methodology's metrics diverge between regimes in ways that synthetic-only evaluation would have hidden. §5.4 provides a single-deployment external-validity check on user queries. §5.5 closes with a year-to-year benchmark-stability check supporting the public benchmark release.

5.1 Cross-Domain Profiles

Table 3 reports baseline metrics for the three courses under an *identical* retrieval configuration ($k = 7$, same generator, same entailment judge).

Moving from Psychology to Calculus, Coverage@7 collapses from 76.8% to 20.0%, Gen. Accuracy drops from 82.1% to 6.5%, and the abstain-per-probe rate rises from 0.0% to 68.5%; Health Economics sits between the two on every metric. The methodology's metrics surface several distinct facets of this spread that a single accuracy number would conflate. On Psychology, Gen. Accuracy exceeds Coverage (82.1% vs. 76.8%): the system answers correctly on more probes than direct retrieval support would predict, suggesting that the generator contributes beyond retrieval — through reasoning, parametric knowledge, or interpolation across the retrieved set. On Health Economics, Gen. Accuracy tracks Coverage closely (52.0% vs. 53.0%), indicating no surplus generator contribution beyond what retrieval provides. DG separates two operationally distinct failure modes that a single accuracy number conflates: *anti-discrimination*, where the system is more credulous on distractors than on the correct option (DG < 0 — e.g., it says No to the correct claim while accepting some distractor as supported; the random-retriever control in §5.2 exhibits this, with DG $= -0.05$), and *non-discrimination*, where the system refuses to commit either way (DG ≈ 0 via mass abstention). On our courses, the hard end (Calculus) is non-discrimination: DG is 0.04 not because the system is being fooled by distractors, but because it almost never accepts the correct option and abstains on most probes. FYR confirms this from the distractor side — distractor credulity is only 2%, so the dominant failure mode is not false acceptance of distractors but refusal to engage with the option set.

In our three-course instantiation, the same retriever, the same generator, and the same evaluation pipeline produce this spread. The structurally different signal profiles across courses suggest that, in our setting, the position of the exam content on a corpus-specific axis — roughly, how often the exam's answers appear in the corpus directly versus must be derived from it — shapes the methodology's output strongly. We frame this as one signal the methodology surfaces on the systems we evaluated, not as a universal property of closed-domain RAG; cross-system generalization is future work.

The magnitude of this variation under identical pipeline choices was larger than expected and translates directly to what students experience from the deployed tutor: Psychology users encounter a system that answers most queries with grounded confidence; Calculus users of the same tutor face one that refuses to commit on more than two thirds of its option-level probes. The methodology's metrics surface this course-by-course fit at evaluation time, identifying which courses the system serves well and which it does not before deployment exposes the gap to users.

5.2 Sensitivity and Specificity

This section validates the methodology's diagnostic utility: when we change a component of the RAG system, do the metrics reflect the change in interpretable ways? A useful methodology must be both *sensitive* (metrics respond when the system changes) and *specific* (movements localize to the component that changed); without both, the metrics can detect that something is different but cannot attribute the change to a particular pipeline component. We test these properties on Psychology 2011 by holding the corpus and the rest of the pipeline fixed and varying one component at a time. Table 4 reports a compact view; the random-retriever negative control collapses every retrieval metric to zero and Gen. Accuracy to 27%, near the always-Yes floor of $1/n = 25\%$.

The configuration ablations move the relevant per-component metrics as expected (retrieval-side changes shift Coverage@k and α-nDCG@k; generator swap shifts Gen. Acc. while leaving retrieval metrics unchanged). The Discrimination Gap is the most informative of these: DG tracks the Gen. Acc. trend on k (0.58 $\rightarrow$ 0.69) and falls 21 points on the Haiku-3.5-generator swap (0.66 $\rightarrow$ 0.45) while retrieval metrics are unchanged — localizing the regression

Table 3: Cross-domain baseline on exam-based MCQs at identical retriever and generator ($k = 7$). The same RAG system produces qualitatively different signal profiles across courses.

Course	Cov.@7	α-nDCG@7	Recall@7	A_{strict}	Gen. Acc.	DG	FYR	Abstain
Psychology 2011	76.8%	0.81	0.92	74.2%	82.1%	0.66	19%	0.0%
Health Econ.	53.0%	0.61	0.71	51.4%	52.0%	0.45	11%	28.6%
Calculus 1B	20.0%	0.28	0.34	18.3%	6.5%	0.04	2%	68.5%

Table 4: Sensitivity & specificity on Psychology 2011, with each variant differing from the baseline ($k = 7$, Voyage-3 1024-dim, 1024-token chunks, Claude Sonnet 4.5) along a single axis. Bold marks the baseline row.

Configuration	Cov.@k	α-nDCG	Gen. Acc.	A_{strict}	DG
Baseline	**76.8%**	**0.81**	**82.1%**	**74.2%**	**0.66**
Retrieval depth					
$k = 3$	68.0%	0.74	78%	65.3%	0.58
$k = 10$	81.9%	0.84	84%	78.4%	0.69
$k = 20$	81.5%	0.82	83%	77.9%	0.67
Embedding dimension					
Voyage-3 (256-dim)	75.3%	0.78	82%	72.1%	0.65
Chunking strategy					
256-token chunks	72.0%	0.75	79%	69.1%	0.61
Generator LLM					
Haiku 3.5 generator	76.8%	0.81	69.6%	65.8%	0.45
Corpus / negative control					
Extended materials	82.5%	0.85	84%	79.8%	0.70
Random retriever	0.0%	0.00	27%	0.0%	−0.05

to the generator's discriminative behavior, not its retrieval. The random-retriever control collapses DG to slightly negative (−0.05), confirming that without useful retrieved context the system cannot tell correct options from expert-designed distractors. DG thus responds in the right direction to every axis, satisfying the same sensitivity-and-specificity criterion as the rest of the methodology's metrics.

5.3 Synthetic vs. Exam-Based Evaluation

We compare exam-based metrics with synthetic RAGAS-generated questions over the same corpora and identical RAG configuration. The two regimes broadly agree on the *direction* of sensitivity under configuration changes (both correctly register k-variation, model swaps, and the random control). They disagree, however, in the readings they produce on common evaluation metrics, in ways that are structured by which corpus is being evaluated. Table 5 summarizes.

The headline observation is that the methodology's metrics expose corpus-system fit problems that synthetic evaluation does not. On Calculus, abstention under exam-based evaluation is 68.5%, against only 8.7% under synthetic — a practitioner relying on synthetic evaluation alone would observe a system that commits on ∼91% of probes and achieves 70.7% generation accuracy, while exam-based evaluation reveals that the same system declines to commit on more than two thirds of exam-derived probes and achieves only 6.5% generation accuracy. This gap is the result the synthetic-vs-exam comparison is built to surface, and it is not localizable from retrieval-side metrics alone: retrieval metrics show that the exam setting is harder (e.g., Calculus Coverage@7 drops from 38.4% under synthetic to 20.0% under exam-based), but not whether the failure is abstention, distractor credulity, or correct-option rejection. Coverage@7 on Psychology differs by only −4.8 pp between regimes, consistent with the embedding-favored retrievability of corpus-derived questions, but does not on its own signal that the regimes are producing materially different pictures of system behavior.

The mechanism is structural rather than implementation-specific. Synthetic questions are generated from the corpus and can only ask about what the corpus directly states. Expert-written exams are authored about the material, can ask about content the corpus is supposed to teach but does not directly state, and come with expert-designed distractors that test discriminative reasoning rather than fact lookup. On Calculus the contrast is sharpest: RAGAS could not build its knowledge graph over the Calculus corpus and fell back to passage-local question generation, producing notably easier questions — but the failure mode generalizes. A passage-local generator would degrade silently to the same easy-question regime; the constraint applies to passage-derived synthetic generators broadly (KG-based, passage-local, instructor-style LLM generation, adversarial-distractor approaches), since all derive their questions from the indexed corpus and therefore cannot construct conditional-reasoning questions that are not statable from any single passage. Health Economics presents a different relationship between the two regimes, but the direction of the broader observation is unchanged: synthetic and exam-based evaluation are complementary rather than interchangeable — synthetic reports on retrieval over corpus-derived content, exam-based reports on curriculum-aligned discriminative behavior — and the methodology's metrics, particularly DG and Abstention, are where the divergence is visible.

5.4 Relevance to Real User Queries

Section 5.3 showed that synthetic and exam-based evaluation produce substantively different readings of system behavior on the same RAG system. This raises a deployment-relevance question: which of the two sets of queries do real user queries more closely resemble? If user queries align with exam-derived questions, the exam-based metrics are the deployment-relevant ones; if they align with synthetic, synthetic is the closer match. We test this on the production-deployed Calculus 1B course — the case where the synthetic–exam divergence is largest (§5.3) and where the question is most pointed.

Table 5: Synthetic vs. exam-based evaluation under an identical RAG configuration ($k = 7$, Voyage-3 1024-dim, Claude Sonnet 4.5). Δ = Exam − Synth (sign chosen so positive = exam exposes harder behavior).

	Psychology			Health Econ.			Calculus 1B		
Metric	Exam	Synth	Δ	Exam	Synth	Δ	Exam	Synth	Δ
Coverage@7	76.8%	81.6%	−4.8	53.0%	65.1%	−12.1	20.0%	38.4%	−18.4
α-nDCG@7	0.81	0.85	−0.04	0.61	0.72	−0.11	0.28	0.46	−0.18
A_{strict}	74.2%	78.5%	−4.3	51.4%	63.7%	−12.3	18.3%	36.1%	−17.8
Gen. Accuracy	82.1%	89.0%	−6.9	52.0%	81.4%	−29.4	6.5%	70.7%	−64.2
DG	0.66	0.78	−0.12	0.45	0.62	−0.17	0.04	0.21	−0.17
FYR	19%	11%	+8.0	11%	7%	+4.0	2%	3%	−1.0
Abstention Rate	0.0%	0.0%	0.0	28.6%	5.2%	+23.4	**68.5%**	**8.7%**	**+59.8**

Setup. We collected 150 raw student queries from the system's Calculus 1B deployment and excluded social and system messages (greetings, presence checks, and other queries that did not invoke the RAG pipeline), leaving $|U| = 80$ substantive course queries. For each query $u \in U$ we computed its cosine similarity to the nearest exam-derived question e in the original Calculus 1B final-exam stems set E, and the analogous similarity to the nearest RAGAS-generated synthetic question $s \in S$. Each query was classified as *exam-like* or *synthetic-like* by two independent methods: an embedding classifier (assigning the verdict with the higher cosine) and an LLM-as-judge classifier shown all three sets and prompted to judge style, intent, complexity, length, language, and content alignment per query. Table 6 reports results.

Table 6: User-query relevance on Calculus 1B ($|U| = 80$ substantive queries).

Quantity	Value
Mean cosine to exam set E / synthetic set S	0.590 / 0.416
Paired difference Δ (per query)	$\mathbf{0.174}$, $p = 0.002$
Embedding-classifier exam-like verdict	50/80 (62.5%)
LLM-classifier exam-like verdict	50/80 (62.5%)
Inter-classifier agreement	**80/80 (100%)**

Qualitative examples. The classification split is intuitive in practice. Queries about applying or computing on the material tend to align with exam-derived questions, while definitional or lookup queries tend to align with synthetic. Representative examples from the deployment:

- *Exam-like:* "Evaluate $\sum_{n=1}^{\infty} \frac{1}{n^2 \log n}$ for convergence."
- *Synthetic-like:* "What is the definition of a Cauchy sequence?"

Interpretation. On this one application-oriented deployment, real student queries align systematically more with exam-derived questions than with corpus-derived synthetic questions ($\Delta = 0.174$, $p = 0.002$): exam-based evaluation is closer to what students actually ask than synthetic generation is. A minority of queries align with synthetic (the definition-lookup pattern above); both regimes capture some real user behavior, but exams capture the dominant one. This demonstration is scoped to one deployment; we do not extend the claim to deployment contexts that are not application-oriented (casual conversation, operational support, open-domain lookup).

5.5 Year-to-Year Stability

A benchmark whose metrics drift sharply with the choice of exam year would be a poor target for future comparison. We test stability by running the 2010 Psychology exam (32 questions) against the same 2011 corpus used for our baseline evaluation and comparing to the 2011 exam (42 questions) results. The 2010 run yields Coverage@7 = 78.3% vs. 76.8% on 2011 (+1.5 pp on the headline retrieval metric); Gen. Acc. = 84.4% vs. 82.1% (+2.3 pp); DG = 0.68 vs. 0.66 (+0.02). The methodology therefore produces consistent signal across exam years for the same corpus across all three headline metrics, suggesting that exam-specific idiosyncrasies do not dominate the evaluation. This stability is a precondition for the publicly released benchmark and supports rotating exam years over time as a contamination-mitigation strategy without invalidating year-to-year comparisons. More broadly, on courses with stable year-over-year syllabi, this is exactly how students themselves prepare: they study with prior-year exams against the current course material and are expected to be able to solve them. The RAG system, given the same corpus, is asked to do the same task.

6 Discussion

Component-level diagnosis. A conventional end-to-end accuracy number is sufficient to rank systems but insufficient to localize *why* one outperforms another. The configuration ablations make this concrete. Replacing Sonnet 4.5 with Haiku 3.5 as the generator drops Gen. Accuracy by 12.5 pp and DG by 21 points while leaving retrieval metrics essentially unchanged; an end-to-end metric would register the regression without locating it. The inverse holds for the Voyage-3 256-dim projection, which shifts retrieval metrics while Gen. Accuracy stays at 82%. The methodology's metric set localizes the regression in each case to the responsible component. Abstention, reported alongside, adds a further diagnostic axis: on Calculus the system does not merely answer incorrectly — under exam-based evaluation it refuses to answer 68.5% of the time, against only 8.7% under synthetic, a qualitatively different failure mode a practitioner must address differently from low accuracy and one synthetic-only evaluation would have masked. Production usage on the deployed Calculus tutor is qualitatively consistent

with this pattern: many real student queries receive non-answers rather than substantive responses, matching what the probe-level abstention measurement predicts.

Exam-answerability contract. MCQ exam questions come with a built-in answerability guarantee: instructors design exams to be answerable from the course material together with the student's reasoning. This contract obviates the need for the oracle-retriever upper-bound evaluation that Thakur et al. [19] use to estimate retrieval headroom from gold-answer queries. Whether the corpus actually captures that material is itself a property the methodology surfaces — corpus gaps appear as evaluation failures rather than precondition violations.

Stated vs. derivable, and identifying course–system fit. A priori, a practitioner cannot easily predict which courses a deployed RAG system will serve well: under an identical retriever, generator, and prompting pipeline, the same system delivers grounded confident answers on Psychology and abstention-dominated failure on Calculus. The cross-domain spread (§5.1) showed that these qualitatively different profiles align with a corpus-specific axis — roughly, how often the exam's answers are stated in the corpus versus only derivable from it. More importantly, the methodology gives practitioners a tool for identifying which courses are well-served by the existing RAG system (high Coverage, high DG, low abstention) and which are not (low Coverage, abstention-dominated, or distractor-credulous) — a per-course fit signal that single-corpus end-to-end accuracy cannot provide, and one that helps prioritize where corpus improvement or methodological change is needed.

Practical recommendation. For practitioners choosing between exam-based and synthetic RAG evaluation, our recommendation turns on two factors: artifact availability and deployment-query alignment. *When MCQs are available and real user queries resemble exam-style questions* (typically: education, professional certification, medical licensing), our methodology is the stronger choice: it leverages a curriculum-aligned, expert-validated evaluation signal, and the Discrimination Gap and per-option Yes/No structure surface distractor-credulity failures that retrieval-side metrics and end-to-end accuracy both miss. *Otherwise* — when MCQs are unavailable, or when real user queries diverge from exam-style content even with MCQs available — synthetic generation retains substantial value as a low-overhead default. In practice, the two regimes probe different query distributions — exam-based tests application-oriented reasoning while synthetic tends toward definitional/lookup tasks (since synthetic queries derive from corpus passages) — so MCQs alone may not capture all real user behavior. Our Calculus deployment illustrates this: a minority of real student queries (definitional lookups) fell outside the exam-aligned class but were captured by synthetic, so running both regimes is valuable even when MCQs are available.

Limitations. (1) Two of the methodology's three LLM components — nugget extraction (via the Pradeep et al. Nuggetizer) and support judgment (via the GPT-4o protocol of Thakur et al.) — have prior human-validation studies on technical-documentation corpora that we inherit [13, 19, 20]; domain transfer to educational content and validation of the option-rephrasing step (introduced here) are future work. (2) We instantiate the methodology on a single deployed RAG system across three courses; cross-system generalization of the empirical profiles is future work. (3) We do not disentangle whether low-Coverage failures reflect genuinely missing corpus evidence or cases where the evidence exists but requires reasoning beyond the current prompt; separating these factors is future work. (4) MCQs are expert-written probes, not natural user queries; the user-query alignment of §5.4 is validated on one deployment, and complementing MCQ-based measurement with query-log-based measurement is appropriate where deployment data is available. (5) The evaluation signal is only as strong as the source exam: ambiguous questions, weakly designed distractors, or factual errors propagate directly into the metrics.

7 Conclusion

We introduced an evaluation methodology for closed-domain RAG that repurposes expert-written multiple-choice exams as a ready-made evaluation signal. The methodology has two intertwined contributions. First, it enables RAG evaluation on artifacts that already exist in the deployment context — expert-written, instructor-validated exams produced by the same authority that defines what users should be able to learn from the corpus. Synthetic generation and exam-based evaluation answer different questions about a system: synthetic asks whether the system can retrieve and use what its corpus directly contains; exam-based asks whether the system handles the curriculum-aligned discriminative tasks the corpus is supposed to teach. Both questions matter; neither substitutes for the other. Second, the per-option pipeline together with the *Discrimination Gap* provide a metric set that measures discriminative judgment between expert-designed alternatives — a property unique to MCQs and absent from open-ended evaluation. A system can score high on standard retrieval and nugget-recall metrics yet still be fooled by plausible distractors; DG and Generation Accuracy surface exactly this failure mode.

We instantiated the methodology on a production RAG-based digital tutor across three undergraduate courses and release a reproducible benchmark on MIT OpenCourseWare's Introduction to Psychology. In our instantiation, exam-based and synthetic evaluation produced complementary but non-interchangeable signals: synthetic questions better reflected corpus-derived lookup behavior, while exam-derived probes exposed application-oriented failures, abstention patterns, and distractor credulity that synthetic-only evaluation would have hidden. The contrast was especially pronounced in Calculus, where exam-derived probes revealed a course-system fit problem that was largely masked by synthetic evaluation.

These findings suggest a practical role for expert-written MCQs in closed-domain RAG evaluation. When such artifacts are available and resemble the tasks users expect the system to support, they provide a reusable, expert-validated benchmark for measuring not only whether the system retrieves relevant evidence, but whether it can discriminate correct claims from plausible alternatives. MCQ-based evaluation should therefore complement, not replace, synthetic and query-log-based evaluation: together, these regimes give practitioners a fuller picture of how a deployed RAG system behaves across corpus-derived lookup, curriculum-aligned reasoning, and real user demand.

GenAI Usage Disclosure

The authors used ChatGPT (OpenAI) to improve the clarity, grammar, and readability of selected author-written passages. The tool was not used to generate research ideas, methods, experimental results, analyses, or conclusions. All suggested revisions were reviewed and edited by the authors, who take full responsibility for the final content of the paper.

References

[1] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 659–666, 2008.

[2] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL): System Demonstrations*, pages 150–158, 2024.

[3] Thomas M. Haladyna, Steven M. Downing, and Michael C. Rodriguez. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3):309–333, 2002.

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

[5] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[6] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 15696–15707, 2023.

[7] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.

[8] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522, 2023.

[9] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10862–10878, 2024.

[10] Andrew M. Olney. Generating multiple choice questions from a textbook: LLMs match human performance on most metrics. In *Proceedings of Empowering Education with LLMs — The Next-Gen Interface and Content Generation (AIED 2023 Workshop)*, CEUR Workshop Proceedings, Tokyo, Japan, July 7, 2023.

[11] Christian Grévisse, Maria Angeliki S. Pavlou, and Jochen G. Schneider. Docimological quality analysis of LLM-generated multiple choice questions in computer science and medicine. *SN Computer Science*, 5(5):636, 2024.

[12] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning (CHIL)*, pages 248–260, 2022.

[13] Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. The great nugget recall: Automating fact extraction and RAG evaluation with large language models. *arXiv preprint arXiv:2504.15068*, 2025.

[14] Hila Reicher, Yarden Frenkel, Maor Juliet Lavi, Rami Nasser, Yuval Ran-Milo, Ron Sheinin, Mark Shtaif, and Tova Milo. A generative AI-empowered digital tutor for higher education courses. *Information*, 16(4):264, 2025.

[15] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 338–354, 2024.

[16] Manish Singh and Manish Shrivastava. Options-aware dense retrieval for multiple-choice query answering. *arXiv preprint arXiv:2501.16111*, 2025.

[17] Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O. Arik, Danqi Chen, and Tao Yu. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *International Conference on Learning Representations (ICLR)*, 2025.

[18] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[19] Nandan Thakur, Jimmy Lin, Sam Havens, Michael Carbin, Omar Khattab, and Andrew Drozdov. FreshStack: Building realistic benchmarks for evaluating retrieval on technical documents. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025. arXiv:2504.13128.

[20] Nandan Thakur, Ronak Pradeep, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. Support Evaluation for the TREC 2024 RAG Track: Comparing Human versus LLM Judges. *arXiv preprint arXiv:2504.15205*, 2025.

[21] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. UMBRELA: UMbrela is the (open-source reproduction of the) Bing relevance assessor. *arXiv preprint arXiv:2406.06519*, 2024.

[22] Jonas van Elburg, Peter van der Putten, and Maarten Marx. Can we evaluate RAGs with synthetic data? *arXiv preprint arXiv:2508.11758*, 2025.

[23] Ellen M. Voorhees. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, 2003.

[24] Ellen M. Voorhees. I come not to bury Cranfield, but to praise it. In *HCIR 2009: Bridging Human-Computer Interaction and Information Retrieval*, 2009.

[25] Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-tau Yih, and Xin Luna Dong. CRAG – comprehensive RAG benchmark. In *Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[26] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 535–546, 2021.

[27] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.