Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Assuring Institutional RAG for Compliance-Literacy Support: Replication Reliability, Warrant Stability, and Evidence-Driven Revision

Jacob Holster

jbh6331@psu.edu

Pennsylvania State University

Research Article

Keywords: retrieval-augmented generation, educational artificial intelligence, compliance literacy, replication reliability, warrant stability, assurance case

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11576603/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Assuring Institutional RAG for Compliance-Literacy Support: Replication Reliability, Warrant Stability, and Evidence-Driven Revision

Jacob Holster, Ph.D.

Assistant Professor of Music Education and AI

School of Music, College of Arts and Architecture, The Pennsylvania State University, University Park, PA, United States

Corresponding author. Jacob Holster, jbh6331@psu.edu, 101 East North Hills Place, State College PA, 16803, https://orcid.org/0000-0001-9366-1630.

Abstract

Preparing a human-subjects research protocol is a consequential but unevenly taught form of academic and professional writing. Researchers must translate study designs into ethical, regulatory, and institutional language, while general-purpose language models offer fluent assistance in a domain where fabrication and unsupported authority carry regulatory consequences. This article reports the design and technical assurance evaluation of IRB Helper, an internally deployed retrieval-augmented generation system intended to support researcher compliance literacy through institutionally grounded policy guidance. A static 60-item stress suite was executed 25 times across initial development and three replicated system configurations. Two prospectively specified revision cycles linked observed failures to targeted changes and then tested those changes under repeated execution. Retrieval-score reliability remained high across configurations, G(1) = .89–.93, while benchmark-score reliability remained moderate, G(1) = .61–.64. Retrieved-document overlap increased from .658 to .701 and cited-warrant overlap from .449 to .602, although near-ceiling pass rates concealed eight corpus-covered questions routed away from retrieval. No probable fabrication or detector false positive was observed across the 660 responses following allowlist regeneration. The cycles also showed that technically valid fixes can target the wrong pipeline layer, safety mechanisms can introduce new user-visible variability, and evaluation instruments require the same versioning and audit discipline as the systems they assess. The study contributes replication reliability and warrant stability as assurance properties for researcher-built, institutional-adjacent educational AI, therefore establishing technical and informational preconditions for subsequent evaluation with learners.

Keywords: retrieval-augmented generation; educational artificial intelligence; compliance literacy; replication reliability; warrant stability; assurance case

Assuring Institutional RAG for Compliance-Literacy Support: Replication Reliability, Warrant Stability, and Evidence-Driven Revision

1. Introduction

Prospective ethics review in the United States is rooted in documented failures of human-subjects protection. The National Research Act of 1974, the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979), and the Common Rule (Protection of Human Subjects, 2018) established a regulatory infrastructure intended to ensure that research involving human participants is justified, consented, and accountable. Few researchers dispute that charge. Whether the resulting oversight achieves its protective purpose is nonetheless difficult to establish. For instance, the U.S. Government Accountability Office (2023) concluded that federal agencies assess regulatory compliance without validated measures that define participant protection.

At the same time, the process through which investigators prepare studies for review remains unevenly supported. Empirical and critical accounts of institutional review boards (IRBs) have documented long and variable review timelines, inconsistent determinations across boards, as well as concerns about overreach and administrative costs (Abbott & Grady, 2011; Fost & Levine, 2007; Klitzman, 2015; Schneider, 2015; Silberman & Kahn, 2011). The IRB process can, however, function as a supportive framework for study conceptualization, helping researchers clarify procedures, risks, benefits, consent processes, and data protections before a study begins (Grady, 2015; Orimadegun, 2020; Protection of Human Subjects, 2018).

In this article, researcher *compliance literacy* refers to the capacity to translate a study design into ethical, regulatory, and institutional language while retaining the reasoning needed to justify those choices. Protocol writers must describe procedures, weigh risks and benefits, specify recruitment and consent processes, anticipate data-security concerns, and present the study in language legible to reviewers and participants. These demands require learning a situated institutional genre whose conventions, warrants, and consequences are not equally visible to all researchers. The steepest learning curve often falls on graduate students, investigators in unfunded fields, community-engaged researchers, and others with limited access to specialized regulatory expertise (Grohmann, 2025; Onakomaiya et al., 2023). Genre and academic-literacies perspectives further position protocol preparation as professional learning distributed across templates, training modules, informal peer networks, and access to institutional analysts (Lea & Street, 1998; Miller, 1984; Wingate, 2018).

Large language models (LLMs) present both an obvious and hazardous response to this need. Protocol preparation involves writing over a bounded but institutionally specific body of rules, making generative

assistance attractive. The same fluency, however, can conceal fabrication, unsupported inference, and misplaced authority. General-purpose models may produce plausible guidance without access to the policies governing a particular institution, and fabrication is especially consequential when a query concerns regulatory obligations (Dahl et al., 2024; Ji et al., 2023). An answer that misstates local policy or cites a nonexistent regulatory pathway can consume review resources and, if uncorrected, undermine the protections review is intended to provide. As such, this article reports the design and technical assurance evaluation of IRB Helper, an internally deployed RAG application intended to support human-subjects research preparation through institutionally grounded, source-linked guidance. Moreover, the study asks what evidence should be established before a researcher-built AI system can responsibly be evaluated as professional-learning infrastructure.

2. Background

2.1. Prospective ethics review and the preparation burden

Not all IRB burden occurs at the point of review. A substantial portion occurs before review begins, when investigators must translate a study idea into protocols, consent forms, recruitment materials, data-security plans, and responses to anticipated reviewer concerns. Further, administrative efficiency in IRB submission is not equivalent to effective participant protection or compliance literacy development as the process can invite ambiguity in analyst and researcher interpretations of materials. Considering the efficiency of LLMs, researchers must continue to learn compliance frameworks as faster preparation cannot distinguish between procedural substitution and genuine understanding (Parreco et al., 2019; Rosenfeld, 2020). The human factor, represented by intensive regulatory support, has been associated with shorter total turnaround among supported applications, yet receiving support at all showed no adjusted association with review time, and time saved reveals nothing about what investigators came to understand (Desai et al., 2017). The Consortium to Advance Effective Research Ethics Oversight (AEREO) has therefore argued that IRB quality should center participant perspectives and that review decisions should at minimum meet a basic standard of reasonableness by accounting for relevant perspectives, considering key factors, and providing defensible justifications (Anderson et al., 2023; Anderson et al., 2024). The same standard is a fair demand of the proposals investigators submit, and of any system that helps produce or verify them.

The Havasupai case illustrates the consequences of ambiguity. Biological samples provided by tribe members in the context of diabetes research were later used for studies of schizophrenia, migration, and inbreeding without participant knowledge, under a consent document whose broad reference to behavioral and medical disorders did not produce a shared understanding of those uses (Garrison & Cho,

2013). Formal documentation and institutional procedure, in this case, did not ensure shared understanding, community acceptability, or restriction of the samples to uses participants recognized as authorized. IRB chairs and researchers later described heightened attention to culturally sensitive consent and secondary use, but most did not report substantially changing their own approach to compliance work (Garrison & Cho, 2013). The 2018 revision of the Common Rule, which expanded exemption categories for much noninterventional social and behavioral research, then widened the space in which ethically consequential decisions occur without deliberation (Owczarzak & Smith, 2022).

The goal of protocol preparation support, then, is to help investigators make their study decisions and understand the ethical implications of those decisions. Further, a system that provides preparation support might center compliance literacy while providing an efficient drafting process. Compliance literacy may ultimately emerge from a pedagogical process where researchers learn the relationships among writing moves across their discipline and IRB expectations in relation to their approach to protocol and supporting documents, supported by available IRB guidance (Marcos Miguel & Noy, 2023). These accounts do not, however, indicate that completing a protocol develops compliance literacy.

Scaffolding and cognitive apprenticeship provide design lenses for distinguishing assistance from substitution (Collins et al., 1989; Collins & Kapur, 2014; Wood et al., 1976). An LLM-assisted system might surface relevant policy, explain section expectations, connect study decisions, and support revision before submission. It might instead produce a polished artifact while obscuring the reasoning its user needs to understand and defend. Regulatory assistance can range from answering questions to completing materials on an investigator's behalf (Desai et al., 2017), and these forms of support may generate similar documents while distributing cognitive and ethical work very differently. A defensible educational design must therefore make relationships among study design, institutional policy, ethical obligation, and written representation inspectable while preserving investigator responsibility.

2.2. Large language models, RAG, and assurance in compliance-critical learning

The rapid adoption of general-purpose LLMs in academic work (Kasneci et al., 2023; van Dis et al., 2023) has outpaced the development and evaluation of local domain applications. Hallucination, the generation of fluent but unfaithful or fabricated content, is well documented across natural-language-generation tasks (Ji et al., 2023), and its consequences scale with the authority attached to the generated text. Recent work has begun to examine LLM support on both sides of IRB practice. On the investigator side, Godwin et al. (2024) developed an open-source IRB Draft Generator that transforms brief structured inputs—including the study hypothesis, inclusion and exclusion criteria, and design type—

into a draft human-subjects protocol using an LLM and a literature-search component. The system was proposed as a means of reducing clerical burden and increasing application consistency, although the report did not evaluate protocol accuracy, ethical adequacy, or researcher learning. On the reviewer side, Fukataki et al. (2025) tested GPT-4 and GPT-4o on Japanese-language protocols and consent forms, finding that GPT-4o achieved 80% accuracy in extracting research objectives and background and 100% accuracy for research design and participant-related risks and benefits, with customized prompts producing more consistent outputs than standard prompts across repeated trials. The authors also noted the absence of gold-standard data and independent raters for system evaluation.

Together, these studies establish the feasibility of LLM-assisted protocol drafting and review-oriented information extraction but leave unresolved how a deployed institutional system should demonstrate acceptable behavior before either investigators or reviewers rely on it. Additionally, in protocol preparation, a polished submission supported by LLMs may conceal whether the investigator understands why a safeguard is necessary. The design problem is therefore determining which work an LLM-assisted system should complete, which reasoning it should make visible, and which decisions must remain with the investigator or institutional authority.

Retrieval-augmented generation (RAG) conditions model responses on an external document corpus rather than relying only on a parametric trained knowledge set (Lewis et al., 2020). RAG can also improve data provenance by connecting generated claims to the governing source material through citations. However, a system may retrieve tangential or incomplete material or generate claims that exceed the evidence provided. As such, comprehensive RAG evaluation might distinguish among retrieval relevance, answer accuracy, faithfulness to context, and source attribution (Yu et al., 2024). Citation presence is also insufficient as a source may exist without warranting the claim attached to it (Wallat et al., 2025). These distinctions also have a replication dimension in that a system intended to scaffold professional reasoning should not expose learners to materially different responses across similar queries. Before usability, pedagogy, or learning can be meaningfully studied, the system's evidence must therefore be characterized for the qualities of reproducibility, provenance, and boundedness.

2.3. Purpose and research questions

IRB Helper integrates citation-backed policy question answering, protocol-writing assistance, and template-constrained document generation. The broader application is intended to support researcher compliance literacy while preserving investigator responsibility and institutional authority. The present study evaluates only the policy-question-answering pipeline. The protocol builder and document-

generation functions are described as part of the system but remain outside the scope of the present study.

The evaluation adopts an assurance-case framework in which a bounded claim about a system is connected to explicit evidence within a specified operating context (Rushby, 2015). The claim examined here is that an Artificial Intelligence (AI) system intended for compliance-literacy support should be evaluated for reproducible retrieval, bounded response behavior, traceable documentary warrants, and an assurance instrument capable of distinguishing system failures from evaluator failures. Whether IRB Helper improves learning, protocol quality, ethical judgment, or participant protection is consequential to the evaluation of this claim.

The article makes four contributions. First, it frames IRB preparation as compliance literacy enacted through a situated institutional genre. Second, it presents the design commitments and architecture of an institutional RAG system intended to support that learning. Third, it reports a 25-execution assurance record spanning system development with two prospectively specified revision cycles. Fourth, it introduces replication reliability and warrant stability as assurance properties for researcher-built educational AI and examines how targeted revisions can improve one pipeline property while revealing or causing failures elsewhere.

> **RQ1.** Across development and three replication configurations, how did benchmark performance, citation-identifier flags, response classifications, retrieval reliability, and evidentiary-path stability vary under repeated execution?
>
> **RQ2.** How did two prospectively specified revision cycles affect their targeted assurance properties, and what residual failure mechanisms remained after verification?

A targeted interpretive review of warrant variation in the final configuration supplements the quantitative analyses by examining whether changing cited authorities represent defensible alternatives, complementary evidence, or substantively problematic inconsistency. Evaluation culminates as an assurance claim for ethical future research with pedagogical foci and human-subjects involvement.

3. System design

The architecture of IRB Helper operationalizes the bounded assurance claim introduced in Section 2. Rather than treating the appearance of relevant retrieval as sufficient evidence of trustworthiness, the system links characteristic failure modes of general-purpose LLM assistance to five design commitments to be designed-for and evaluated (Figure 1).

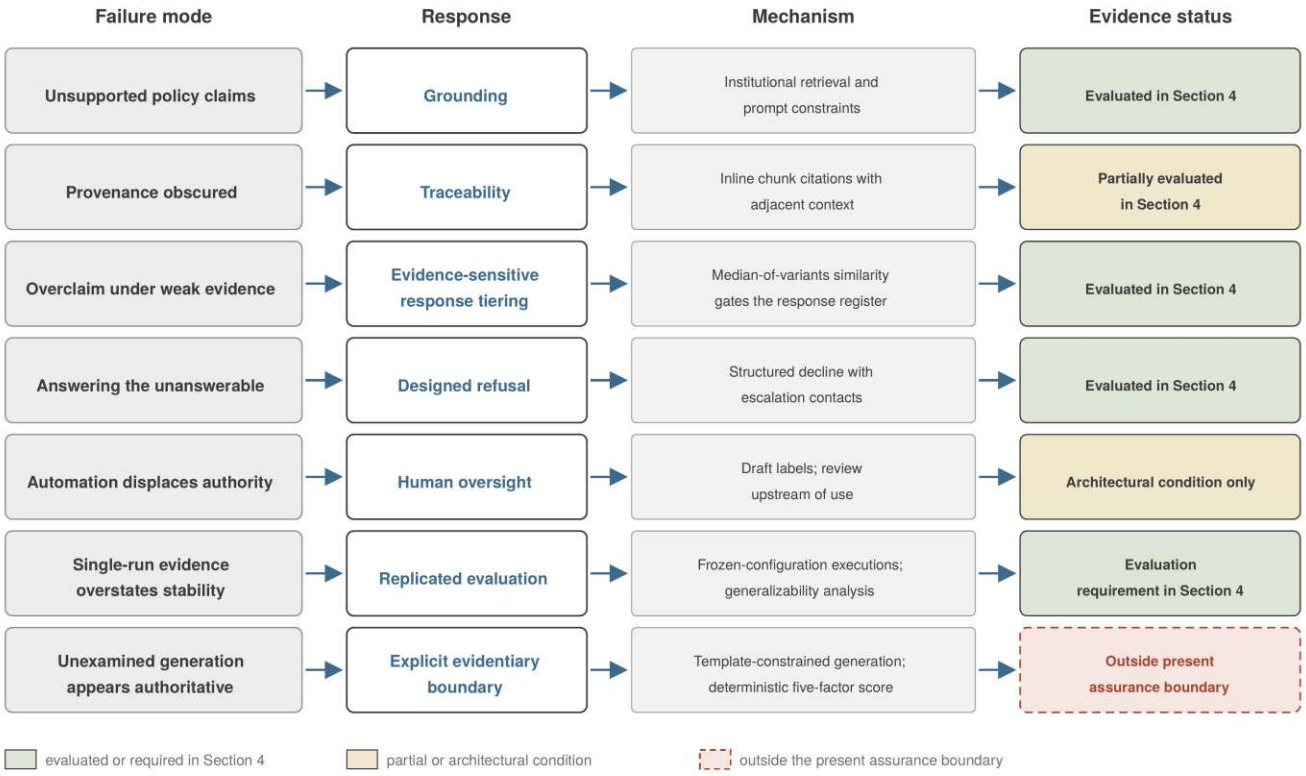

Figure 1. *Failure modes of general-purpose LLM assistance mapped to IRB Helper's design responses, mechanisms, and evidence status.* Rows one through five represent design commitments; row six represents the replication requirement introduced by the assurance study; row seven marks application functions outside the present assurance boundary.

Grounding. Policy-specific claims are intended to derive from retrieved institutional or federal text rather than parametric model knowledge. When the corpus does not provide sufficient evidence, the system should then qualify, redirect, or decline instead of presenting general model knowledge as institutional policy.

Traceability. Responses include inline citations linked to specific source passages so that investigators and IRB analysts can inspect the evidence supporting a claim.

Evidence-sensitive response tiering. Retrieval evidence governs the tone and scope of the response. In the final evaluated configuration, the response tier is assigned from the median cosine similarity across three query-variant retrievals, using thresholds of .50, .35, and .20. Higher tiers authorize direct, citation-backed synthesis; lower tiers require progressively more qualified guidance, explicit source limitations, referral, or structured decline. A one-step conservative downshift is applied when the three variant scores disperse across tiers.

Human oversight. IRB Helper is positioned as preparation support upstream of mandatory human review, not as a substitute for investigator responsibility or IRB authority. Interface language identifies

generated content as draft or preparatory material, and consequential claims remain subject to verification against official sources and applicable institutional review.

Designed refusal. When a request cannot be supported by the available corpus or exceeds the system's legitimate scope, the intended behavior is a structured decline accompanied by an appropriate escalation path. Designed refusal is intended to model an epistemic behavior relevant to compliance literacy where insufficient document evidence escalates to human support.

3.1. Architecture overview

IRB Helper is a full-stack web application. The client is a React and TypeScript single-page application delivered as a static bundle. Server-side logic runs as a suite of Deno serverless edge functions covering policy chat, document embedding, protocol-section assistance, document generation, quality-score computation, an automated stress-test procedure, and URL-based policy ingestion. Persistent state is held in a managed PostgreSQL 15 database with the pgvector extension for approximate vector similarity search, row-level security enforced on all user-scoped tables, and role-based access control implemented through SECURITY DEFINER functions.

LLM access is abstracted through a unified gateway supporting multiple providers. The production generation model for policy chat is google/gemini-2.5-flash-lite, with temperature constrained to a low range for compliance-critical generation. Embeddings are produced by OpenAI's text-embedding-3-small, truncated to 768 dimensions via the API's dimensions parameter and stored in vector(768) columns under an IVFFlat cosine index. The same embedding model is used at ingestion and at query execution, including hypothetical-document variants (Section 3.3). Production answers are synthesized by a deliberately small, low-cost model for generalizability.

3.2. Corpus and ingestion pipeline

The first iteration, v0.9's knowledge base comprised 78 ingested documents and 1,097 chunks of text, approximately 1.66 million characters and 415,000 tokens by standard heuristic. The corpus spans the institution's HRP-series policies, procedures, templates, and guidance. An audit identified duplicate ingestion of the HRP-103 Investigator Manual, the corpus's largest document at 101 chunks. The first revision cycle removed the duplicate while preserving the earlier document record, leaving 77 unique documents and 996 chunks from v1.0 onward. No other document, chunk boundary, or ingestion setting was changed during that correction..

Ingested documents are parsed to plain text with structural metadata preserved, split at semantic boundaries into chunks of approximately 500 to 1,500 tokens, embedded, and stored with source

metadata. Metadata extraction at ingestion captures contact information including emails, phone numbers, and URLs enabling escalation behaviors to emerge from established protocols.

Parsing bounds what the system can retrieve. Complex PDF and DOCX structures may lose tables, headings, or other signposting during plain-text conversion. Consequentially, material that is not ingested cleanly cannot later be retrieved or cited faithfully.

3.3. Query processing, routing, and retrieval

A user query first passes through a pre-retrieval routing layer (Figure 2). Urgent and identified out-of-scope requests may be directed to structured-decline or contact responses. Section 4 shows that this routing layer also captured eight corpus-covered suite questions, producing a consequential false-boundary condition that remained unresolved in the final evaluated configuration.

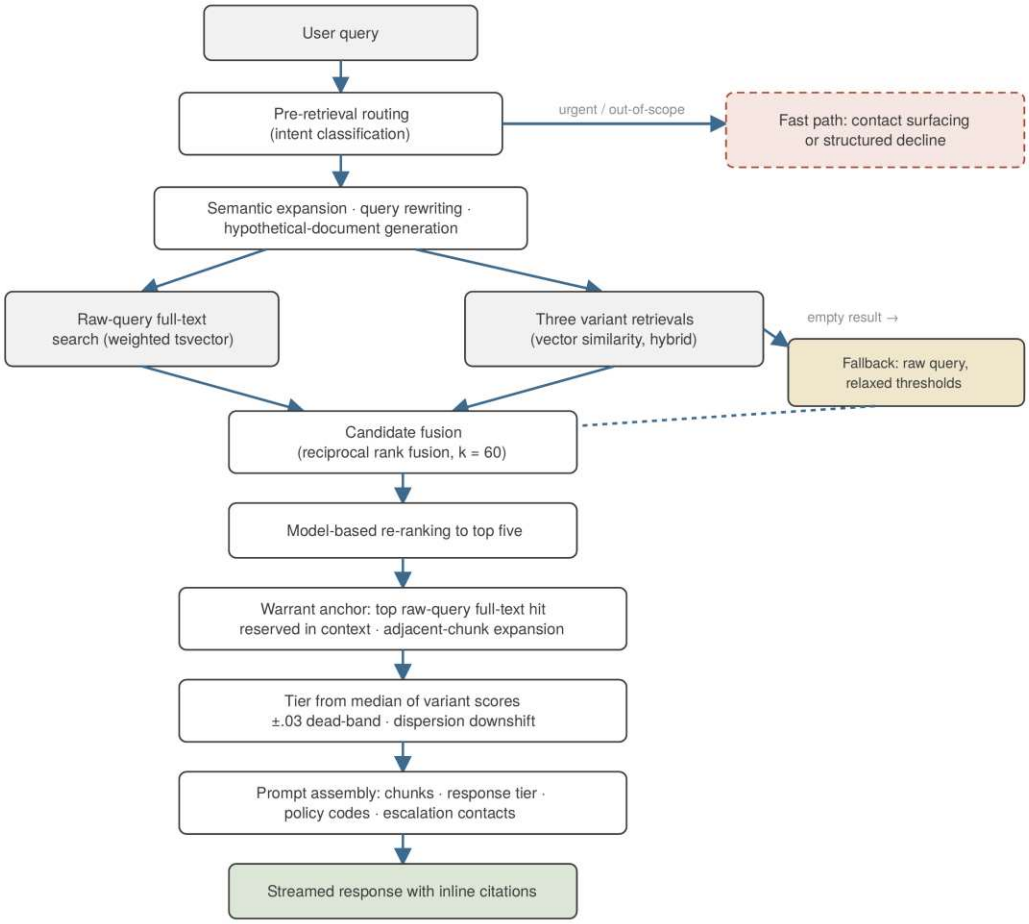

Figure 2. *End-to-end query processing in IRB Helper, final evaluated configuration (v1.1).*

Queries proceeding to retrieval undergo semantic expansion, query rewriting, and hypothetical-document generation (HyDE; Gao et al., 2023). Three query variants are retrieved independently. Full-text search operates on the raw user query, while expansions are confined to semantic retrieval. Hybrid

candidates combine vector similarity and weighted full-text search using reciprocal rank fusion ($k = 60$; Cormack et al., 2009), with a rescue search available for sparse-result, low contextual similarity cases. When a first retrieval produces no chunks, a fallback retries the raw query under more relaxed thresholds.

A model-based reranking pass then reduces the candidate pool to the five most relevant chunks. A deterministic warrant anchor reserves a synthesis-context position for the top raw-query full-text result when available, and adjacent-chunk expansion supplies surrounding material to reduce the risk that policy qualifications are detached from the retrieved passage. The response tier is assigned from the median of the three variant scores and may be conservatively downgraded when those scores disperse. Retrieved chunks, the selected response tier, recognized policy codes, and extracted contacts are assembled into the synthesis prompt, and the answer is returned with inline citations.

3.4. Protocol builder and document generator

The protocol builder function implements the institutional HRP-591 template as a guided workflow, moving from study title and personnel through structured context elicitation, optional upload of existing study documents, and LLM-assisted drafting section by section. Sections are linked in a dependency graph so that revisions to upstream sections such as methodology propagate suggested revisions to dependents such as risks and consent procedures. Each drafted section is scored on a 0–100 scale by a deterministic heuristic (Figure 3) summing five factors: presence of substantive content (20 points), meeting a section-typed minimum length (20 points), keyword coverage against terms extracted from the institutional guidance for that section (up to 30 points), absence of placeholder patterns such as bracketed [TBD] markers (15 points), and presence of paragraph or list structure (15 points). Scores of 70 or higher are labeled good, 40–69 needs work, and below 40 incomplete.

An investigator who receives a low score can explore which factor contributed, and the scoring can be inspected by an IRB analyst without invoking model behavior. The heuristic is interface feedback intended to prompt reflection and revision. A conversational refinement loop translates natural-language revision requests, such as making a section appropriate for adolescent participants, into structured edits that preserve compliance-relevant content. Other functions provide indicators of progress and confirmation of section completion.

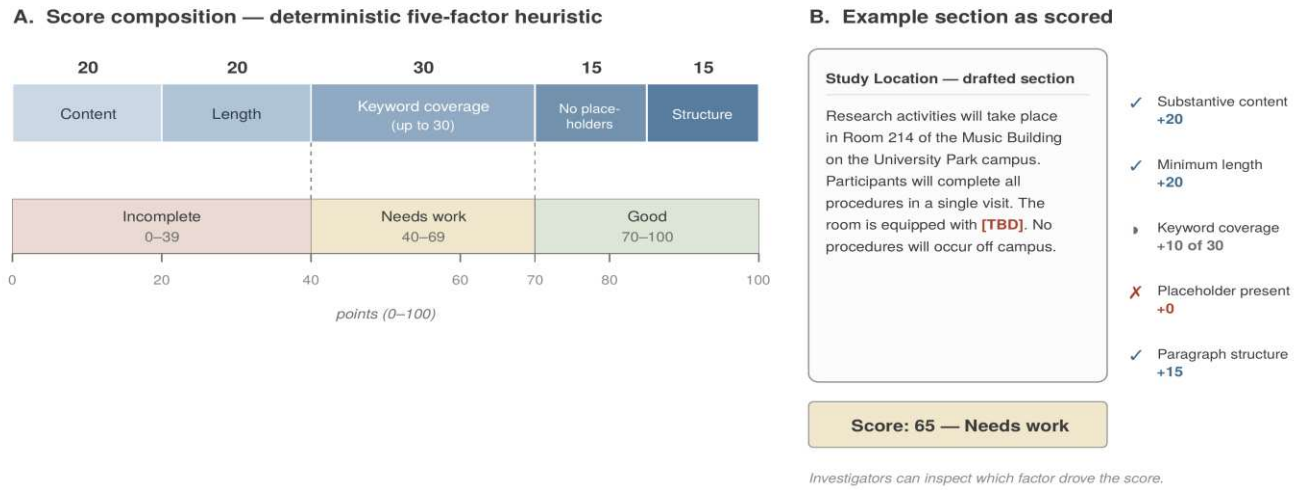

Figure 3. *Five-factor heuristic scale for the protocol builder.*

The document generator produces study-specific consent forms, parental permission and assent materials, and recruitment text. Generation is mapped to institution templates and the user's structured study context is interpolated into template language. Users are reminded that nothing the system produces becomes a participant-facing artifact until it has been reviewed and has received any applicable institutional determination or approval. Document generation also remains outside the present assurance boundary and should not be interpreted as evaluated participant-facing performance. As previously stated, the protocol builder and document generator are not assessed in the assurance study but included here to establish the application's intended pedagogical and compliance literacy contexts.

3.5. Implementation and deployment status

User data including protocols, generated documents, and chat history are isolated per authenticated user through row-level security policies enforced at the database layer, and, by design, no cross-user data enters the model context. Model requests are transmitted to providers through the gateway on a per-request basis in that the application persists no protocol content with providers. Provider-side retention is governed by vendor terms. The application is not publicly available; as of June 2026, it operates as a restricted internal deployment, and recorded use reflects internal and development activity prior to human-subjects involvement.

4. Technical assurance study

The technical assurance study evaluated whether IRB Helper's policy-question-answering pipeline and the evidence used to characterize it behaved reproducibly across repeated execution. The study comprised an initial development window, a five-execution v0.9 baseline evaluation block, two

prospectively specified revision cycles, and five-execution verification evaluation blocks for v1.0 and v1.1. The same 60-item suite and deterministic scoring instrument were retained throughout to preserve longitudinal continuity. Because the suite informed development and revision, the results represent regression assurance and mechanism testing and are not generalizable.

The evaluation is deliberately bounded as it asks whether a researcher-built AI system intended for compliance-literacy support is sufficiently reproducible, traceable, bounded, and diagnosable to justify subsequent evaluation with learners. Data analyzed included developer-authored queries, system outputs, and institutional policy documents. No human participants were involved.

4.1. Evaluation design

The evaluation used the application's automated stress-test procedure, comprising 60 queries spanning basic retrieval (n = 8), semantic understanding (n = 8), complex scenarios (n = 7), cross-document synthesis (n = 6), edge cases (n = 6), and five policy-specific domains of five questions each, covering HIPAA, conflict of interest, continuing review, multi-site research, and data-safety monitoring. Ten items addressed multi-site research and data-safety monitoring, topics for which controlling source documents were deliberately absent from the evaluation corpus. These items assessed whether the system appropriately refused queries. The remaining 50 items are described below as corpus-covered items. Adversarial probes included empty input, nonsensical text, and a request for the deliberately nonexistent document HRP-999. All queries passed through the same production endpoint used by the application, exercising the deployed query-processing, retrieval, re-ranking, generation, citation, and refusal pipeline described in Section 3. Responses received scores from 0 to 5 under a deterministic rubric described in the supplementary materials. The rubric rewarded expected-policy alignment, synonym-expanded keyword coverage, retrieval evidence, response form, and appropriate escalation while penalizing failed retrieval and citations to unrecognized HRP-pattern identifiers. A secondary allowlist accommodated authentic institutional documents referenced within corpus materials but not separately ingested.

The application's operational criterion required a pass score of 4 for basic-retrieval items and 3 for other categories. Structured declines on designed-gap and nonexistent-document items received a base score of 4 when the response identified its evidentiary limitation and supplied an appropriate escalation route, with half-point citation and contact bonuses able to raise the rounded total to 5. A uniform criterion required a score of 4 for every item and provided a more conservative comparison across categories. The scores are regression-benchmark indicators. Put otherwise, the rubric does not independently establish

substantive correctness. Section 4.6 provides a bounded interpretive review of final-configuration warrant variation.

The complete suite was executed eight times from January 12 through January 21, 2026. During this development period, semantic expansion, retrieval logic, prompt constraints, identifier allowlists, and related system components were iteratively refined. A maintenance execution followed on April 2. Five executions on July 6 and 7 (runs 10–14), conducted with no changes to the prompt, corpus, model designation, retrieval configuration, scoring rules, or application code between runs, characterized the post-tuning build, designated v0.9.

Two prospectively specified revision cycles followed, each specified in advance with fixes tied to observed failure mechanisms. The first cycle (v0.9 → v1.0) comprised five fixes motivated by the baseline replication findings: 1) an empty-retrieval fallback retrying the raw, unrewritten query; 2) rewrite ensembling assigning the response tier from the median of three variant retrievals, with a conservative downshift when variants disperse widely and the selected variant persisted as query-side provenance; 3) a gibberish gate capping the tier at minimal when a query contains no in-vocabulary content tokens, with a short adjudicated set of colloquial-to-regulatory expansion pairs; 4) programmatic regeneration of the referenced-document allowlist; and 5) removal of the duplicate HRP-103 ingestion, leaving 77 unique documents. The updated v1.0 build was stress-tested five times consecutively (runs 15–19).

The second cycle (v1.0 → v1.1) started with four fixes including 1) full-text search restricted to the raw query, with expansions confined to the semantic side; 2) a deterministic ±.03 dead-band resolving boundary-adjacent tier medians conservatively; 3) a deterministic warrant anchor guaranteeing the top full-text hit for the raw query a place in the synthesis context, with a chunk-id tie-break in re-ranking; and 4) a terminology-discipline clause in the synthesis prompt whose specified policy-listing line does not appear in any subsequent output and is treated as not deployed. A single diagnostic execution followed the first fix (run 20, reported but excluded from block statistics), and the completed v1.1 build was executed five times consecutively (runs 21–25).

Suite-level stability was summarized using operational and uniform pass rates, mean benchmark scores, and mean maximum retrieval scores. The response tier is assigned from the median cosine similarity across three query-variant retrievals, gated at 0.50, 0.35, and 0.20, with a one-step conservative downgrade when variant scores disperse widely. The logged best-similarity value is the maximum similarity among the sources ultimately provided to synthesis and is related to, but distinct from, the

gated quantity. Item-level consistency within each five-execution block was therefore assessed on three levels where instrumentation permits, including classification of the logged best-similarity proxy at the production gates, classification of the gated median itself, and the served tier recorded by v1.0 on. Agreement is reported as the proportion of corpus-covered items classified identically across executions and as mean pairwise agreement. Stability of the system's documentation path was assessed among items using pairwise Jaccard coefficients, or the intersection of two executed sets divided by their similarity, averaged across execution pairs—for retrieved document sets, cited policy-code sets, and response-token sets (citation). The term warrant stability describes overlap among the policy codes cited across repeated executions. Warrant stability indicates whether the same documentary authorities recur, not whether those authorities adequately support the generated claims, which the targeted review in Section 4.6 examines.

Questions-by-occasions generalizability analyses were conducted on the 50 corpus-covered items within each block, with variance decomposed into question, occasion, and question-by-occasion-plus-residual components estimated from the expected mean squares of the crossed two-way design; single-occasion coefficients were estimated for the maximum retrieval score and the benchmark score, and a decision study projected reliability under averaging. Automated citation-identifier flags across all 25 executions were adjudicated retrospectively against corpus text under a single criterion. An identifier absent from both the detector allowlists and corpus text was classified a probable model fabrication; an identifier present in corpus text but absent from the detector allowlist at the time of the run was classified an evaluator error; HRP-999 was treated separately as a correct denial.

4.2. Development and v0.9 baseline replication

Adjudication also exposed a coverage cost of the tuning window that no single execution had revealed. Eight corpus-covered items including the adverse-event reporting question, four of the five continuing-review items, the exempt-versus-expedited distinction, the Limited Data Set question, and the add-a-site question retrieved at least weak evidence in runs 1 through 3, began returning empty in run 4, and have retrieved nothing in any execution since run 8. The same constraints that suppressed fabrication converted tangential retrieval into no retrieval for 16% of the corpus-covered suite. These items continue to pass the operational criterion (M = 4.00 in the final block, 100% contact provision) because the system answers them with hedged, source-labeled general guidance and escalation contacts, the designed boundary behavior. Design phases of the present study are described in Table 1.

Table 1

Design phases of the technical assurance study

Phase	Runs	Configuration	Purpose
Development	1–8	Changing	Regression development
Maintenance check	9	Post-tuning	Reliability pilot
Baseline replication	10–14	v0.9	Reliability baseline
First verification	15–19	v1.0	Revision-cycle verification
Diagnostic	20	v1.1a	Revision-cycle support
Second verification	21–25	v1.1	Final verification

Note. Per-execution results for all 25 runs, including corpus-adjudicated fabrication and allowlist-error counts, appear in supplementary Table S1.

Across the five v0.9 executions, uniform-criterion pass rates ranged from 90.0% to 95.0%; the operational-criterion rate was 98.3% in four executions and 100% in one. Mean benchmark scores ranged from 4.35 to 4.43 and mean maximum retrieval scores from .514 to .532. Aggregate stability, however, concealed item-level variation. Classification of the logged retrieval proxy at the production gates was high (identical in all five executions for 80.0% of corpus-covered items; mean pairwise agreement .908); the consequential failures were the empty retrievals described above. Benchmark scores were identical across all five executions for only 48.0% of items, and six questions crossed the uniform pass/fail threshold. Retrieved document sets had a mean pairwise Jaccard coefficient of .658 and were identical across all five executions for 32.0% of items; cited policy-code sets, .449 and 36.0%; response tokens, .473, with the least stable decile sharing .211.

The maximum retrieval score showed high single-occasion reliability, G(1) = .89 (question variance .094, question-by-occasion-plus-residual .011), whereas the benchmark score showed moderate single-occasion reliability, G(1) = .61 (.262, .167), with no detectable occasion main effect. The exemplar query, “How do I report a problem in my study?” spanned the entire response range including a returned empty retrieval once within this block. The expected-policy component (i.e., whether any policy was expected and then retrieved) was consistent for 98% of items and contact provision was consistent throughout, while the query keyword-match count was identical across executions for only 50% of items.

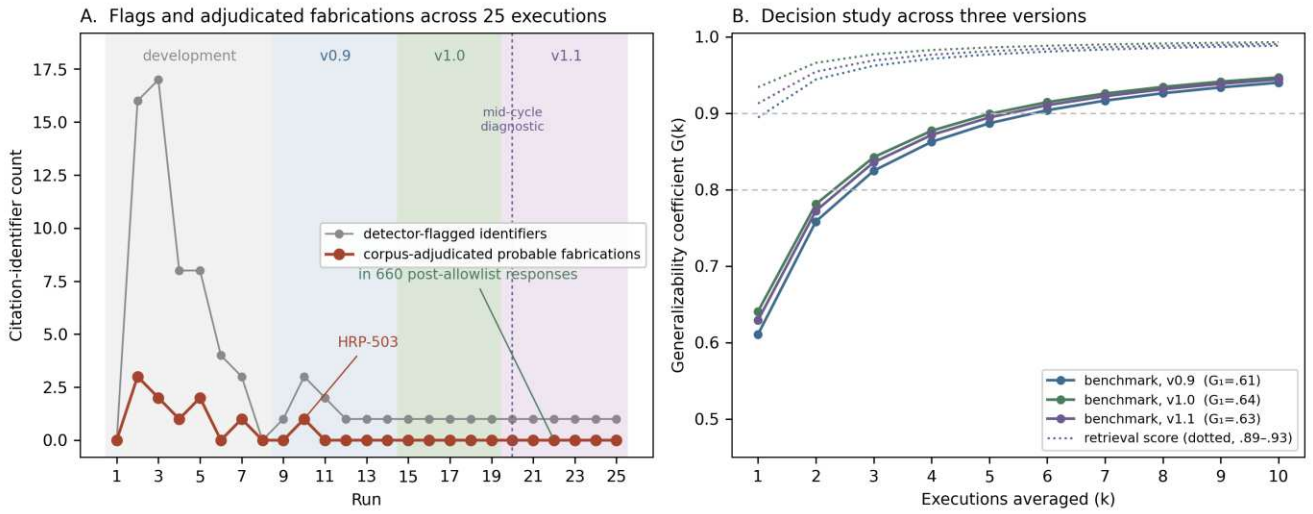

Figure 4. *Development, adjudication, and reliability across three versions.* (A) Detector flags and corpus-adjudicated fabrications across all 25 executions, with the mid-cycle diagnostic run marked; no probable fabrication or detector false positive was observed across the 660 responses following allowlist regeneration. (B) Decision study for the three replication blocks: benchmark-score reliability (G_1 = .61, .64, .63) and retrieval-score reliability (.89–.93) as a function of executions averaged.

4.3. First evidence-driven revision cycle: v1.0

The five v1.0 fixes carried six prospectively specified predictions (Table 2; supplementary Figure S1). One was met in full; the value of prospective specification is that the other five are reported exactly as found. P6 was met without qualification: across the 300 frozen-configuration responses, adjudication found zero probable fabrications and zero detector false positives, the regenerated allowlist having eliminated the evaluator-error class, and the single fabrication observed in the baseline block (HRP-503, run 10) did not recur. P1 was not met, and the analysis reclassified the failure it targeted in that intermittent empty retrieval persisted at a low rate, and the fallback could not touch the deterministic eight-item class, for which a relaxed-threshold retry finds nothing to relax into. P2 — registered against tier gates subsequently shown to be measurement artifacts (Section 4.4) — was not met as registered; the corrected classification analysis appears below and in Table 3. P3 fell narrowly short (.669 against .70) and P4 improved but missed by a distance (.491 against .75). P5 improved directionally but missed (G(1) = .64 from .61, residual variance down 19%; retrieval reliability .89 to .93), and averaging the five executions yields benchmark reliability of .90 (Figure 4B).

Unregistered outcomes moved in the intervention's favor. The gibberish gate performed perfectly, logging .000 in every execution for a probe that had previously drawn scores up to .74. The uniform floor rose to 93.3% and the band narrowed to 3.4 points; the operational criterion reached 60 of 60 in three executions. Identical benchmark scores rose from 48% to 64% of items. Deduplication cut the

Investigator Manual's share of retrieval slots from 35.9% to 23.0%, and the latency cost of ensembling (1.57×) remained within its prospectively specified bound.

Table 2

Prospectively specified predictions and outcomes across both revision cycles

Cycle	Prediction	Target	Baseline (5 runs)	Observed (5 runs)	Met
1	P1. Empty retrievals, corpus-covered items	0 intermittent events	3	2	No
1	P2. Tier agreement, first three executions [a]	≥ .85	.66	.46	No
1	P3. Retrieved document-set Jaccard	≥ .70	.658	.669	No
1	P4. Cited-warrant Jaccard	≥ .75	.449	.491	No
1	P5. Benchmark-score G(1)	≥ .70	.611	.641	No
1	P6. Fabrications and detector false positives	0 and 0	1 fab., 2 FP	0 and 0	Yes
2	P1′. Deterministic empty retrievals (8-item class)	0 or named ingestion list	40	40	No
2	P2′. Served-tier pairwise agreement	≥ .85	.706 [a]	.738	No
2	P3′. Cited-warrant Jaccard	≥ .65	.491	.602	No
2	P4′. Keyword-component consistency	≥ 70%	48%	58%	No
2	P5′. Benchmark-score G(1)	≥ .75	.641	.629	No
2	P6′. Fabrications and false positives	0 and 0	0 and 0	0 and 0	Yes

Note. Targets were committed in the change specifications before any code was modified; both specifications are supplementary material, and shortfalls are reported as found. Baselines are five-execution block values from the unified export. [a] Specified and computed at tier gates subsequently shown to be measurement artifacts; the corrected classification analysis appears in Sections 4.4–4.5 and Table 3, where the gated median's pairwise agreement (.990) and the served tier's (.738) bracket the dispersion downshift's contribution. Cycle 1 P1 counts exclude engineered probes and the eight-item deterministic class, which pre-dates the revisions; averaging the five v1.0 executions yields benchmark G(5) = .90. The two persistently zero-hit topics on the raw-query full-text check — adding a research site, and enrollment during pending continuing review — are the named ingestion candidates; the remaining six items retrieve full-text hits and are blocked by routing, not coverage.

4.4. Diagnosis and second revision cycle: v1.1

In the second cycle code inspection settled the tier architecture as described in Section 4.1 and revealed that tier thresholds previously inferred from the logged score distribution were binning artifacts, produced by hardcoded mass at 1.0 from direct-lookup queries and by repeated-query clustering rather than by any threshold operating on that quantity. The correction was applied retrospectively: at the production gates, proxy-tier classification had been stable in every block (all-runs-identical .80, .88, and .88 across v0.9, v1.0, and v1.1), and an earlier boundary-compression account of tier instability dissolves with the artifact. The evaluation instrument, in other words, required the same auditing as the system, and the assurance process is what made the analyst’s error catchable in evidence. Second, event logs confirmed that v1.0’s empty-retrieval fallback had fired on the one exemplar failure it was built for and recovered nothing: the retry re-entered the same pipeline whose downstream stages caused the failure, localizing the fault below the rewrite layer. Third, stage-level candidate counters localized that fault: full-text search received the raw query joined with the entire expansion-term bag under conjunctive matching, so multi-clause procedural queries had to match every expansion term to match

anything — the synonym map built to bridge register gaps had become the mechanism starving retrieval. Fourth, the scorer's decline override was confirmed from code as a composition effect rather than a rubric conflict (Section 4.1), closing a documented discrepancy.

Five executions of v1.1 again met one of six prospectively specified predictions in full (Table 2; supplementary Figure S2), and again the shortfalls carry the findings. P6′ was met and extended the record: zero probable fabrications and zero detector false positives across the 360 diagnostic and verification responses, bringing the post-allowlist total to 660 consecutive responses without either failure; latency (6.8 s against a 7.3 s bound) was within limits. P1′ was not met, and its failure completed the fault localization the fallback had begun: the eight-item class returned zero sources in every execution because those questions never reach retrieval at all — a pre-retrieval route diverts them to the graceful-decline path, upstream of the full-text repair, the fallback, and every retrieval-side remedy. Two cycles of correctly implemented fixes have now each addressed a genuine mechanism one layer below the actual fault, which is itself a finding: in a stochastic pipeline, fault localization is iterative, and each verification pays for the next diagnosis.

The full-text repair was nonetheless real where retrieval runs. The exemplar query, which had returned empty retrieval intermittently across twelve prior executions, returned weak but nonzero evidence in every v1.1 run, passed the operational criterion in all five, and was rescued to moderate or high tiers by the variant ensemble even when its primary variant found almost nothing; cross-document synthesis, previously the weakest and least consistent category, rose to $M = 4.10$ and passed the operational criterion six of six in every execution (Table S2). P2′ resolved into the cycle's most instructive result. The quantity the code actually gates — the median variant score — was nearly deterministic under replication (pairwise tier agreement .990; identical across all five executions for 97.4% of items). The tier users are served agreed at only .738 pairwise. The gap is the dispersion downgrade: the first cycle's own conservatism mechanism, firing on 36 of 250 live responses and on different items in different runs, is now the dominant source of served-tier variance — a stabilizer introduced to absorb rewrite stochasticity that instead relocated it into the classification itself. The dead-band, by contrast, fired three times in 250 responses: correctly implemented for a boundary-compression problem the corrected gate analysis had dissolved. P3′ narrowly missed (.602 against .65) while representing the largest single-cycle stability gain in the program (from .491), with the retrieved-document Jaccard crossing the first cycle's unmet target (.701) and the warrant anchor performing exactly to specification — sixteen items anchored, the identical chunk in all five executions for every one. P4′ improved but missed (keyword-component consistency 58% against 70%), with attribution limited by the terminology clause's

unverified deployment. P5′ was flat (G(1) = .63; retrieval .91), the uniform band widened slightly (90.0–96.7%), the operational criterion held 59–60 of 60, and contact provision on the two items expecting it slipped once in ten opportunities.

4.5. Assurance evidence across configurations

Table 3 assembles the program's stability metrics across the three replication blocks. Stability accumulated where the fixes had mechanisms to reach: document, warrant, and token overlap rose cycle over cycle, and identical benchmark scores rose from 48% to 66% of items. It stayed flat where the residual variance lives downstream of retrieval, with benchmark reliability near .63 against retrieval reliability near .91 and the keyword component still the localized source. Stable retrieval strength did not imply stable provenance, cited authority, or generated language; multiple policy documents may legitimately support one answer, and whether the remaining warrant turnover reflects documentary plurality or instability in claim-level support is examined in Section 4.6.

Table 3

Stability of assurance evidence across the three replication blocks (corpus-covered items)

Metric	v0.9	v1.0	v1.1
Uniform pass-rate band (%)	90.0–95.0	93.3–96.7	90.0–96.7
Proxy-tier classification, identical across all 5	.80	.88	.88
Gated-median tier, identical across all 5	—	—	.97
Served-tier pairwise agreement	—	—	.74
Retrieved-document Jaccard	.658	.669	.701
Cited-warrant Jaccard	.449	.491	.602
Response-token Jaccard (p10)	.473 (.211)	.480 (.177)	.510 (.253)
Identical benchmark scores, all 5	.48	.64	.66
Benchmark G(1) / retrieval G(1)	.61 / .89	.64 / .93	.63 / .91
Probable fabrications / detector false positives	1 / 2	0 / 0	0 / 0
Mean latency (s)	3.9	6.1	6.8

Note. Tier instrumentation (gated median and served tier) was added in v1.1; the served-tier deficit against the gated median is attributable to the stochastic dispersion downshift (Section 4.4). Fabrication and false-positive counts are corpus-adjudicated.

Across the 660 responses produced after allowlist regeneration in runs 15–25, including the mid-cycle diagnostic execution, no probable fabrication or detector false positive was observed. Across the broader 960-response post-tuning record in runs 10–25, one probable fabrication and two evaluator allowlist errors were observed, all during the v0.9 block. These demonstrate that a clean individual execution cannot establish elimination of a low-frequency stochastic failure and that the assurance instrument requires the same versioned maintenance as the corpus, prompts, routing rules, and model configuration it evaluates.

4.6. Targeted warrant review

To interpret the remaining cited-warrant turnover, the six corpus-covered questions that crossed the uniform pass/fail threshold in the v1.1 block were reviewed across all five executions, yielding 30 response instances. For each question, cited policy authorities and the claims attached to them were compared across executions. Warrant changes were classified as: a) defensible alternatives, where different sources independently supported the same claim; b) complementary, where sources supplied distinct but compatible parts of the answer; c) materially incomplete, where an authority supported the general direction but omitted a consequential condition or limitation; d) conflicting, where executions implied incompatible policy relationships; or e) unsupported, where the cited authority did not warrant the attached claim. This review interprets warrant stability but does not constitute independent expert validation of the system.

[Results paragraph to be added once the 30-instance review is complete.]

4.7. Summary

Across 25 executions, three frozen replication blocks, and two prospectively specified revision cycles, the study produced both accumulation and recursion. Document-set, cited-warrant, and lexical-response stability improved across configurations, while retrieval-score reliability remained consistently high and benchmark-score reliability remained moderate. The cycles also revealed that a fix may work as implemented while targeting the wrong pipeline layer, and that a conservative mechanism can relocate variability into the user-visible response. Eight corpus-covered questions remained routed to a graceful-decline pathway before retrieval, demonstrating that near-ceiling operational pass rates can conceal loss of grounded coverage. Finally, retrospective adjudication corrected both detector errors and an analytic threshold artifact. Assurance for a stochastic educational AI system is therefore a continuing, versioned discipline of repeated execution, prospective prediction, failure localization, and evaluator correction.

5. Discussion

5.1. What replicated assurance revealed

Aggregate pass rates remained near ceiling while the underlying assurance properties moved differently. Retrieval scores were highly reliable in a single execution, yet benchmark scores, retrieved documents, cited warrants, and generated wording were less stable. Targeted revisions improved document and warrant overlap, but benchmark reliability remained largely unchanged because substantial variation occurred downstream of retrieval in generated wording and its interaction with the scoring rubric.

The study also exposed a coverage failure concealed by the operational criterion. Eight corpus-covered questions were routed to structured decline before retrieval, yet continued to pass because the rubric rewarded transparent limitation, general guidance, and escalation. Designed boundary behavior had therefore expanded into a false boundary. This result demonstrates why corpus coverage, refusal quality, and pass rates must be evaluated separately: a system can decline responsibly while declining a question it should have been able to answer.

Finally, the assurance instrument itself became an object of evaluation. Detector allowlists misclassified authentic identifiers, and tier thresholds initially inferred from the logged score distribution proved to be analytic artifacts. Repeated frozen execution made these errors visible and correctable. Responsible evaluation of stochastic educational AI must therefore version scoring rules, allowlists, derived variables, and interpretive assumptions alongside models and corpora.

5.2. Assurance as a prerequisite for compliance-literacy support

The educational claim of this article is deliberately prospective. IRB Helper is intended to support compliance literacy by making policy warrants, evidentiary limits, and escalation pathways visible, but the present study does not show that users learn from those features. It establishes a prior condition: learners should not be asked to rely on an AI-supported environment whose guidance, authority, and boundaries have not been characterized under replication. This sequencing is also regulatory: prospective review cannot be repaired after the fact (Tsan, 2020), so a system should be characterized well enough to identify foreseeable risks before any participant is asked to rely on it.

The findings identify three properties relevant to educational AI design (Table 4). First, provenance must be inspectable, because a stable retrieval score can conceal changing sources and warrants. Second, uncertainty mechanisms must themselves be evaluated, because the dispersion downshift introduced to regulate confidence became a source of user-visible tier variability. Third, system refusal must distinguish a genuine knowledge boundary from a routing or coverage failure. These properties do not guarantee learning, but their absence can corrupt the informational environment in which learning is supposed to occur.

Table 4

Assurance claims, present evidence, and the boundary between technical characterization and educational effectiveness

Assurance claim	Mechanism	Present evidence	Residual limitation
Evaluation evidence is reproducible	Frozen repeated executions; G-study	Section 4	Three local configurations
Retrieval behavior is characterizable	Versioned corpus, repeated retrieval	Section 4	Developer-authored suite

Assurance claim	Mechanism	Present evidence	Residual limitation
Evidentiary paths can be compared	Document and warrant overlap	Section 4	Overlap does not establish truth
Unsupported questions are bounded	Structured decline and escalation	Section 4	Eight corpus-covered items falsely routed
Citation identifiers are governed	Corpus adjudication and allowlist hashing	Section 4	Low-frequency failures cannot be ruled out
Human authority is preserved	Draft labels and escalation	Architectural evidence	No user-behavior evidence
Protocol-writing support benefits learners	Protocol builder and feedback	Not evaluated	Requires human-participant study
Protocol-quality heuristic is valid	Deterministic five-factor score	Not evaluated	Requires analyst validation

5.3. Designed refusal and false evidentiary boundaries

Designed refusal remains an important educational hypothesis, but the present findings complicate it. A system that visibly reaches the limit of its evidence may model the epistemic behavior novice researchers need: recognizing insufficient support and consulting an appropriate human authority. The eight-item routed-empty class shows the inverse risk. When the system mistakes an corpus-covered question for an evidentiary boundary, a polished decline can teach the learner that institutional guidance does not exist or that the issue is inherently indeterminate. Refusal quality must therefore be judged jointly with coverage and routing accuracy rather than treated as success whenever the language is appropriately cautious.

Refusal also operates as institutional communication. The clarity, consistency, and perceived relevance of IRB communication shape researchers' trust, their compliance, and the way faculty socialize students into research practice (Kramer et al., 2009), and an interface that mediates institutional guidance conveys an institutional posture whether or not it intends one: whether uncertainty is acceptable, whether consulting a person is normal, and whether compliance is ethical reasoning or bureaucratic navigation. The inverse risk deserves equal weight. A fluent, uniformly authoritative system could make local policy appear more determinate, more comprehensive, or more unanimous than it is, and traceability and evidence-sensitive response tiering therefore also regulate the voice in which the institution appears to speak. Table 5 extends this reading across the design commitments, pairing each feature with the learner activity it is intended to occasion and the offloading risk it cannot by itself prevent; the same feature can scaffold learning or enable displacement depending on how the user engages with it, and distinguishing those outcomes is precisely the work of the planned human-participant study.

Table 5

Design features, intended learner activity, and offloading risk

System feature	Intended learner activity	Offloading risk
Source-linked answers	Inspect and compare policy evidence	Citations ignored; provenance becomes decoration

System feature	Intended learner activity	Offloading risk
Evidence-sensitive tiering	Judge the strength of available evidence	Tier labels read as authority
Designed refusal	Recognize limits and seek human consultation	Referral becomes task abandonment
Section dependencies	Trace relationships across study decisions	Connections made invisibly on the user's behalf
Document generation	Adapt study information for prospective participants	Fluent wording accepted without comprehension

Note. The same feature can scaffold learning or enable offloading depending on how the user engages with it; distinguishing these outcomes is an aim of the planned human-participant study (Section 6).

5.4. Equity and researcher-built educational AI

Because RAG systems ingest corpora without retraining, the architecture lowers one technical barrier to local adaptation: institutional policies can be incorporated, and policy revisions absorbed, without retraining the underlying model. It does not eliminate the institutional labor of corpus curation and cleaning, template and workflow mapping, vocabulary alignment, escalation-contact configuration, response-tier configuration, benchmark construction, validation, and ongoing governance. The researchers who most need such scaffolding, doctoral students, faculty in unfunded fields, and investigators at institutions without dedicated regulatory staff, are precisely those least served by the current market of static templates and generic chatbots. We do not claim the present system resolves that asymmetry. We claim only that the architecture lowers one of the barriers to resolving it; the remaining barriers are institutional rather than architectural.

5.5. Institutional governance and sustainability

What would an office of research protections need in order to sanction a system like this one? Four operational answers follow from the architecture. First, versioned corpus governance. Policies revise, and a grounded system is only as current as its last ingestion, so corpus synchronization must be an owned institutional process with visible version stamps rather than a developer habit. Second, analyst-facing review workflows. Traceability is designed for verification, and the natural verifiers are the analysts who already read protocols. Giving them the same citations users see converts oversight from sampling to inspection. Third, audit logging. Query and feedback records make the system's behavior reviewable after the fact, the same property boards require of investigators. Fourth, sustainability. Recurring inference costs, template rigidity set against generative flexibility, and prompts that can silently break when a template updates are maintenance realities, and institutions should fund such systems as services rather than accept them as gifts. None of these is exotic. All of them are the ordinary governance any institution already applies to systems of record.

6. Limitations and future work

Eight limitations bound the claims. First, no human participants evaluated IRB Helper; the study establishes technical and informational assurance properties, not usability, learning, protocol quality, cognitive offloading, or participant protection. Second, the stress suite was developer-authored and informed development and revision, so the findings demonstrate regression assurance rather than held-out generalization. An analyst-authored question set should be evaluated separately before broader use. Third, the corpus represents one institution, and portability is architectural rather than demonstrated. Fourth, the automated benchmark combines retrieval, expected terminology, response form, and escalation behavior; it is not an independent measure of substantive correctness and remains sensitive to generated wording. Fifth, eight corpus-covered items remained diverted before retrieval in the final configuration, meaning the system should not be exposed to learners without revising and re-evaluating that routing rule. Sixth, some instrumentation was introduced only in later versions, limiting direct comparison of served response tiers across all blocks; one proposed synthesis-prompt change also could not be verified as deployed. Seventh, the targeted warrant review is author-conducted and interpretive rather than independent expert validation. Eighth, the assurance case is informational rather than comprehensive: application security, vendor retention, prompt injection, access control, malicious corpus alteration, and integration with the institutional IRB platform were not evaluated.

The next research phase should occur only after the routed-empty failure is corrected and the configuration is frozen again. A human-participant study can then examine whether graduate students use the protocol writer effectively, produce stronger protocol language under blinded analyst review, understand and articulate the warrants for their decisions, recognize when escalation is appropriate, and transfer that reasoning to novel cases without the system. The protocol-quality heuristic should be validated against analyst judgment before it is treated as more than revision-oriented interface feedback.

7. Conclusion

IRB preparation is a consequential form of professional learning, and generative AI creates both an opportunity to make institutional guidance more accessible and a risk of replacing difficult reasoning with fluent but unsupported text. The central question is what evidence should be required before such guidance becomes part of an educational environment.

Across 25 executions and two prospectively specified revision cycles, IRB Helper's retrieval evidence proved substantially more reliable than its composite benchmark performance. Document and warrant stability improved across versions, but high pass rates concealed a deterministic loss of grounded coverage, safety mechanisms introduced new user-visible variability, and the evaluation instrument

produced errors requiring retrospective correction. These findings show that grounding, confidence mechanisms, refusal, and benchmark scores cannot be treated as self-validating system features.

The study contributes replication reliability and warrant stability as assurance properties for researcher-built educational AI. More broadly, it positions assurance as an iterative discipline in which the system, its revisions, and the instrument used to evaluate them remain open to correction. A grounded system owes learners a reproducible and defensible account of why each answer should be trusted.

Funding Declaration

There was no funding for this work.

Clinical Trial Number

Not applicable

Ethics, Consent to Participate, and Consent to Publish

The experiments involved artificial systems only and collected no human-participant or personal data; human-subjects review was therefore not applicable.

References

Abbott, L., & Grady, C. (2011). A systematic review of the empirical literature evaluating IRBs: What we know and what we still need to learn. *Journal of Empirical Research on Human Research Ethics, 6*(1), 3–19. <https://doi.org/10.1525/jer.2011.6.1.3>

Anderson, E. E., Hurley, E. A., Serpico, K., Johnson, A., Rowe, J., Singleton, M., Bierer, B. E., Cholka, B., Chaudhari, S., & Fernandez Lynch, H. (2023). Engaging key stakeholders to overcome barriers to studying the quality of research ethics oversight. *Research Ethics, 19*(1), 62–77. <https://doi.org/10.1177/17470161221138028>

Anderson, E. E., Johnson, A., & Fernandez Lynch, H. (2024). Inclusive, engaged, and accountable institutional review boards. *Accountability in Research, 31*(8), 1287–1295. <https://doi.org/10.1080/08989621.2023.2220884>

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Lawrence Erlbaum Associates.

Collins, A., & Kapur, M. (2014). Cognitive apprenticeship. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 109–127). Cambridge University Press. <https://doi.org/10.1017/CBO9781139519526.008>

Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 758–759). Association for Computing Machinery. <https://doi.org/10.1145/1571941.1572114>

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis, 16*(1), 64–93. <https://doi.org/10.1093/jla/laae003>

Desai, P., Nasa, P., Soo, J., Jia, C., Berbaum, M. L., Fischer, J. H., & Johnson, T. P. (2017). Effects of regulatory support services on institutional review board turnaround times. *Journal of Empirical Research on Human Research Ethics, 12*(3), 131–139. <https://doi.org/10.1177/1556264617704294>

Fost, N., & Levine, R. J. (2007). The dysregulation of human subjects research. *JAMA, 298*(18), 2196–2198. https://doi.org/10.1001/jama.298.18.2196

Fukataki, Y., Hayashi, W., Nishimoto, N., & Ito, Y. M. (2025). Developing artificial intelligence tools for institutional review board pre-review: A pilot study on ChatGPT's accuracy and reproducibility. *PLOS Digital Health, 4*(6), Article e0000695. <https://doi.org/10.1371/journal.pdig.0000695>

Gao, L., Ma, X., Lin, J., & Callan, J. (2023). Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1762–1777). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.99

Garrison, N. A., & Cho, M. K. (2013). Awareness and acceptable practices: IRB and researcher reflections on the Havasupai lawsuit. *AJOB Primary Research, 4*(4), 55–63. <https://doi.org/10.1080/21507716.2013.770104>

Godwin, R. C., Bryant, A. S., Wagener, B. M., Ness, T. J., DeBerry, J. J., Horn, L. L., Graves, S. H., Archer, A. C., & Melvin, R. L. (2024). IRB-draft-generator: A generative AI tool to streamline the creation of institutional review board applications. *SoftwareX, 25*, Article 101601. <https://doi.org/10.1016/j.softx.2023.101601>

Grady, C. (2015). Institutional review boards: Purpose and challenges. *Chest, 148*(5), 1148–1155. https://doi.org/10.1378/chest.15-0706

Grohmann, S. (2025). *EthicAlly: A prototype for AI-powered research ethics support for the social sciences and humanities* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2508.00856>

Hadden, K. B., Prince, L. Y., Moore, T. D., James, L. P., Holland, J. R., & Trudeau, C. R. (2017). Improving readability of informed consents for research at an academic medical institution. *Journal of Clinical and Translational Science, 1*(6), 361–365. <https://doi.org/10.1017/cts.2017.312>

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. <https://doi.org/10.1145/3571730>

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>

Klitzman, R. (2015). *The ethics police? The struggle to make human research safe*. Oxford University Press.

Kramer, M. W., Miller, V. D., & Commuri, S. (2009). Faculty and institutional review board communication. *Communication Education, 58*(4), 497–515. <https://doi.org/10.1080/03634520903006208>

Lea, M. R., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education, 23*(2), 157–172. <https://doi.org/10.1080/03075079812331380364>

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

Marcos Miguel, N., & Noy, S. (2023). Human subjects research guidelines for undergraduate researchers: An analysis of institutional review board (IRB) websites at top liberal arts colleges in the United States. *Journal of Empirical Research on Human Research Ethics, 18*(4), 263–277. <https://doi.org/10.1177/15562646231190818>

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech, 70*(2), 151–167. <https://doi.org/10.1080/00335638409383686>

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. U.S. Department of Health, Education, and Welfare.

Ogloff, J. R. P., & Otto, R. K. (1991). Are research participants truly informed? Readability of informed consent forms used in research. *Ethics & Behavior, 1*(4), 239–252. <https://doi.org/10.1207/s15327019eb0104_2>

Onakomaiya, D., Pan, J., Roberts, T., Tan, H., Nadkarni, S., Godina, M., Park, J., Fraser, M., Kwon, S. C., Schoenthaler, A., & Islam, N. (2023). Challenges and recommendations to improve institutional review boards' review of community-engaged research proposals: A scoping review. *Journal of Clinical and Translational Science, 7*, Article e93. <https://doi.org/10.1017/cts.2023.516>

Orimadegun, A. E. (2020). Protocol and researcher's relationship with institutional review board. *African Journal of Biomedical Research, 23*(Suppl. 2), 15–20.

Paasche-Orlow, M. K., Taylor, H. A., & Brancati, F. L. (2003). Readability standards for informed-consent forms as compared with actual readability. *New England Journal of Medicine, 348*(8), 721–726. https://doi.org/10.1056/NEJMsa021212

Parreco, L., Rooney, L., Hampp, S., Brown, A., & Minasian, L. (2019). A rare opportunity: Examining the experience of a new institutional review board. *Journal of Empirical Research on Human Research Ethics, 14*(3), 274–285. <https://doi.org/10.1177/1556264619841815>

Protection of Human Subjects, 45 C.F.R. § 46 (2018).

Rosenfeld, S. J. (2020). Institutional review board assessment—Balancing efficiency and quality. *Ochsner Journal, 20*(1), 50–55. <https://doi.org/10.31486/toj.19.0075>

Rushby, J. (2015). *The interpretation and evaluation of assurance cases* (Technical Report SRI-CSL-15-01). SRI International. [https://www.csl.sri.com/users/rushby/papers/sri-csl-15-1-assurance-cases.pdf](https://www.csl.sri.com/users/rushby/papers/sri-csl-15-1-assurance-cases.pdf?utm_source=chatgpt.com)

Schneider, C. E. (2015). *The censor's hand: The misregulation of human-subject research*. MIT Press. https://doi.org/10.7551/mitpress/9780262028912.001.0001

Silberman, G., & Kahn, K. L. (2011). Burdens on research imposed by institutional review boards: The state of the evidence and its implications for regulatory reform. *The Milbank Quarterly, 89*(4), 599–627. https://doi.org/10.1111/j.1468-0009.2011.00644.x

Tsan, M.-F. (2020). Comment on the proposed institutional review board retrospective review of research. *Journal of Empirical Research on Human Research Ethics, 15*(3), 229–231. <https://doi.org/10.1177/1556264620908153>

U.S. Government Accountability Office. (2023). *Institutional review boards: Actions needed to improve federal oversight and examine effectiveness* (GAO-23-104721). [https://www.gao.gov/products/gao-23-104721](https://www.gao.gov/products/gao-23-104721?utm_source=chatgpt.com)

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature, 614*(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7

Wallat, J., Heuss, M., de Rijke, M., & Anand, A. (2025). Correctness is not faithfulness in retrieval-augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval* (pp. 22–32). Association for Computing Machinery. <https://doi.org/10.1145/3731120.3744592>

Wingate, U. (2018). Academic literacy across the curriculum: Towards a collaborative instructional approach. *Language Teaching, 51*(3), 349–364. <https://doi.org/10.1017/S0261444816000264>

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. <https://doi.org/10.1111/j.1469-7610.1976.tb00381.x>

Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024). *Evaluation of retrieval-augmented generation: A survey* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2405.07437>