Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL36386

A Related Line of Research RegTutor: Self-correcting Agentic Retrieval-Augmented Generation for Medical-Device Compliance Training

Hanyi Jin

Faculty of Data Science, City University of Macau, Macau, China
jackingooll286@gmail.com

Abstract. Medical-device compliance training requires feedback grounded in authoritative regulations. However, existing retrieval-augmented generation approaches often lack explicit evidence assessment and post-generation validation, weakening citation traceability in high-stakes regulatory tutoring. To address this limitation, we propose RegTutor, a self-correcting agentic retrieval-augmented generation framework for medical-device compliance training. Specifically, RegTutor retrieves regulatory evidence from 21 CFR Parts 803, 807, and 820, as well as FDA guidance, evaluates evidence sufficiency, rewrites inadequate queries, generates tutoring feedback, and validates citation format and claim-evidence alignment before finalizing its response. Extensive results showed that RegTutor achieved 100.0% answer correctness, citation accuracy, and helpfulness on a predefined 30-item controlled evaluation. In a balanced 300-item operational rerun, all RegTutor runs completed successfully, and citation accuracy reached 99.0%. On a separately frozen 60-item manually reviewed blind evaluation, RegTutor achieved 100.0% answer correctness, citation accuracy, and helpfulness among 59 completed runs. A secondary replication with GLM-4.7 further demonstrated 100.0% answer correctness and 98.3% citation accuracy. These findings demonstrate the value of self-correction and citation validation for traceable regulatory tutoring

Keywords: Medical-device compliance, Large language models, Agentic RAG, Citation validation, Interactive compliance training

1. Introduction

Medical-device compliance training requires learners to interpret FDA requirements for reporting, registration and listing, and quality management under 21 CFR Parts 803, 807, and 820 [1]. Regulatory-affairs specialists, quality personnel, new employees, and relevant medical or biomedical trainees therefore need systematic practice in applying authoritative sources to realistic situations. Such training requires not only access to regulations, but also explanations of why an answer is correct or incorrect.

Large language models (LLMs) can make regulatory learning more accessible, but direct use in compliance education is risky. They may generate unsupported claims, fabricate citations, or attach a

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GL36386

genuine citation to an unsupported statement [2]. Although retrieval-augmented generation can provide external evidence to a language model [3], retrieval alone does not ensure that generated claims are supported by the retrieved passages. Moreover, most RAG applications are designed for question answering rather than instructional tutoring; they rarely generate exercises, assess learner responses, identify misconceptions, or provide corrective explanations.

This study proposes RegTutor, a self-correcting agentic RAG framework for interactive medical-device compliance training. RegTutor retrieves regulatory evidence, assesses its sufficiency, rewrites queries when necessary, generates tutoring feedback, and validates citation format and claim-evidence alignment before finalization. Failed checks trigger targeted revision and revalidation, ensuring that each final regulatory claim remains traceable to retrieved evidence.

This study makes three contributions: (1) it presents an interactive compliance-training system that supports regulatory question generation, answer grading, and evidence-grounded explanatory tutoring rather than regulatory question answering alone; (2) it introduces a self-correcting RAG workflow that combines iterative retrieval with claim-level citation validation; and (3) it evaluates Naive RAG, Standard RAG, and RegTutor using controlled evaluation, full-dataset operational testing, held-out blind validation, and secondary cross-model replication.

2. Related work

2.1. Compliance QA and retrieval-augmented generation

RAG connects language-model outputs with retrievable external knowledge, improving the inspectability and updateability of responses to knowledge-intensive tasks [3]. In regulated domains, this capability is especially valuable because users must be able to inspect the legal or technical basis for a response. However, a retrieval result does not by itself guarantee that the generated answer accurately reflects the retrieved source, and an inline citation may still be incomplete, misplaced, or unsupported.

Existing compliance-oriented RAG systems are primarily designed for information access and question answering. In a typical interaction, a user asks a question and the system returns an answer along with retrieved context. However, retrieval alone does not provide a pedagogical cycle of prompting, assessing, and explaining learner responses. The reverse tutoring scenario—where the system poses regulatory exercises, evaluates learner responses, diagnoses misconceptions, and delivers evidence-grounded corrective feedback—is what RegTutor is designed to target.

2.2. Agentic RAG and citation attribution

Agentic RAG enables a system to inspect intermediate evidence and take corrective actions rather than relying on a single retrieve-then-generate pass. ReAct interleaves reasoning with information-seeking actions [4], SELF-RAG combines retrieval, generation, and self-reflection [5]; and CRAG evaluates retrieved documents and triggers corrective retrieval when evidence quality is inadequate [6]. These approaches make RAG better able to detect insufficient or irrelevant evidence and to improve the generation process iteratively.

A related line of research concerns the attribution and evaluation of generated claims. A related line of work evaluates whether RAG responses remain faithful to retrieved evidence [7]. However, existing work does not combine iterative evidence correction and claim-level citation validation for regulatory compliance training, where every corrective statement to a learner must be grounded in an authorized provision. This integrated tutoring traceability is the focus of RegTutor.

DOI: 10.54254/2755-2721/2026.GL36386

3. Method

3.1. Task formulation

RegTutor supports three instructional modes: question generation, answer grading, and explanatory tutoring. In question-generation mode, the system creates a regulatory exercise grounded in retrieved provisions and FDA guidance documents. In grading mode, it compares a learner response with a reference answer and identifies whether the required regulatory conclusion is preserved. In explanatory tutoring mode, which is the primary focus of this study, the system produces corrective feedback that explains the relevant requirement, identifies the learner's misconception, and attaches inline citations to supporting evidence.

Formally, given a regulatory question q, a reference answer a, and a learner response s, the system produces feedback f together with grounded citations $C = \{(c_i, r_i)\}_{i=1}^{m}$, where c_i is a regulatory claim in f and r_i is the specific retrieved provision or FDA guidance that supports it. The objective is that every material claim in the final feedback can be verified against its cited evidence. Although the three modes share the same retrieval and validation mechanism, our experiments primarily evaluate explanatory tutoring because it directly assesses whether the corrective feedback provided to learners is both useful and traceable.

$$f, C = \text{RegTutor}(q, a, s), \quad C = \{(c_i, r_i)\}_{i=1}^{m} \tag{1}$$

3.2. System overview

RegTutor is implemented as a single-agent state machine with six functional stages and two conditional correction loops, as illustrated in Figure 1. Given a regulatory learning task, the system first retrieves relevant passages from the regulatory corpus. It then assesses whether the retrieved evidence is sufficient to support the intended feedback. If the evidence is insufficient, the system rewrites the query and repeats the retrieval process. When sufficient evidence is available, RegTutor generates instructional feedback with inline citations.

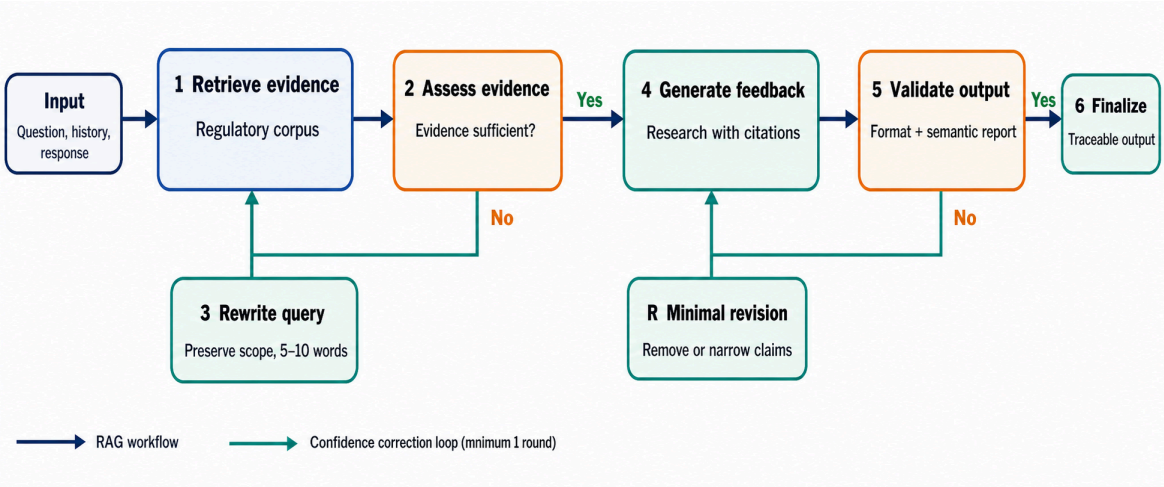

Figure 1. RegTutor evidence-grounded self-correcting workflow. The system retrieves regulatory evidence, assesses evidence sufficiency, rewrites queries when necessary, generates tutoring feedback, validates claim-citation alignment, and performs bounded minimal revision before finalization

DOI: 10.54254/2755-2721/2026.GL36386

The generated feedback then undergoes validation. The system checks citation format, verifies that cited sources belong to the retrieved evidence set, and assesses whether each cited passage supports its associated claim. If validation fails, RegTutor performs a minimal revision and repeats the validation check. Both the retrieval loop and the validation loop arelimited to three rounds to prevent excessive iterations. By recording the retrieved evidence, validation results, and revision rounds, the workflow provides an auditable process for producing regulatory tutoring feedback.

3.3. Iterative retrieval

Regulatory questions may contain ambiguous terminology, incomplete wording, or an incorrect source scope. After initial retrieval, RegTutor assesses whether the retrieved passages support both the required regulatory conclusion and the explanatory reasoning. If evidence is insufficient, rewrite_query reformulates the query while preserving explicit source constraints, such as a CFR part or FDA guidance document, and repeats retrieval for up to three rounds. If sufficient evidence is still unavailable, the system restricts feedback to the available evidence rather than producing an unsupported conclusion.

3.4. Self-correcting validation

Before finalization, RegTutor decomposes generated feedback into material claims and validates each claim against three conditions: citation validity, source eligibility, and semantic support. The system checks whether a citation follows the required format, refers to an eligible retrieved source within the required regulatory scope, and substantively supports its associated claim. If a claim fails validation, revise_feedback makes a minimal correction by correcting the citation, removing the unsupported statement, or narrowing an overly broad claim; it cannot introduce new unsupported requirements or unverified sources. The revised feedback is revalidated for up to three rounds, making citations verifiable evidence links rather than decorative references.

4. Experiments

4.1. Experimental setup

Datasets. We constructed RegTutor-Bench, comprising 300 medical-device compliance questions evenly distributed across 21 CFR Parts 803, 807, and 820. For the balanced v2 split, 60 questions were reserved for development and the remaining 240 formed the test pool. The predefined 30-item controlled evaluation set was drawn exclusively from this test pool, and the held-out blind evaluation used 60 separately authored questions.

Baselines. We compared RegTutor with Naive RAG and Standard RAG using the same regulatory corpus and underlying language model. Naive RAG followed a retrieve-then-generate workflow without citation constraints. Standard RAG prompted the model to provide citations but did not validate claim–citation support. RegTutor additionally performed evidence assessment, query rewriting, and self-correcting citation validation.

Implementation Details. The controlled, full-dataset, and held-out blind evaluations used deepseek-v4-flash through the DeepSeek Chat Completions API. A secondary cross-model replication used GLM-4.7 on a separate frozen 60-item set. Within each comparison, the model, regulatory corpus, output limit, temperature, and item order were held constant; only the retrieval, prompting, and validation workflow differed

DOI: 10.54254/2755-2721/2026.GL36386

4.2. Main results

Table 2 reports the rerun on a predefined 30-item controlled evaluation set drawn only from the balanced v2 test split and addresses RQ1. Naive RAG achieved 100.0% answer correctness but produced no strictly correct citations (0.0%), demonstrating that apparently correct tutoring feedback is not necessarily traceable to regulatory evidence. Standard RAG increased citation accuracy to 96.7% by prompting the model to provide citations.

RegTutor achieved 100.0% Answer Correctness, Citation Accuracy, and Helpfulness, improving citation accuracy over Standard RAG by 3.3 percentage points. Thus, the complete workflow produced fully correct and evidence-supported feedback on this frozen controlled sample.

Compared with Standard RAG, RegTutor improved controlled-set citation accuracy from 96.7% to 100.0%, reflecting the combined effect of evidence assessment and validation.

Table 1. Results on the precommitted controlled evaluation set (n = 30). Best results are shown in bold

Method	Answer Correctness	Citation Accuracy	Helpfulness
Naive RAG	100.0%	0.0%	96.7%
Standard RAG	100.0%	96.7%	100.0%
RegTutor	100.0%	100.0%	100.0%

4.3. Component analysis

The three-system comparison provides evidence for the contribution of citation prompting and self-correcting validation, addressing RQ2. Moving from Naive RAG to Standard RAG increases citation accuracy from 0.0% to 96.7%, showing that explicit citation prompting substantially improves traceability. Moving from Standard RAG to RegTutor further increases citation accuracy from 96.7% to 100.0%. This remaining improvement is attributable to RegTutor's evidence-sufficiency assessment, bounded query rewriting, claim-level validation, and minimal-revision loop.

This comparison does not separately isolate the effect of iterative retrieval and self-correcting validation, since RegTutor introduces both mechanisms jointly. Nevertheless, it demonstrates the practical value of the complete workflow: citation prompting alone improves traceability, whereas evidence assessment and validation are needed to make the resulting citations considerably more reliable for regulatory tutoring.

4.4. Full-dataset, held-out blind, and cross-model evaluation

To assess operational behavior beyond the controlled comparison, we reran all three methods on the complete 300-item balanced v2 benchmark. Table 3A shows that Standard RAG achieved 87.7% citation accuracy, whereas RegTutor achieved 99.0% citation accuracy. All three methods completed all 300 items without API or structured-output failures. Compared with Standard RAG, RegTutor required substantially more computation, using approximately 3.60 times as many tokens and 2.08 times the average latency per completed item.

We further conducted a separately frozen external blind evaluation using 60 newly authored and manually reviewed questions, with 20 questions from each of 21 CFR Parts 803, 807, and 820. The questions had no overlap with the 300-item operational benchmark, development data, pilot

DOI: 10.54254/2755-2721/2026.GL36386

examples, or the controlled evaluation set. As shown in Table 3B, Naive RAG and Standard RAG completed all 60 runs, while RegTutor completed 59 runs and encountered one structured-output error. Among its completed runs, RegTutor achieved 100.0% Answer Correctness, Citation Accuracy, and Helpfulness. Standard RAG achieved 88.3% citation accuracy, whereas Naive RAG achieved 3.3%. These results provide stronger robustness evidence than the previous small-scale evaluation, while also demonstrating that structured-output handling remains an operational limitation.

To assess whether the workflow can operate under a different model family, we conducted a secondary GLM-4.7 cross-model replication on a separately frozen 60-item candidate set balanced across Parts 803, 807, and 820. The set was derived from the same eCFR snapshots and kept separate from the original 300-item benchmark and held-out blind set. Because item-level human review was pending at execution, this is interpreted as a secondary candidate validation rather than an independently human-reviewed blind test. We ran six sequential 10-item checkpoints; all 180 method runs completed without API or structured-output failures. As shown in Table 4, RegTutor achieved 100.0% Answer Correctness, 98.3% Citation Accuracy, and 91.7% Helpfulness. These results demonstrate that the workflow can be executed reliably and retain high evidence-grounded performance under GLM-4.7.

Table 2. Full-dataset operational evaluation and held-out blind validation. Best results are shown in bold

(A) Full-dataset operational evaluation (n = 300)						
Method	Completed	Answer	Citation	Helpful	Latency (s)	Tokens
Naive RAG	300	99.7%	0.0%	99.3%	3.400	377,776
Standard RAG	300	99.3%	87.7%	100.0%	3.355	543,429
RegTutor	300	97.3%	99.0%	98.7%	6.982	1,956,348
(B) External blind evaluation (n = 60; completed runs shown)						
Method	Completed	Answer	Citation	Helpful	Latency (s)	Tokens
Naive RAG	60/60	100.0%	3.3%	100.0%	3.748	60,621
Standard RAG	60/60	100.0%	88.3%	100.0%	3.534	96,000
RegTutor	59/60	100.0%	100.0%	100.0%	6.256	286,860

Table 3. Secondary GLM-4.7 cross-model replication (60-item frozen candidate set)

Method	Completed	Answer	Citation	Helpful	Latency (s)
Naive RAG	60/60	100.0%	0.0%	96.7%	26.266
Standard RAG	60/60	98.3%	98.3%	96.7%	18.646
RegTutor	60/60	100.0%	98.3%	91.7%	97.909

5. Conclusion

This study presented RegTutor, a self-correcting agentic RAG framework for medical-device compliance training. RegTutor combines evidence retrieval, evidence-sufficiency assessment, query

DOI: 10.54254/2755-2721/2026.GL36386

rewriting, instructional feedback generation, and claim-level citation validation to produce traceable corrective feedback. Unlike conventional retrieve-then-generate workflows, it revises insufficient retrieval results and unsupported claims before finalizing learner-facing responses.

Across the 300-item balanced operational evaluation, RegTutor completed all runs and achieved 99.0% citation accuracy, substantially outperforming Standard RAG (87.7%). On the separately frozen 60-item external blind evaluation, RegTutor achieved 100.0% answer correctness, citation accuracy, and helpfulness among 59 completed runs. These results indicate that citation prompting alone is insufficient for reliable regulatory tutoring and that explicit evidence assessment and validation can improve traceability.

The study remains limited by the modest size of the manually reviewed blind set and one structured-output failure. Future work should include larger independently reviewed benchmarks, expert and learner studies, and improved structured-output handling to assess real compliance-training outcomes.

References

[1] U.S. Food and Drug Administration. (n.d.). Title 21—Food and drugs: Parts 803, 807, and 820. *Electronic Code of Federal Regulations*. Retrieved July 31, 2026, from https://www.ecfr.gov/current/title-21

[2] 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

[3] 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.

[4] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=WE_vluYUL-X

[5] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*. https://openreview.net/forum?id=hSyW5go0v8

[6] Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2024). Corrective retrieval augmented generation. *arXiv*. https://doi.org/10.48550/arXiv.2401.15884

[7] Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150–158). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-demo.16