ELECTRA: A Grounded Retrieval-Augmented Generation System for Electric Distribution Standards

Michael Warren[1]*, Reddy Mandati[1], Vladyslav Anderson[1], Bobby Besharati[1], Tim Krall[1]
[1]Exelon Analytics, Philadelphia, PA, USA

***Abstract*—ELECTRA is a retrieval-augmented generation (RAG) system for electric distribution standards knowledge retrieval deployed across four Exelon operating companies (BGE, ComEd, PECO, and PHI). Building on a previously deployed gas standards RAG architecture, ELECTRA introduces Docling-based document processing for AI-powered layout and table extraction, a dual text–image index, and a direct-quoting prompting strategy that enforces exact citations with page-level references. An LLM-as-judge evaluation of 932 expert-verified question–answer pairs compares the baseline and enhanced architectures across groundedness, similarity to ground truth, and coherence. The enhanced system increases groundedness by 36.9% (from 3.458 to 4.734 on a 5-point scale), with 86.9% of responses showing improved source attribution, while incurring a modest 7.3% reduction in coherence scores. In production use, page-level quotations reduce standards verification time from tens of minutes to under a minute per query and provide auditable, safety-critical traceability for electric distribution design and compliance workflows.**

***Index Terms*—**Decision support systems, document handling, electricity distribution, electrical safety, information retrieval, knowledge based systems, machine learning, natural language processing, power distribution reliability, standards

NOMENCLATURE

- ELECTRA – Electric distribution standards retrieval system at Exelon
- RAG – Retrieval-augmented generation
- RAQS – RAG Answer Quality Score
- CQS – Context Quality Score
- BM25 – Okapi BM25 relevance ranking function
- RRF – Reciprocal Rank Fusion
- NESC – National Electrical Safety Code

I. INTRODUCTION

Electric distribution standards documentation forms the backbone of utility operations, encompassing technical specifications, regulatory mandates, and safety protocols that govern the design, construction, operation, and maintenance of power distribution systems [1]. At Exelon Corporation [2], engineers across four operating companies BGE, ComEd, PECO, and PHI routinely navigate thousands of pages of standards to ensure compliance with federal and state regulations, as well as industry best practices. These standards directly govern electric distribution design and construction decisions, where NESC and company-specific requirements define clearances, ratings, and safety margins for field installations.

Traditional methods of accessing this documentation involve manual keyword searches through PDF repositories, a process that is both time-consuming and error-prone. Employee turnover has compounded this challenge, resulting in attrition of experienced personnel and difficulties for newer staff in adequately applying industry standards and quickly referencing safety-critical information. This knowledge accessibility gap creates operational inefficiencies and potential compliance risks, especially as engineers experiment with large language models in reliability-critical workflows [3].

RAG systems offer a solution by combining LLM capabilities with domain-specific document retrieval [4]. Building upon prior work deploying the Gastor system for gas standards at BGE (and now implemented at PECO and PHI) [5], this paper presents ELECTRA, a RAG-based knowledge system for electric distribution standards. ELECTRA extends the foundational RAG architecture with enhancements in document processing, citation accuracy, and response groundedness.

This paper makes four contributions. First, it introduces a Docling-based document processing pipeline that preserves

979-8-3315-8138-1

layout, hierarchy, and table structure for electric distribution standards. Second, it implements a dual text–image retrieval architecture that combines embeddings for narrative text and engineering diagrams to support multimodal queries. Third, it defines and deploys a RAG Answer Quality Score (RAQS) that combines context quality, citation usage, and answer confidence to provide per-response quality feedback. Fourth, it reports a production A/B evaluation over 932 expert-verified electric distribution standards question–answer pairs, showing a 36.9% improvement in groundedness and substantial reductions in engineer verification time when answering standards-related design and compliance questions.

II. BACKGROUND

A. Electric Distribution Standards Challenge

Electric distribution standards encompass the technical specifications that govern how electrical power is safely and reliably delivered from substations to end customers. At Exelon, engineers must navigate standards that incorporate requirements from the National Electrical Safety Code (NESC) [1], and other sources, for the safety of electric supply and communication utility systems along with company-specific design guidelines that reflect regional conditions and operational practices.

The challenge lies in the volume and complexity of this documentation. A typical operating company could maintain over 7,500 pages across roughly 15 or more core standards documents, covering topics from overhead line construction to underground cable installation. When an engineer needs to verify a specific requirement such as the minimum clearance for a transformer cabinet door or the conductor ampacity for a given configuration they must search through multiple PDFs, interpret cross-references, and synthesize information from tables and diagrams. This manual process typically requires 10-30 minutes per question and is susceptible to oversight errors.

B. RAG Architecture for Standards Retrieval

RAG addresses these challenges by augmenting large language models with access to authoritative external knowledge bases [4]. Unlike general-purpose LLMs that rely solely on training data, RAG systems retrieve relevant document chunks at query time, enabling precise, traceable responses grounded in source material.

The core RAG pipeline consists of three stages: (1) document preprocessing and vector embedding; (2) retrieval, where user queries are matched against the vector database using similarity search; and (3) generation, where retrieved context is combined with the query to produce an augmented prompt for the LLM.

The baseline ELECTRA implementation, following the Gastor architecture [5], employs hybrid search combining keyword matching with semantic vector search, followed by re-ranking using BM25 and Reciprocal Rank Fusion (RRF) algorithms [6]. This foundation presented opportunities for enhancement through structured document processing and multimodal embeddings, enabling richer context retrieval and more precise source attribution. These enhancements are presented in Section III.

III. METHODOLOGY

A. Document Processing with Docling

The primary innovation in document preprocessing is the adoption of Docling, an open-source PDF processing toolkit developed by IBM Research [7]. Traditional PDF parsing tools extract raw text but lose structural information: tables become jumbled text streams, hierarchical headers mix with body content, and diagrams are discarded entirely. This structural loss is particularly problematic for electrical standards, which contain specification tables (voltage ratings, ampacity values), wiring diagrams, and cross-references that are critical for accurate retrieval. Docling addresses these limitations through two specialized AI models:

DocLayNet: Provides page layout analysis using an object detection architecture trained on 81,000 manually labeled technical document pages. It identifies text blocks, headers, figures, and reading order across complex multi-column layouts [7].

TableFormer: Recognizes table structure, preserving row/column relationships even for complex tables with merged cells, hierarchical headers, and borderless formats.

B. Dual-Index Multimodal Architecture

Building upon the Docling-processed documents, the system employs a dual-index retrieval architecture:

The generated text index uses Cohere embed-v3-english embeddings (1024 dimensions) [8]. Text chunks maintain metadata including document name, page number, section hierarchy, and adjacent page references.

The generated image index uses Cohere embed-v4 embeddings (1536 dimensions). Each extracted figure is embedded along with its caption and contextual description, enabling queries like "show transformer cabinet configuration" to retrieve relevant visual references.

An intelligent routing node analyzes each query to determine whether image retrieval is necessary, preventing latency overhead for text-only questions while enabling multimodal responses when appropriate. Results from both indices are

979-8-3315-8138-1

merged and re-ranked using scaled score fusion before context generation as demonstrated in Fig. 1.

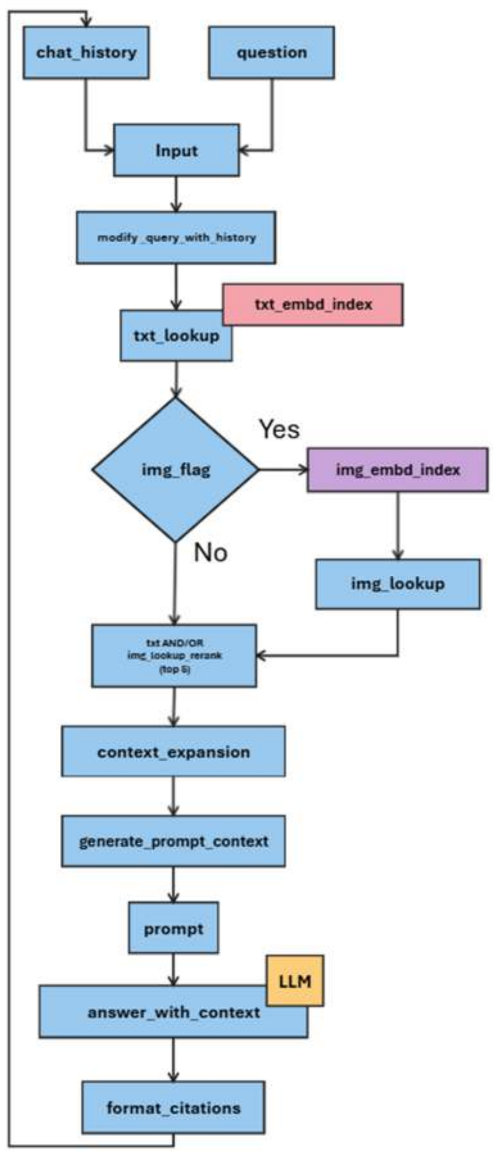

Figure 1 - Dual Index Architecture

C. Direct Quoting System Prompt

One of the most impactful enhancements is the direct quoting system prompt, which fundamentally changes how the LLM generates responses. The baseline system would paraphrase retrieved information, resulting in responses like:

"According to the installation requirements, the current transformer cabinet's hinged entry point must have a sealing hasp with a 7/16" hole."

While this paraphrased response is coherent, it lacks traceability. Engineers cannot quickly verify the claim without searching the source document. Even when engineers were satisfied with answers like this, user feedback frequently requested more explicit source attribution. The enhanced system prompt enforces a strict citation format:

"The [Document Name] states: '[exact quote]' [docN] (Page X)."

This constraint produces responses like:

"The Service and Meter Requirements Book specifies: 'TRANSFORMER CABINET SHALL ALSO HAVE A HINGED DOOR WITH SEALING HASP (7/16" HOLE)' [doc9] (Page 104)."

The direct quote preserves exact specifications (7/16" vs. approximate interpretations) and provides immediate page reference for verification. This approach trades some response fluency for dramatically improved groundedness and verifiability.

D. Context Expansion and RAG Quality Scoring

The enhanced architecture includes context expansion, where the system automatically retrieves adjacent pages (±4 pages from initially retrieved chunks) to capture specifications that span multiple pages or require surrounding context. Additionally, if the initial retrieval contains text results but the query appears to involve physical configuration, the system automatically searches the image index for supplementary diagrams.

To provide real-time quality assessment, we define a RAG Answer Quality Score (RAQS), conceptually inspired by RAGAS [9]:

$$RAQS = 0.5CQS + 0.3CU + 0.2AC \quad (1)$$

The RAG Answer Quality Score (RAQS) is defined in (1) as a weighted combination of context quality (CQS), citation usage (CU), and answer confidence (AC)

Where the Context Quality Score (CQS) combines retrieval score (weighted mean of rerank scores with position decay), coherence score (document concentration and page proximity), and coverage score (diversity of sources with multimodal bonuses). Citation Usage measures the proportion of available sources referenced in the response. Answer Confidence applies inverse hedging penalties for uncertain language phrases like "cannot find" or "no information available." Weights were chosen to prioritize context quality over verbosity and tuned qualitatively based on pilot use.

This quality scoring provides engineers with an immediate indicator of response reliability and guides the generation process toward more grounded outputs.

E. Evaluation Framework

979-8-3315-8138-1

To quantify the impact of these enhancements, we implemented an LLM-as-Judge evaluation framework using GPT-4o [10]. This methodology follows the RAG evaluation paradigm used by Microsoft Azure AI Foundry [11], which defines groundedness as measuring "how consistent the response is with respect to the retrieved context." The evaluation compares the baseline system against the enhanced system across three metrics on a 1-5 scale:

Groundedness: How well claims are supported by cited sources, with emphasis on direct quotes and document references.

GPT Similarity: Semantic similarity between generated responses and expert-verified ground truth answers.

Coherence: Logical flow, clarity, and professional structure of the response.

The ground truth dataset comprises 932 question-answer pairs covering ComEd electric distribution standards, with questions categorized by type (factual, reasoning, comparative) and each answer verified by Engineering Applications domain experts with source page references.

IV. CHALLENGES

A. Document Currency and Version Control

Maintaining current documentation in the knowledge bases requires coordination across four operating companies (BGE, ComEd, PECO, PHI), each with independent standards update cycles. When a standard is revised, whether due to regulatory changes, equipment upgrades, or lessons learned from field operations, the corresponding vector embeddings must be regenerated and the indexes updated promptly.

Despite system disclaimers noting that the technology is evolving, users rely on the retrieved information for safety-critical decisions [12]. The current mitigation involves version tracking metadata, change notification protocols with document owners, and periodic reindexing schedules.

B. User Education and Prompt Formulation

A significant challenge is equipping users with effective prompting skills [13]. Initial usage patterns showed engineers formulating broad, keyword-style queries such as "transformer requirements" rather than specific questions like "What is the minimum clearance requirement for a 500 kVA pad-mounted transformer adjacent to a building wall?"

Broad queries trigger retrieval from multiple documents and sections, often returning information that is technically correct but not from the user's expected source. This challenge is compounded by misconceptions about RAG systems; some users believed the system would "learn" from their usage, conflating RAG with model fine-tuning.

Change management efforts focus on corpus transparency (prominently displaying which documents are indexed), prompting guidance (encouraging specific questions with context), and technical literacy (explaining that the system retrieves from fixed documents rather than generating novel knowledge).

C. Balancing Groundedness and Response Coherence

An inherent tension exists between maximizing groundedness and maintaining natural response coherence. Direct quoting, while dramatically improving traceability, can produce responses that feel verbose or fragmented when multiple specifications must be combined. For instance, answering a question that requires information from three different sections results in three distinct quoted passages rather than a synthesized summary.

User feedback indicates this trade-off is acceptable for compliance-checking tasks where precision matters, but less ideal for exploratory queries where users seek conceptual understanding. Future work may involve query intent classification to adjust response style accordingly direct quoting for factual queries, synthesized responses for conceptual ones.

V. RESULTS

The enhanced ELECTRA system was evaluated against the baseline using the 932-question ground truth dataset. Table I presents comparative performance metrics.

A. Groundedness Improvement

The most significant result is the 36.9% relative improvement in groundedness (3.458 to 4.734 on a 5-point scale), with 86.9% of responses showing improved scores.

This improvement stems directly from the direct quoting system prompt combined with the Docling-based preprocessing pipeline, which preserves document structure and enables more precise context retrieval.

B. Response Example Comparison

To illustrate the qualitative difference, consider this query: "What specific feature is required for the current transformer cabinet's hinged entry point?"

Baseline Response (Groundedness: 3/5): "According to the installation requirements, the current transformer cabinet's hinged entry point must have a sealing hasp with a 7/16" hole."

Enhanced Response (Groundedness: 5/5, RAQS: 0.59): "The Service and Meter Requirements Book specifies: 'TRANSFORMER CABINET SHALL ALSO HAVE A HINGED DOOR WITH SEALING HASP (7/16" HOLE)' [doc9] (Page 104)."

979-8-3315-8138-1

The enhanced response enables verification in seconds an engineer can navigate directly to Page 104 of the specified document rather than searching through the entire standards collection.

C. Coherence Trade-off Analysis

The 7.3% decrease in coherence scores (4.652 to 4.312) represents an expected trade-off. Direct quotations from technical documents are more verbose than paraphrased summaries, and combining multiple quoted specifications produces responses that prioritize accuracy over natural flow.

However, this trade-off is justified by the use case requirements. Notably, 50.8% of responses maintained the same coherence score, and only 39.2% showed degradation, suggesting the impact is moderate and contextual.

D. Practical Implications

The quantitative improvements translate to measurable operational benefits. With baseline responses, engineers spent considerable time verifying information by searching source documents manually. The enhanced system's direct quotes with page references reduce verification time from an estimated 10-30 minutes to under 1 minute, as users can navigate directly to the cited page.

TABLE I
A/B COMPARISON RESULTS

Metric	Baseline	Enhanced (Updated RAG)	% Questions Improved
Groundedness	3.458	4.734	86.9%
GPT Similarity	2.796	3.107	33.2%
Coherence	4.652	4.312	10.1% degraded

VI. CONCLUSION

This paper presented ELECTRA, an enhanced RAG system for electric distribution standards knowledge retrieval. Building upon the foundational Gastor architecture for gas standards [5], ELECTRA introduces several contributions:

Docling-based document processing with AI-powered layout analysis and table structure recognition, enabling extraction of specifications from complex PDF documents while preserving tabular data integrity and structure.

Dual-index multimodal retrieval combining text embeddings with image embeddings to support queries involving both textual specifications and visual diagrams.

Real-time RAQS quality scoring providing engineers with confidence metrics based on context quality, citation usage, and answer certainty.

The quantitative evaluation across 932 ground truth questions demonstrates substantial improvements: groundedness improved by 36.9% (3.458 to 4.734), with 86.9% of responses showing enhanced source attribution.

ACKNOWLEDGMENT

The authors thank the Electric Distribution Standards teams across all operating companies for their valuable feedback during the evaluation of ELECTRA.

REFERENCES

[1] IEEE Standards Association, "National Electrical Safety Code," ANSI C2-2023, 2023.

[2] Exelon Utilities. [Online]. Available: https://www.exeloncorp.com/.

[3] Y. Hu, K. Kanagarajan, F. Hamim, and F. Karray, "The Use and Misuse of Pre-Trained Generative Large Language Models in Reliability Engineering," in Proc. IEEE Annual Reliability and Maintainability Symposium (RAMS), 2024, pp. 1-7.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459-9474.

[5] R. Mandati, M. Warren, B. Besharati, T. Krall, "Utility-Scale Knowledge Retrieval: Deploying Generative Artificial Intelligence Application for Gas Standards Management," in CIGRE US National Committee 2025 Grid of the Future Symposium, 2025.

[6] X. Li, et al, "BMX: Entropy-weighted Similarity and Semantic-enhanced Lexical Search," arXiv, vol. 2408.06643, Aug. 2024.

[7] C. Auer, M. Dolfi, A. Carvalho, C. Ramis, and P. Staar, "Docling Technical Report," arXiv:2408.09869, Aug. 2024.

[8] Cohere Inc., "Embed Models Documentation," 2025. [Online]. Available: https://docs.cohere.com/docs/embed

[9] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," arXiv:2309.15217, Sep. 2023.

[10] OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, Mar. 2023.

[11] Microsoft, "Evaluate with Azure AI Foundry SDK: RAG Evaluators," Microsoft Learn, 2025. [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-foundry/concepts/evaluation-evaluators/rag-evaluators

[12] M. Ni, Y. Zhao, and J. Guo, "ChatGrid: Intelligent Knowledge Q&A for Power Dispatching Control Based on Large Language Models," in Proc. IEEE International Conference on Information Technology, Networking, Electronic and Automation Control (ITNEC), 2025, pp. 146.

[13] J. Joskowicz and D. Slomovitz, "Engineers' Perspectives on the Use of Gen AI Tools in the Workplace," IEEE Engineering Management Review, vol. 52, no. 2, pp. 94-102, 2024.

979-8-3315-8138-1