SALUS

TECHNICAL WHITE PAPER + BENCHMARK REPORT

Rosie: Technical White Paper & Benchmark Report

Rosie (Domain-Specific AI System for Construction Safety)

Authors and reviewers: Alex Jacobs; Dany Ayvazov

July 2026 | Paper v2.0.0 | Benchmark dataset v1.0

1,023	95.01%	+8.0-20.9 pp	p < 0.001
benchmark questions	Rosie accuracy	reported advantage	pairwise comparisons

A domain-specific retrieval and evaluation report for construction-safety AI. All figures and benchmark claims in this edition are drawn from the July 2026 source paper.

Publication record

This file is the versioned PDF companion to the canonical HTML technical resource. It preserves the supplied July 2026 paper while making its publication version, dataset version, and citation explicit.

Publication	Technical white paper and benchmark report
Paper version	2.0.0 (paper-v2.0.0)
Release date	2026-07-18
Benchmark dataset	SALUS-SafetyQA v1.0
Authors	Alex Jacobs; Dany Ayvazov
Reviewed by	Alex Jacobs; Dany Ayvazov
Review date	2026-07-18
Canonical HTML	https://www.salussafety.io/us/resources/technical/rosie-ai-methodology

SUGGESTED CITATION

Jacobs, A., and Ayvazov, D. (2026). *Rosie: Technical White Paper & Benchmark Report* (Version 2.0.0). SALUS Technologies. https://www.salussafety.io/us/resources/technical/rosie-ai-methodology

VERSION SCOPE

Paper v2.0.0 is the July 2026 report edition. The separately published benchmark dataset remains v1.0. A paper-version change does not imply a dataset-version change.

DATA AND LICENSING

The paper does not specify a reuse license for the report itself. The SALUS-SafetyQA dataset and code are published under CC BY 4.0 at https://github.com/Salus-Technologies/SALUS-SafetyQA.

SOURCE OF RECORD

The supplied authoring source is available at https://docs.google.com/document/d/1LLQ6tFyeZiAAAfbnVJReQ5mPxIEBXGHZ5DJ_J10iTPQ/edit.

Contents

The page numbers below refer to this versioned PDF edition.

Abstract

Construction safety documentation is critical but underserved by general-purpose AI systems, which frequently hallucinate plausible but incorrect safety information in high-stakes scenarios. We present Rosie, a production domain-specific AI system for construction safety that achieves 95.01% accuracy through hybrid retrieval combining dense vectors, BM25 keyword search, and learned reranking with safety-optimized prompting to deliver 100% document-grounded responses.

Our contributions include: (1) SALUS-SafetyQA benchmark-the first comprehensive evaluation benchmark for construction safety AI, containing 1,023 expert-validated multiple-choice questions across 11 question types and 10 document source types, (2) comprehensive comparative evaluation demonstrating 8.0-20.9 percentage point improvements of domain-specific AI (Rosie) over frontier LLMs-GPT-5.6 Sol (87.00%), Gemini 3.5 Flash (86.31%), Claude Opus 4.8 (78.49%), and Claude Sonnet 5 (74.10%)-with rigorous statistical analysis (all $p < 0.001$, McNemar's test), and (3) systematic failure mode analysis identifying where and why general-purpose LLMs fail on safety-critical questions, including hallucination of specifications and poor performance on equipment manuals.

Keywords: *retrieval-augmented generation, construction safety, domain adaptation, benchmark evaluation, failure mode analysis*

1. Introduction

1.1 Motivation

Construction safety documentation presents unique challenges for information retrieval systems:

- **Heterogeneous formats:** SDS, equipment manuals, regulations, and policies follow distinct structures requiring specialized parsing
- **High-stakes accuracy:** Incorrect safety information can result in injuries, fatalities, and significant legal liability
- **Temporal sensitivity:** Regulations and standards update frequently, requiring version-aware retrieval
- **Multi-jurisdictional complexity:** State and federal regulations may conflict or complement each other
- **Technical terminology:** Domain-specific language that general-purpose models frequently misinterpret

Existing solutions fail to address these challenges adequately:

- General-purpose LLMs (ChatGPT, Gemini, Claude, etc.) hallucinate plausible but incorrect safety information
- Traditional search returns relevant documents but requires manual extraction of specific answers
- Rule-based systems lack flexibility for natural language queries and require constant maintenance

1.2 Contributions

This paper makes the following technical contributions:

- SALUS-SafetyQA benchmark - 1,023 expert-validated multiple-choice questions across 11 question types and 10 document source types
- Comprehensive comparative evaluation of domain-specific RAG (Rosie) versus frontier LLMs (GPT-5.6, Claude Opus 4.8 / Sonnet 5, Gemini 3.5 Flash) with rigorous statistical analysis demonstrating 8-21 percentage point improvements
- Analysis of domain adaptation gaps - systematic failure modes in general-purpose LLMs including hallucination of specifications and poor performance on equipment manuals

2. Related Work

2.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) systems enhance large language models with external knowledge retrieval [1, 2]. Recent advances include:

- **Dense retrieval:** DPR [3] and ColBERT [4]
- **Hybrid approaches:** combining dense and sparse signals [5, 6]
- **Reranking:** cross-encoders and learned rankers [7]

Rosie builds on these foundations with safety-specific adaptations for construction contexts.

2.2 Domain-Specific AI Systems

- **Medical:** Med-PaLM 2 achieved expert-level medical question answering [8]
- **Legal:** Reasoning-Focused Legal Retrieval Benchmark [9]

2.3 Safety AI Benchmarks

- Accident cause classification using Word2Vec and deep learning [10]
- Incident report analysis with machine learning [11]
- PPE detection in images [12]

2.4 Benchmark Generation

Our benchmark generation methodology adapts the automated MCQA framework introduced by Gokdemir et al. [13]. To our knowledge, SALUS-SafetyQA is the first comprehensive QA benchmark designed specifically for construction safety.

3. System Overview

3.1 Architecture

Rosie uses a layered retrieval and reasoning pipeline:

- **Hybrid search:** dense and sparse vector search with native metadata filtering for robust coverage
- **Hybrid reranking:** multi-signal rerank combining semantic similarity, keyword match, and learned rankers
- **Query expansion/extraction:** model-based decomposition of user queries into canonical search terms
- **Validation layer:** checks retrieved spans for alignment with the query
- **Answer synthesis:** custom safety assistant prompt that enforces grounding, compliance, and conciseness

3.2 Retrieval Performance

- **Search latency:** <2 seconds (p95)
- **Accuracy:** 95.01% on safety questions
- Scalable to millions of indexed safety documents

4. Dataset & Benchmark Design

4.1 Corpus

The Rosie corpus contains safety-critical documentation across the following types:

- **SDS:** Chemical hazards and PPE requirements (297 questions, 29.0%)
- **REGULATION:** State and federal safety regulations (211 questions, 20.6%)
- **STANDARD:** ANSI/OSHA standards (218 questions, 21.3%)
- **MANUAL:** Equipment operation and maintenance (139 questions, 13.6%)
- **POLICY:** Contractor and GC safety policies (10 questions, 1.0%)
- **FORM_CHECKLIST:** Structured inspection requirements (10 questions, 1.0%)
- **TRAINING_MATERIAL:** Instructional content (77 questions, 7.5%)
- **SAFETY_ALERT:** Incident and hazard bulletins (27 questions, 2.6%)
- **REPORT:** Investigation and compliance findings (24 questions, 2.3%)
- **OTHER:** Miscellaneous safety-related documents (10 questions, 1.0%)

4.2 Question Generation

We developed a custom LLM-based generation pipeline based on Gokdemir et al. [13], adapting their approach for safety-specific requirements. Each example includes:

- Natural-language question
- Expected short answer
- Multiple-choice version with four realistic options
- Question type classification (11 categories)
- Source type and jurisdictional metadata

4.3 Dataset Statistics

Total Questions: 1,023

4.3.1 Distribution by Question Type

Question Type	Count	Percentage
Specification	324	31.7%
Compliance	230	22.5%
What Hazards	160	15.6%
How To	95	9.3%
When Required	58	5.7%
Definition	53	5.2%
Emergency	31	3.0%

Question Type	Count	Percentage
What PPE	28	2.7%
Who Responsible	23	2.2%
Inspection	15	1.5%
Incident	6	0.6%

Question-type distribution

4.3.2 Example Question

```
{
  "mc_question": "In Michigan, what is the required service interval and periodic test
    voltage for a fiberglass live-line tool used for primary employee protection?",
  "mc_options": [
    {"label": "a", "text": "Every year, tested at 50,000 volts per foot for 1 minute.",
      "is_correct": false},
    {"label": "b", "text": "Every 2 years, tested at 100,000 volts per foot for 5
      minutes.", "is_correct": false},
    {"label": "c", "text": "Every year, tested at 100,000 volts per foot for 5
      minutes.", "is_correct": false},
    {"label": "d", "text": "Every 2 years, tested at 75,000 volts per foot for 1
      minute.", "is_correct": true}
  ],
  "mc_correct_answer": "d"
}
```

5. Evaluation Methodology

5.1 Benchmark Harness

Evaluation is performed with a dedicated script that:

- Loads questions from JSON
- Calls the Rosie benchmark endpoint or model provider endpoint
- Records selected answer, correctness, reasoning, retrieval stats
- **Computes metrics:** accuracy and retrieval usage

5.2 Comparative Baselines

We evaluate:

- Rosie (domain expert LLM system)
- GPT-5.6 Sol (OpenAI)
- Gemini 3.5 Flash (Google DeepMind)
- Claude Opus 4.8 (Anthropic)
- Claude Sonnet 5 (Anthropic)

All baselines receive identical multiple-choice prompts.

5.3 Metrics

- **Accuracy:** Percentage of correct answers
- **Retrieval statistics:** Average docs retrieved, percentage using context
- **Statistical significance:** Bootstrap confidence intervals and McNemar's test (alpha=0.05)

6. Results

6.1 Overall Performance

System	Accuracy	95% CI	Notes
Rosie	95.0%	[93.65%, 96.29%]	Hybrid RAG
GPT-5.6 Sol	87.0%	[84.95%, 89.05%]	Zero-shot
Gemini 3.5 Flash	86.3%	[84.26%, 88.37%]	Zero-shot
Claude Opus 4.8	78.5%	[75.95%, 80.94%]	Zero-shot
Claude Sonnet 5	74.1%	[71.46%, 76.74%]	Zero-shot

Table 1: Overall benchmark results

6.2 Performance Visualizations

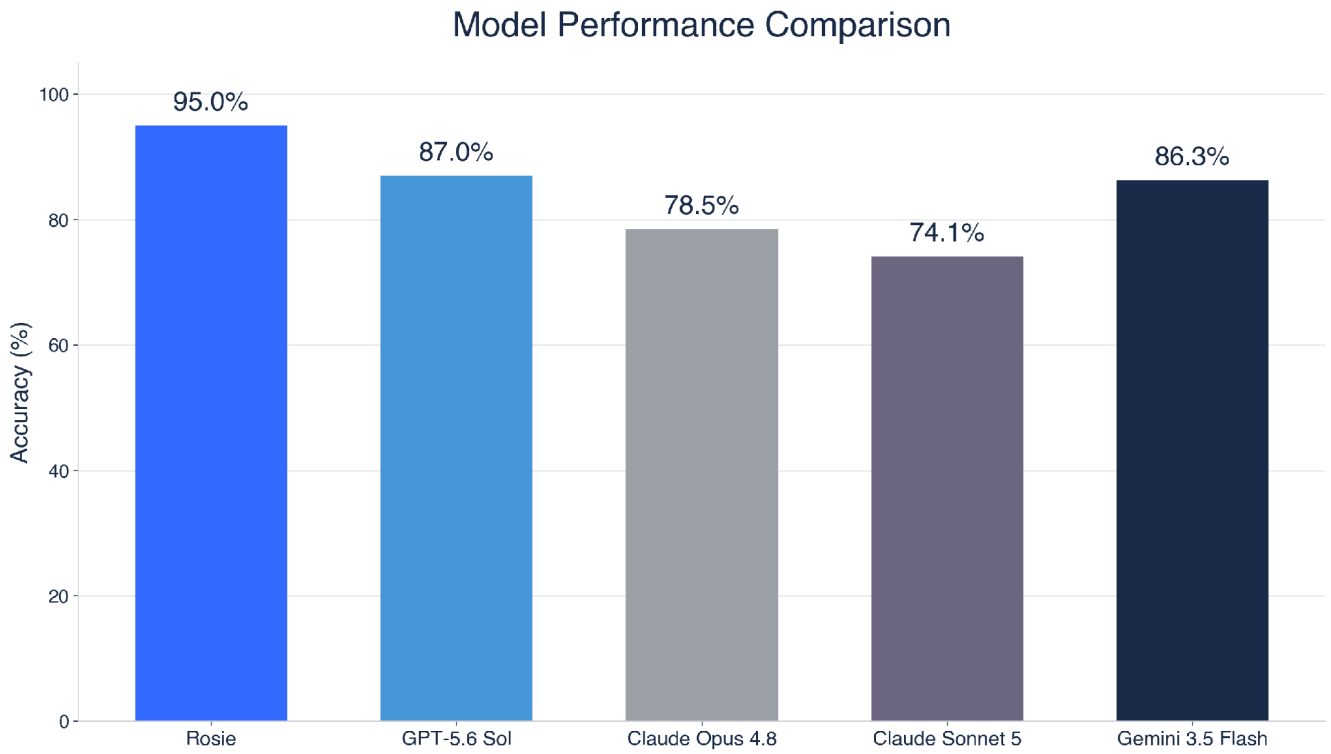

Figure 1: Overall model accuracy comparison across all 1,023 questions.

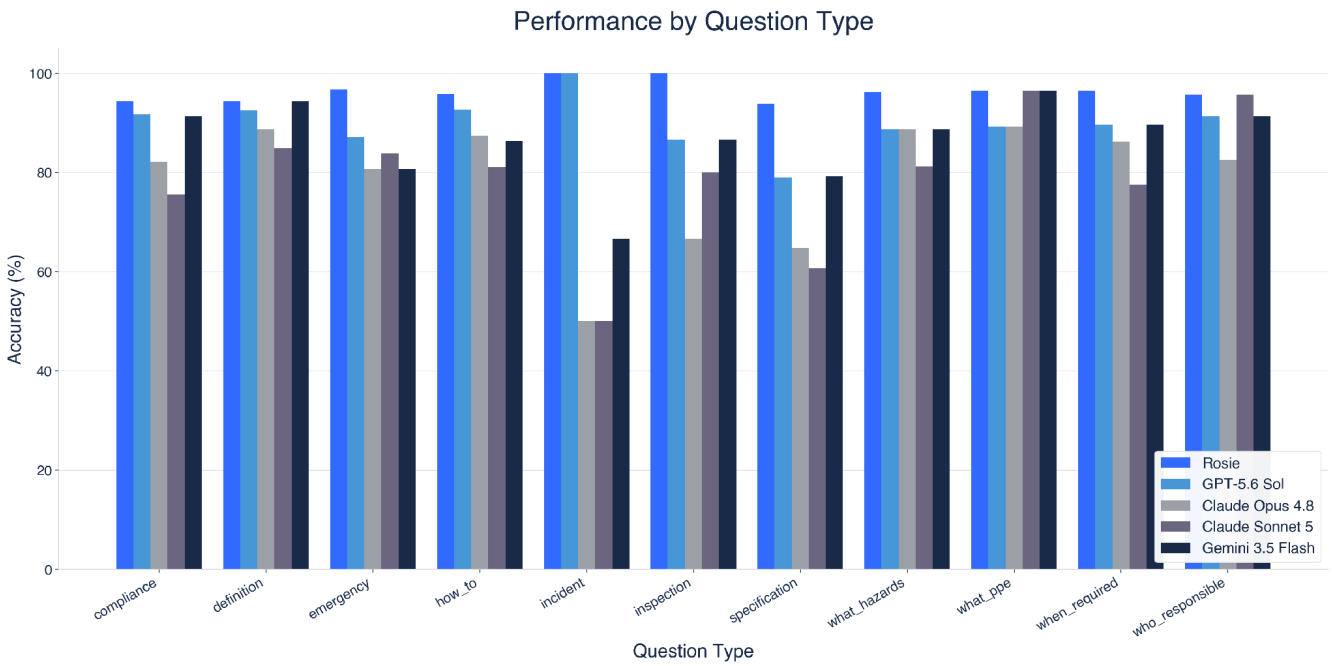

Figure 2: Model accuracy by question type. Rosie is consistent across categories; baselines struggle with specification questions (60.8%-79.3% for baselines versus 93.8% for Rosie).

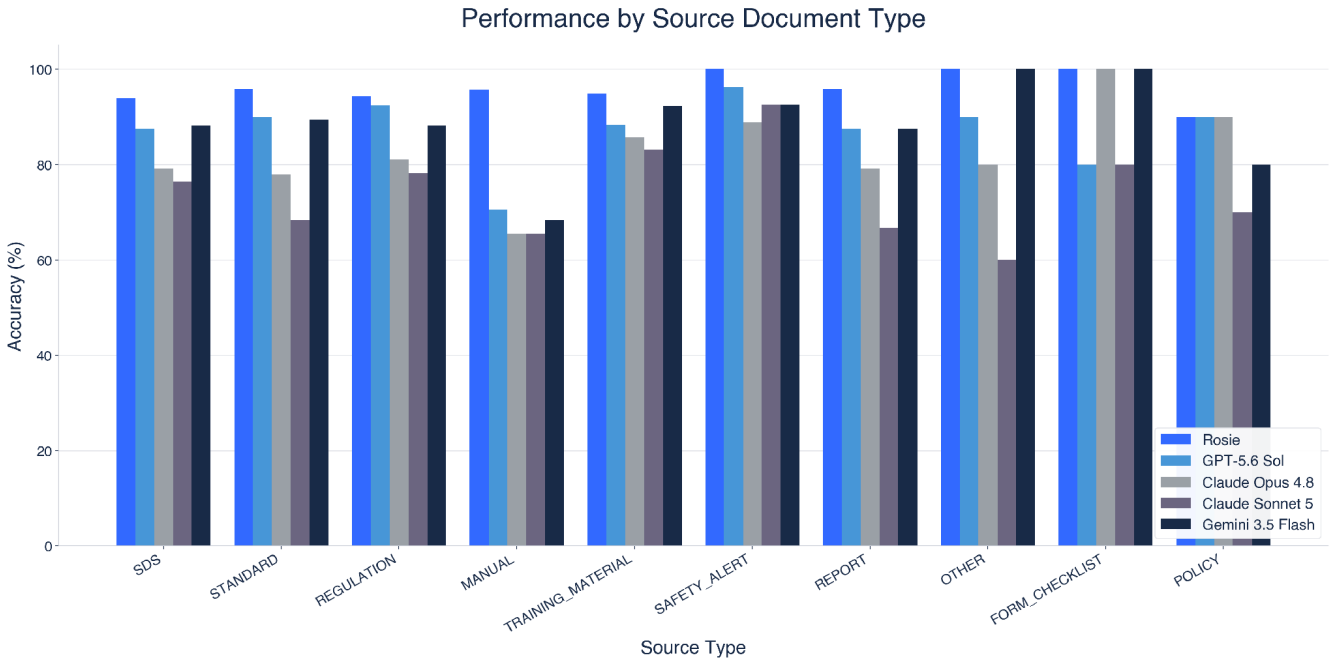

Figure 3: Model accuracy by source document type. Baselines degrade on equipment-manual questions (65.5%-70.5%) compared with Rosie (95.7%).

6.3 Statistical Significance

All pairwise comparisons between Rosie and baseline models show highly significant differences (McNemar's test, $p < 0.001$):

Comparison	Rosie Wins	Baseline Wins	Net Advantage	p-value
vs GPT-5.6 Sol	119	37	+82	$p < 0.001$

Comparison	Rosie Wins	Baseline Wins	Net Advantage	p-value
vs Gemini 3.5 Flash	129	40	+89	$p < 0.001$
vs Claude Opus 4.8	207	38	+169	$p < 0.001$
vs Claude Sonnet 5	245	31	+214	$p < 0.001$

Table 2: McNemar pairwise comparisons

7. Analysis

7.1 Domain Adaptation Impact

Rosie's hybrid RAG architecture demonstrates substantial improvements over frontier LLMs:

- +8.0 pp over GPT-5.6 Sol (95.01% vs 87.00%)
- +8.7 pp over Gemini 3.5 Flash (95.01% vs 86.31%)
- +16.5 pp over Claude Opus 4.8 (95.01% vs 78.49%)
- +20.9 pp over Claude Sonnet 5 (95.01% vs 74.10%)

These improvements are consistent across question types, with Rosie achieving 90%+ accuracy on every category (lowest: specification at 93.8%).

7.2 Category-Specific Performance

7.2.1 Specification Questions (31.7% of benchmark)

Rosie excels at questions requiring precise technical values (93.8% accuracy), while baseline models struggle significantly:

- **GPT-5.6 Sol:** 79.0% (-14.8 pp)
- **Gemini 3.5 Flash:** 79.3% (-14.5 pp)
- **Claude Opus 4.8:** 64.8% (-29.0 pp)
- **Claude Sonnet 5:** 60.8% (-33.0 pp)

This category is the highest failure mode for general-purpose LLMs, which often hallucinate plausible but incorrect numerical values.

7.2.2 Manual-Based Questions

Documents requiring procedural knowledge show the largest gaps:

- **Rosie:** 95.7%
- **GPT-5.6 Sol:** 70.5% (-25.2 pp)
- **Gemini 3.5 Flash:** 68.4% (-27.3 pp)
- **Claude Opus 4.8:** 65.5% (-30.2 pp)
- **Claude Sonnet 5:** 65.5% (-30.2 pp)

7.3 Error Analysis

Rosie's residual errors (approximately 5% of questions) cluster around genuinely ambiguous specifications with multiple plausible interpretations and edge cases involving conflicting jurisdictional requirements.

Baseline LLM errors predominantly involve hallucinated values-inventing plausible but incorrect specifications-and general-knowledge substitution, where models rely on common industry practice instead of the document-specific requirement being asked about.

8. Transparency & Data Availability

8.1 What We Publish

- Full benchmark dataset (1,023 questions) with question text, options, correct answers, and metadata
- Complete evaluation harness (evaluate_benchmark.py) and configuration files
- Statistical analysis code with bootstrap confidence intervals and McNemar's tests

8.2 What We Do Not Publish

- Source document content or page snippets (proprietary corpus)
- Document titles and internal database identifiers
- Raw document files or training data

Note: Questions are designed to be answerable with publicly available safety documentation (SDS, OSHA regulations, equipment manuals) to enable independent evaluation.

8.3 Data Availability

The SALUS-SafetyQA benchmark is released under CC-BY-4.0:

- **Dataset & Code:** https://github.com/Salus-Technologies/SALUS-SafetyQA
- **License:** Creative Commons Attribution 4.0 International

8.4 Limitations

- **English-heavy corpus:** Incomplete coverage for non-US regulations
- **Domain shift risk:** New standards and equipment revisions may not be represented
- **Multiple-choice format:** May overestimate performance relative to open-ended extraction
- **Temporal constraints:** Safety regulations and standards change over time

9. Conclusion

We present Rosie, a domain-specialized construction safety AI system, and introduce the SALUS-SafetyQA Benchmark containing 1,023 expert-validated questions. Results show that tailored RAG pipelines significantly outperform general-purpose LLMs in high-stakes safety domains, with 8-21 percentage point improvements and 100% document grounding.

The benchmark reveals systematic failure modes in frontier LLMs: hallucination of plausible but incorrect specifications and poor performance on equipment manuals. These findings demonstrate the critical need for domain-specific systems in safety-critical applications.

Future work includes expanding jurisdictional coverage, incorporating multimodal inputs (drawings, diagrams), and developing open-ended evaluation protocols.

References

[1] Lewis, P., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020. https://papers.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[2] Guu, K., et al. REALM: Retrieval-Augmented Language Model Pre-Training. ICML, 2020. https://proceedings.mlr.press/v119/guu20a.html

[3] Karpukhin, V., et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP, 2020. https://aclanthology.org/2020.emnlp-main.550/

[4] Khattab, O., and Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. SIGIR, 2020. https://doi.org/10.1145/3397271.3401075

[5] Ma, X., et al. Query Rewriting for Retrieval-Augmented Large Language Models. arXiv:2305.14283, 2023. https://arxiv.org/abs/2305.14283

[6] Ma, X., et al. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. arXiv:2310.08319, 2023. https://arxiv.org/abs/2310.08319

[7] Nogueira, R., and Cho, K. Passage Re-ranking with BERT. arXiv:1901.04085, 2019. https://arxiv.org/abs/1901.04085

[8] Singhal, K., et al. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:2305.09617, 2023. https://arxiv.org/abs/2305.09617

[9] Zheng, C., et al. A Reasoning-Focused Legal Retrieval Benchmark. arXiv:2505.03970, 2025. https://arxiv.org/abs/2505.03970

[10] Zhang, F., et al. A hybrid structured deep neural network with Word2Vec for construction accident causes classification. International Journal of Construction Management, 2019. https://doi.org/10.1080/15623599.2019.1683692

[11] Tixier, A. J.-P., et al. Application of Machine Learning to Construction Injury Prediction. Automation in Construction, 2016. https://doi.org/10.1016/j.autcon.2016.05.016

[12] Fang, Q., et al. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. Automation in Construction, 2018. https://doi.org/10.1016/j.autcon.2017.09.018

[13] Gokdemir, O., et al. Automated MCQA Benchmarking at Scale: Evaluating Reasoning Traces as Retrieval Sources for Domain Adaptation of Small Language Models. SC Workshops '25, 2025. https://doi.org/10.1145/3731599.3767410