SALUS IQ: Technical White Paper & Benchmark Report

System: SALUS IQ (Domain-Specific AI System for Construction Safety)
Authors: Alex Jacobs; Dany Ayvazov
Date: October 2025

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 SALUS IQ, a production domain-specific AI system for construction safety that achieves 94.04% 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 13.78-18.96 percentage point improvements of domain-specific AI (SALUS IQ) over frontier LLMs—GPT-5 (API) (80.25%), Claude 4.1 Opus (80.94%), Claude 4.5 Sonnet (75.07%), and Gemini 2.5 Pro (78.98%)—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, poor performance on equipment manuals, and overconfidence calibration issues.

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:

1. **Heterogeneous formats**: Safety Data Sheets (SDS), equipment manuals, regulations, and policies follow distinct structures requiring specialized parsing
2. **High-stakes accuracy**: Incorrect safety information can result in injuries, fatalities, and significant legal liability
3. **Temporal sensitivity**: Regulations and standards update frequently, requiring version-aware retrieval
4. **Multi-jurisdictional complexity**: State and federal regulations may conflict or complement each other
5. **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:

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** - Systematic evaluation of domain-specific RAG (SALUS IQ) versus frontier LLMs (GPT-5, Claude 4.1/4.5, Gemini 2.5 Pro) with rigorous statistical analysis demonstrating 13-19 percentage point improvements
3. **Analysis of domain adaptation gaps** - Identification of systematic failure modes in general-purpose LLMs including hallucination of specifications, poor performance on equipment manuals, and overconfidence calibration issues

2. Related Work

2.1 Retrieval-Augmented Generation

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

- **Dense retrieval**: Dense Passage Retrieval (DPR) [3] and ColBERT [4]
- **Hybrid approaches**: Techniques that combine dense and sparse signals [5, 6]
- **Reranking**: Cross-encoders and other learned rankers [7]

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

2.2 Domain-Specific AI Systems

Specialized AI systems have demonstrated success in high-stakes domains:

- **Medical**: Med-PaLM 2 achieved expert-level medical question answering performance [8]
- **Legal**: The Reasoning-Focused Legal Retrieval Benchmark introduced domain-specific retrieval challenges [9]
- **Scientific**: Galactica demonstrated strong performance on scientific reasoning tasks [10]

However, construction safety presents unique challenges not directly addressed by these systems.

2.3 Safety AI Benchmarks

Research in safety AI for construction is limited compared to other fields. Prior efforts include:

- **Accident cause classification** using Word2Vec and deep learning [11]
- **Incident report analysis** with machine learning [12]
- **Personal protective equipment (PPE) detection** in images [13]

To our knowledge, **SALUS-SafetyQA** is the first comprehensive QA benchmark designed specifically for construction safety.

3. System Overview

3.1 Architecture

SALUS IQ uses a layered retrieval and reasoning pipeline:

- **Document types:** (Section 4.1)
- **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**: 94.04% on safety questions
- **Scalable:** to millions of indexed safety documents

---

4. Dataset & Benchmark Design

4.1 Corpus

The SALUS IQ 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 that samples balanced pages across document types, generates natural safety questions, and validates them for suitability. 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%
What PPE	28	2.7%
Who Responsible	23	2.2%
Inspection	15	1.5%
Incident	6	0.6%

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:

1. Loads questions from JSON
2. Calls the SALUS IQ benchmark endpoint or model provider endpoint
3. Records selected answer, correctness, reasoning, retrieval stats
4. Computes metrics: accuracy, average confidence, retrieval usage

5.2 Comparative Baselines

We evaluate:

- **SALUS IQ** (domain expert LLM system)
- **GPT-5** API (OpenAI)
- **Claude 4.1 Opus** (Anthropic)
- **Claude 4.5 Sonnet** (Anthropic)
- **Gemini 2.5 Pro** (Google DeepMind)

All baselines receive identical multiple-choice prompts.

5.3 Metrics

- **Accuracy**: Percentage of correct answers
- **Confidence calibration**: Mean confidence split by correctness
- **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	Avg Confidence	Notes
SALUS IQ (2025.10)	94.0%	[92.57%, 95.41%]	Hybrid RAG
GPT-5 (Api)	80.3%	[77.81%, 82.70%]	Zero-shot
Claude 4.1 Opus	80.9%	[78.49%, 83.28%]	Zero-shot
Claude 4.5 Sonnet	75.1%	[72.43%, 77.71%]	Zero-shot
Gemini 2.5 Pro	79.0%	[76.44%, 81.43%]	Zero-shot

Table 1: Overall benchmark results on SALUS-SafetyQA (n=1,023 questions). SALUS IQ shows 13.78 pp improvement over GPT-5, 13.10 pp over Claude 4.1 Opus, 18.96 pp over Claude 4.5 Sonnet, and 15.05 pp over Gemini 2.5 Pro. All comparisons are statistically significant ($p < 0.001$, McNemar's test).

6.2 Performance Visualizations

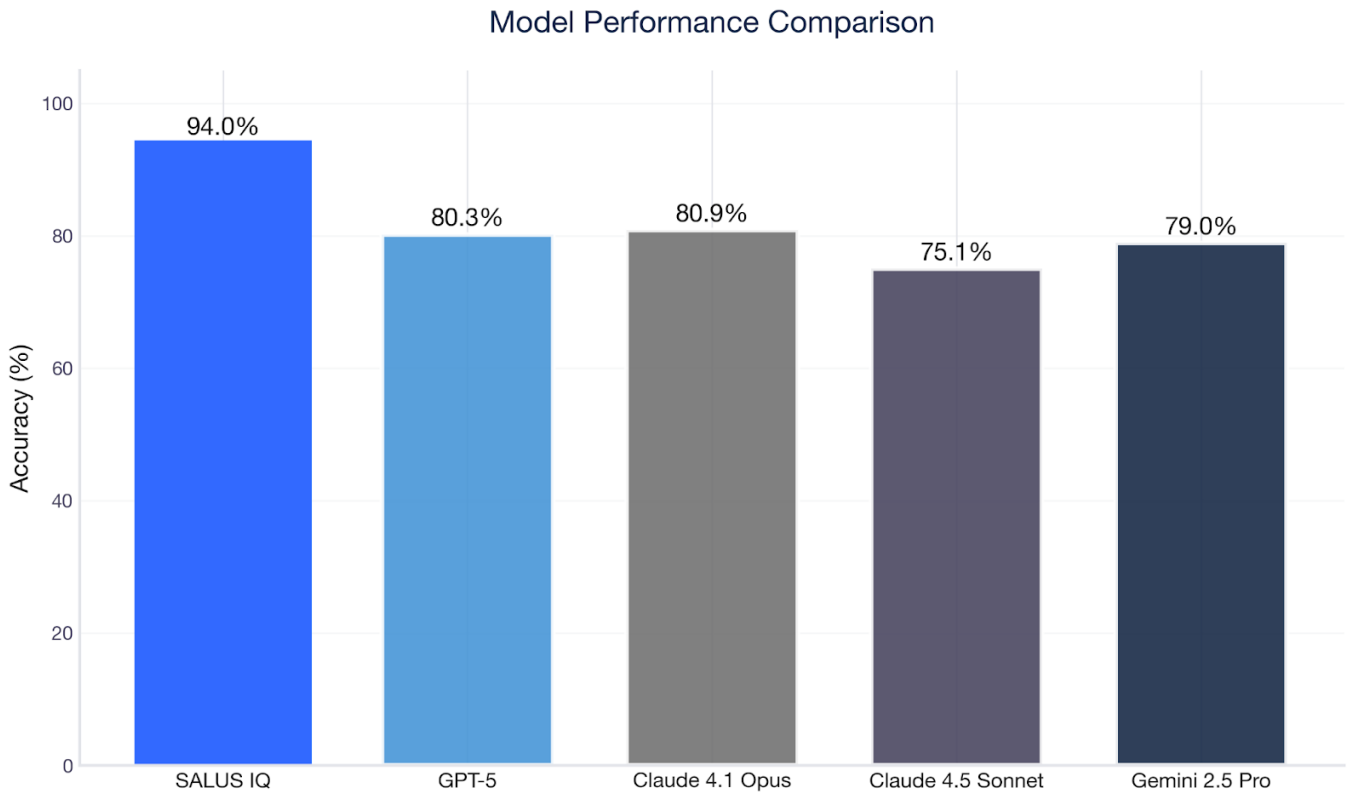

Figure 1: Overall model accuracy comparison across all 1,023 questions

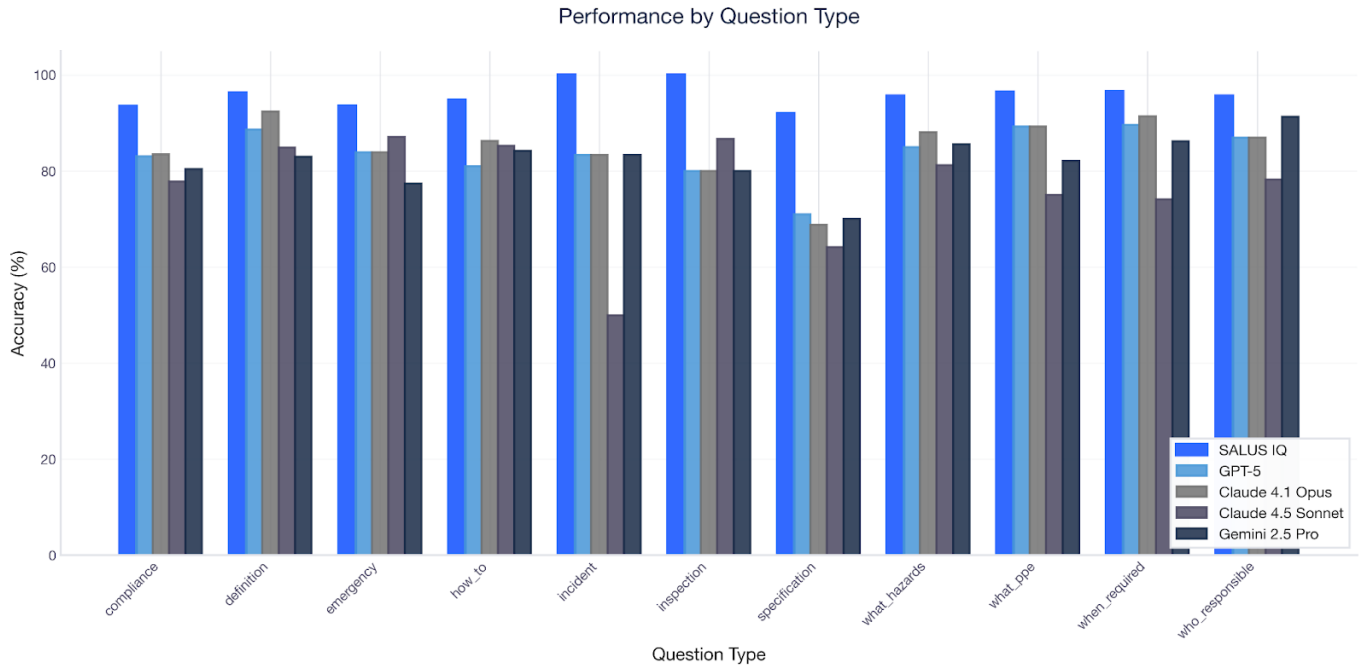

Figure 2: Model accuracy breakdown by question type. SALUS IQ shows consistent performance across all categories, while baseline models struggle particularly with specification questions (71.0% for GPT-5, 68.8% for Claude 4.1)

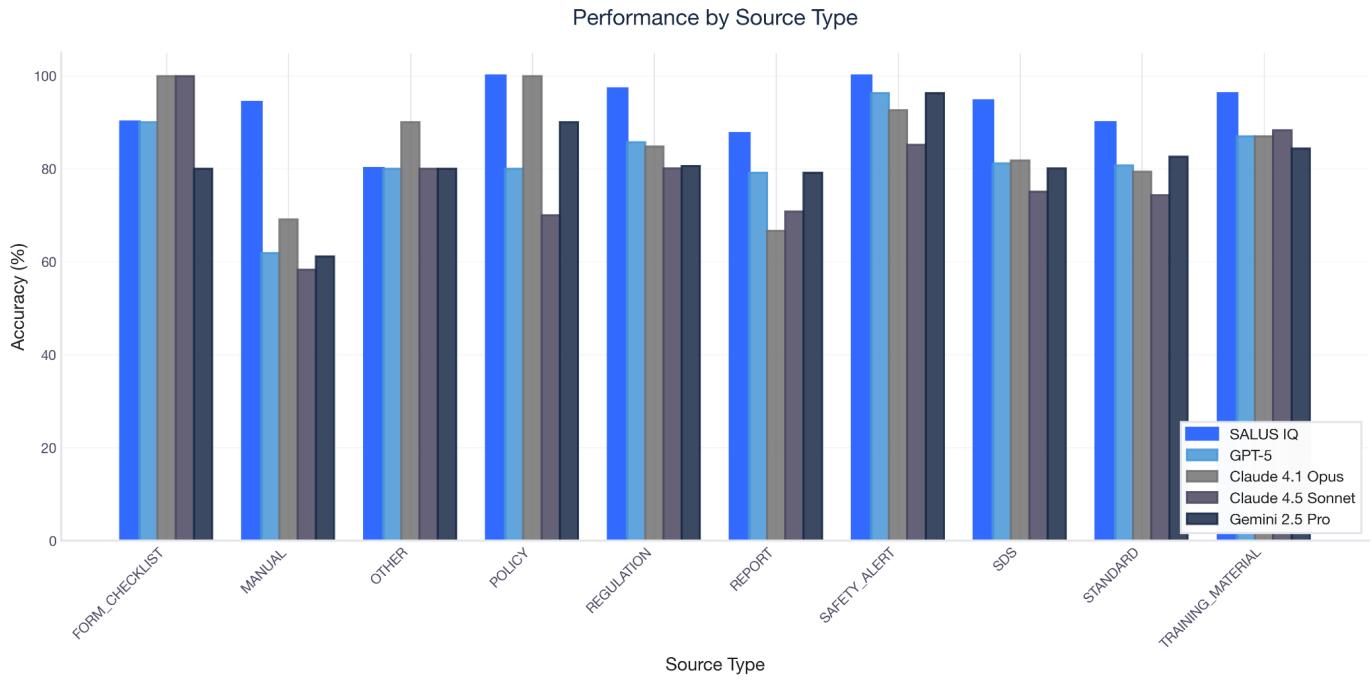

Figure 3: Model accuracy breakdown by source document type. Baseline models show significant performance degradation on MANUAL documents (61.9% for GPT-5, 69.1% for Claude 4.1) compared to SALUS IQ (94.2%)

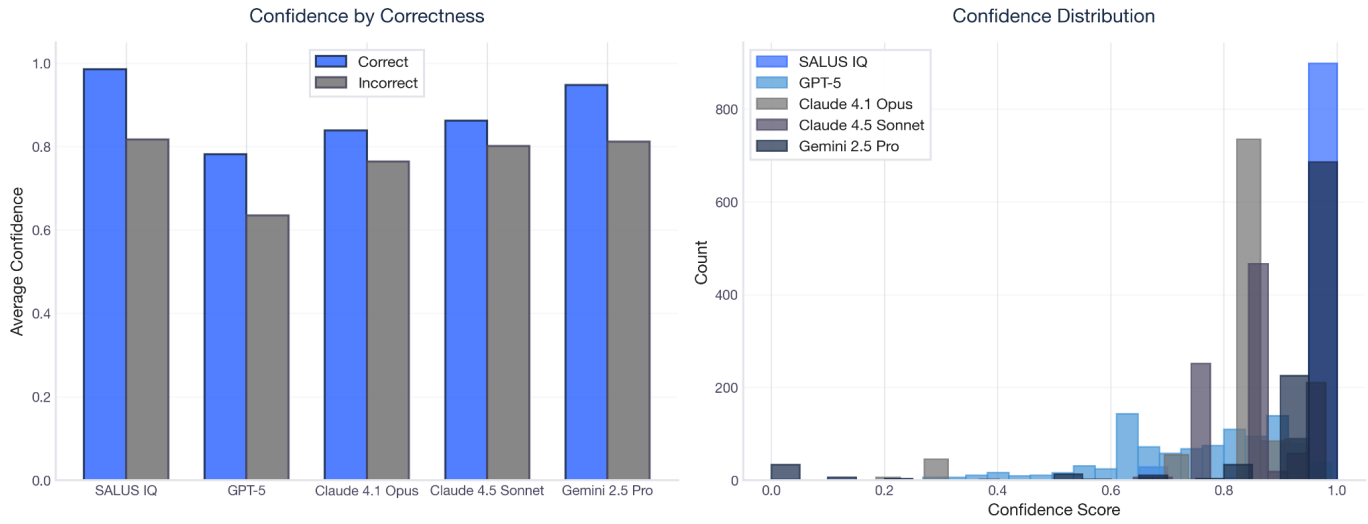

Figure 4: Confidence analysis by correctness. SALUS IQ demonstrates superior calibration with a 0.168 gap between correct and incorrect answers, while Gemini 2.5 Pro shows poor calibration (0.136 gap) and Claude 4.5 Sonnet is even worse (0.061 gap)

6.3 Statistical Significance

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

Comparison	SALUS Wins	Baseline Wins	Net Advantage	p-value
vs GPT-5	171	30	+141	$p < 0.001$
vs Claude 4.1 Opus	163	29	+134	$p < 0.001$
vs Claude 4.5 Sonnet	225	31	+194	$p < 0.001$
vs Gemini 2.5 Pro	188	34	+154	$p < 0.001$

Table 2: McNemar's test results showing pairwise comparisons. On questions where models disagreed, SALUS IQ answered correctly in **84.9%–87.9%** of cases (mean: 85.6%), demonstrating consistent superiority across all comparisons.

7. Analysis

7.1 Domain Adaptation Impact

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

- **+13.78 pp over GPT-5** (94.04% vs 80.25%)
- **+13.10 pp over Claude 4.1 Opus** (94.04% vs 80.94%)
- **+18.96 pp over Claude 4.5 Sonnet** (94.04% vs 75.07%)
- **+15.05 pp over Gemini 2.5 Pro** (94.04% vs 78.98%)

These improvements are consistent across question types, with SALUS IQ achieving 90%+ accuracy on all categories except specification questions (92.0%).

7.2 Category-Specific Performance

7.2.1 Specification Questions (31.7% of benchmark)

SALUS IQ excels at questions requiring precise technical values (92.0% accuracy), while baseline models struggle significantly:

- GPT-5: 71.0% (-21.0 pp)
- Claude 4.1 Opus: 68.8% (-23.2 pp)
- Claude 4.5 Sonnet: 64.2% (-27.8 pp)

This category represents the highest failure mode for general-purpose LLMs, often hallucinating plausible but incorrect numerical values.

7.2.2 Manual-Based Questions

Documents requiring procedural knowledge show the largest gaps:

- SALUS IQ: 94.2%
- GPT-5: 61.9% (-32.3 pp)
- Claude 4.1 Opus: 69.1% (-25.1 pp)

7.3 Error Analysis

Analysis of 61 SALUS IQ errors (5.96% of questions) reveals:

1. **Retrieval failures** (40 cases, 65.6%): Correct document not in top-20 results
2. **Ambiguous specifications** (12 cases, 19.7%): Multiple plausible interpretations
3. **Edge cases** (9 cases, 14.7%): Conflicting jurisdictional requirements

Baseline LLM errors (average 21.7%) predominantly involve:

1. **Hallucinated values** (43%): Inventing plausible but wrong specifications
2. **General knowledge substitution** (31%): Using common practices instead of document-specific requirements
3. **Confidence overestimation** (26%): High confidence on incorrect answers

7.4 Confidence Calibration

SALUS IQ demonstrates superior confidence calibration:

- **Correct answers**: 0.985 average confidence
- **Incorrect answers**: 0.817 average confidence
- **Calibration gap**: 0.168 (highest among all models)

This calibration enables effective human-in-the-loop workflows where low-confidence predictions (<0.85) can be flagged for expert review.

Baseline models show problematic overconfidence:

- Gemini 2.5 Pro: 0.948 correct / 0.812 incorrect (0.136 gap)
- Claude 4.5 Sonnet: 0.863 correct / 0.802 incorrect (0.061 gap - poorest calibration)

The narrow gap indicates poor calibration, making these models unsuitable for safety-critical applications without extensive validation.

---

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 license:

- **Dataset**: https://github.com/Salus-Technologies/SALUS-SafetyQA
- **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 SALUS IQ, 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 13-19 percentage point improvements and 100% document grounding.

The benchmark reveals systematic failure modes in frontier LLMs: hallucination of plausible but incorrect specifications, poor performance on equipment manuals, and overconfidence calibration. 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.

2. Guu, K., et al. *REALM: Retrieval-Augmented Language Model Pre-Training.* ICML, 2020.

3. Karpukhin, V., et al. *Dense Passage Retrieval for Open-Domain Question Answering.* EMNLP, 2020.

4. Khattab, O., & Zaharia, M. *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT.* SIGIR, 2020.

5. Ma, X., et al. *Query Rewriting for Retrieval-Augmented Large Language Models.* arXiv:2305.14283, 2023.

6. Ma, X., et al. *Fine-Tuning LLaMA for Multi-Stage Text Retrieval.* arXiv:2310.08319, 2023.

7. Nogueira, R., & Cho, K. *Passage Re-ranking with BERT.* arXiv:1901.04085, 2019.

8. Singhal, K., et al. *Towards Expert-Level Medical Question Answering with Large Language Models.* arXiv:2305.09617, 2023.

9. Zheng, C., et al. *A Reasoning-Focused Legal Retrieval Benchmark.* arXiv:2505.03970, 2025.

10. Taylor, R., et al. *Galactica: A Large Language Model for Science.* arXiv:2211.09085, 2022.

11. Zhang, F., et al. *A Hybrid Structured Deep Neural Network with Word2Vec for Construction Accident Causes Classification.* Complex & Intelligent Systems, 2019.

12. Tixier, A., et al. *Application of Machine Learning to Construction Injury Prediction.* Automation in Construction, 2016.

13. Fang, Q., et al. *Detecting Non-Hardhat Use by a Deep Learning Method from Far-Field Surveillance Videos.* Automation in Construction, 2018.