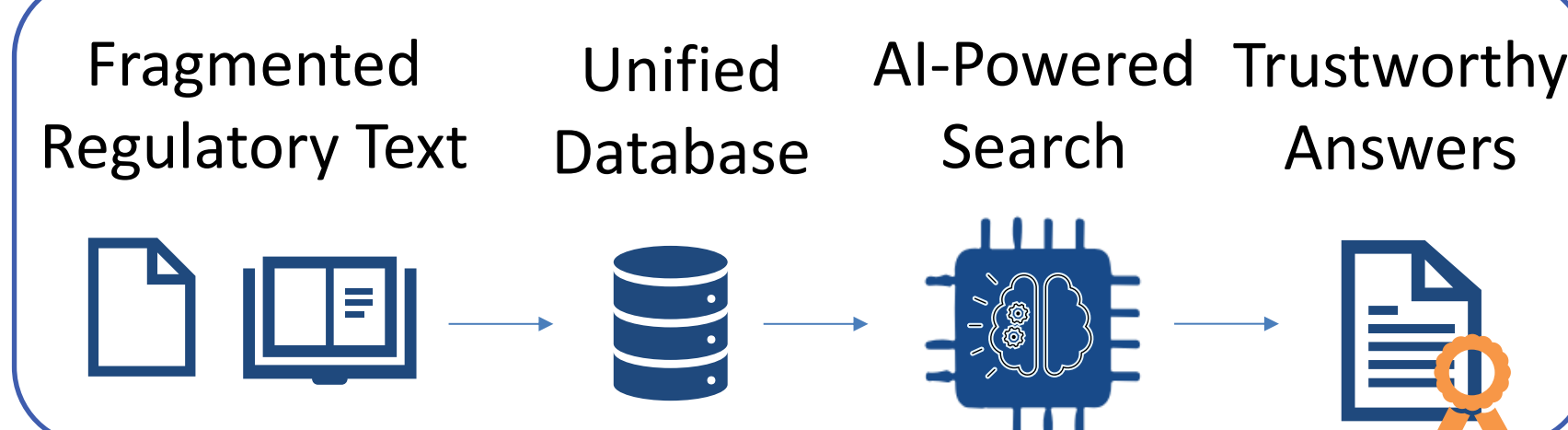
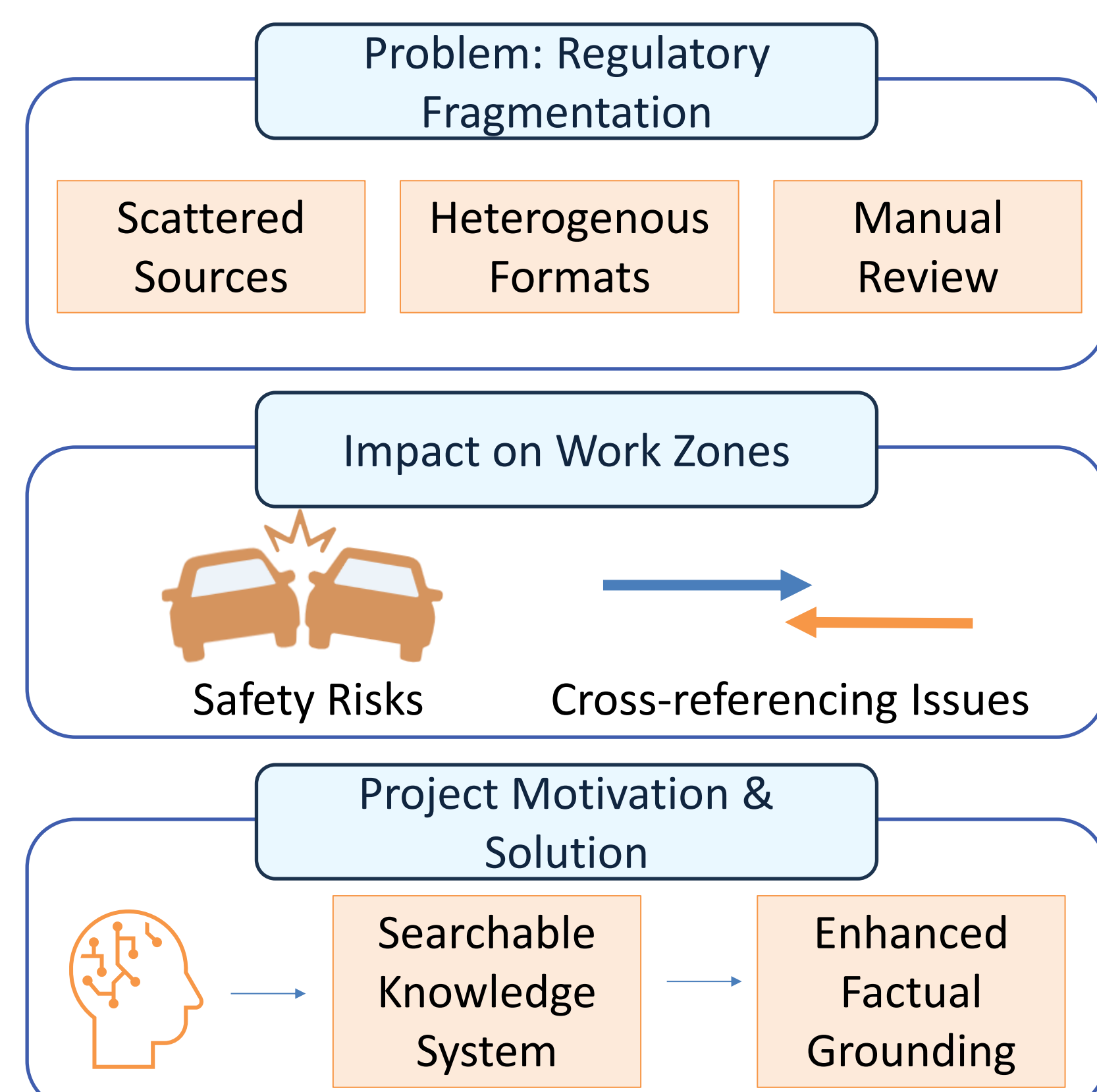


Executive Summary

To address the challenge of fragmented regulations that complicate timely access to safety requirements and increase the risk of human error, this project presents a unified, multimodal, graph-augmented, retrieval-augmented generation (RAG) system that provides a structured and auditable retrieval for Texas work zone safety regulations.



Background and Motivation



Objective

Objective: Build a searchable vector database for rapid, accurate, and traceable work zone safety regulation retrieval in Texas

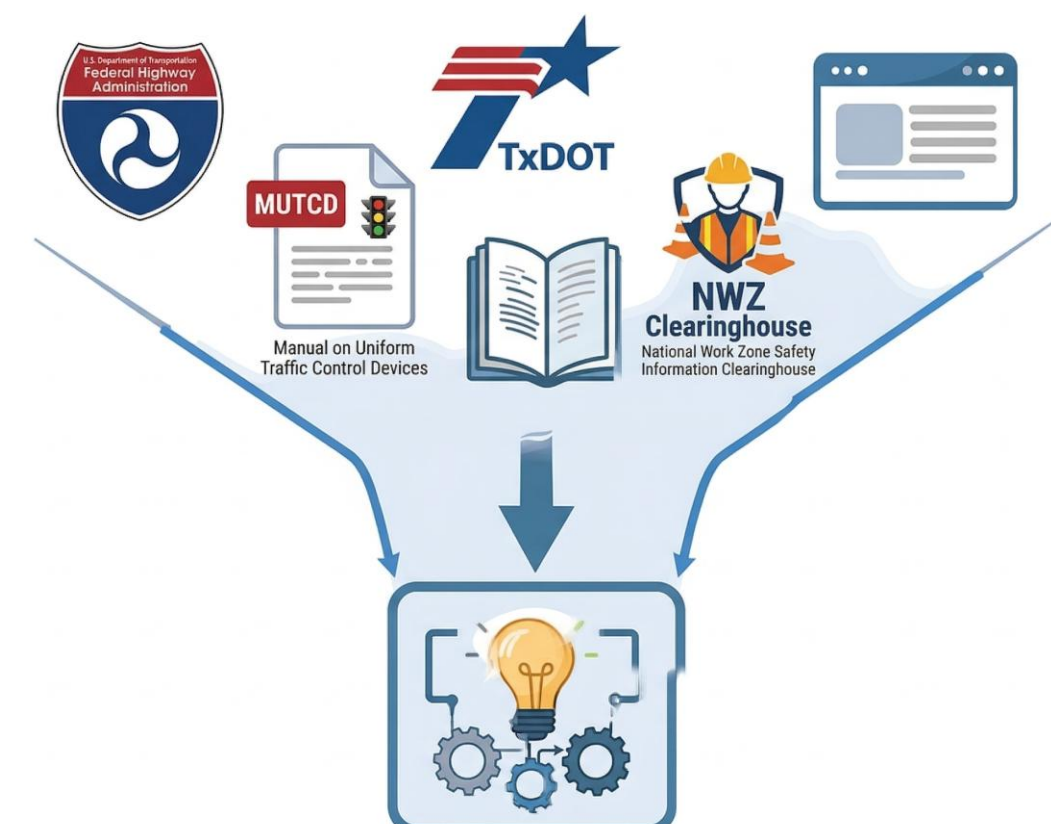


Figure 1. A Centralized Work Zone Safety Regulatory System

Methodology

1. Data Collection



The regulatory corpus was constructed by aggregating 36 authoritative documents from federal and state sources.

2. Document Processing

Text and Layout Extraction

MinerU parsing implements a decoupled vision-language strategy. This involves an initial coarse layout analysis to remove headers, footers, and other structural artifacts. Subsequently, a recognition process extracts dense text, mathematical equations, and structured tables.

Figure & Diagram Transcoding

The Vision Language Model (VLM) (gpt-4o-mini) acts as a vision-to-text transcoder for figures and diagrams. It generates descriptions for visual content, including road work zone scenarios, then fuses it with the textual data streams.

3. Chunking & Metadata Assignment



The text is segmented into distinct atomic chunks with structural constraints to ensure tables and equations remain whole. Precise metadata is assigned and appended to each chunk (including unique doc id, page number, and relevant section heading).

4. Embedding & Knowledge Graph Construction

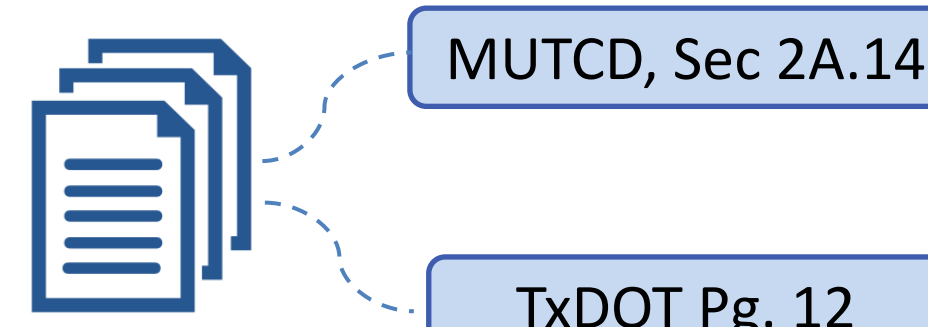
Vector Embedding & Indexing

The system transforms atomic chunks into 3072-dimensional vectors using OpenAI (text-embedding-3-large) model. For multimodal content, a fused embedding is generated by combining the narrative text vector with VLM transcribed descriptions of figures and diagrams.

Knowledge Graph Construction

To enable reasoning across dispersed sections, a graph-based representation is constructed by extracting entities from headers and nodes from glossaries. Logical edges are then inferred from regulatory language and document structures to map cross-references. This allows the system to perform multi-hop reasoning.

5. Retrieval & Generation



The system works by first extracting entities from a user query. It then does both similarity-based search and multi-hop knowledge graph expansion. The model (GPT-4o-mini) finally combines this information to produce a traceable and auditable response.

Results and Discussion

Evaluation and Results

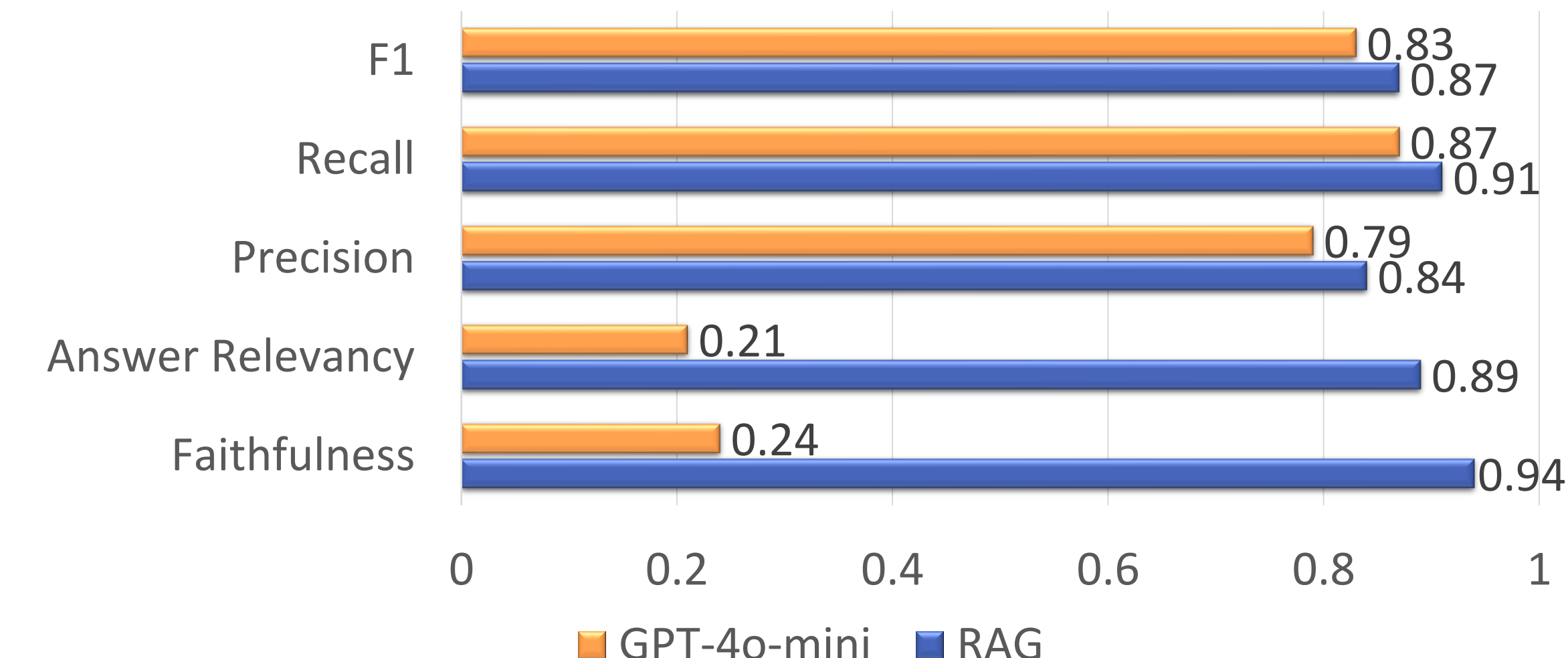
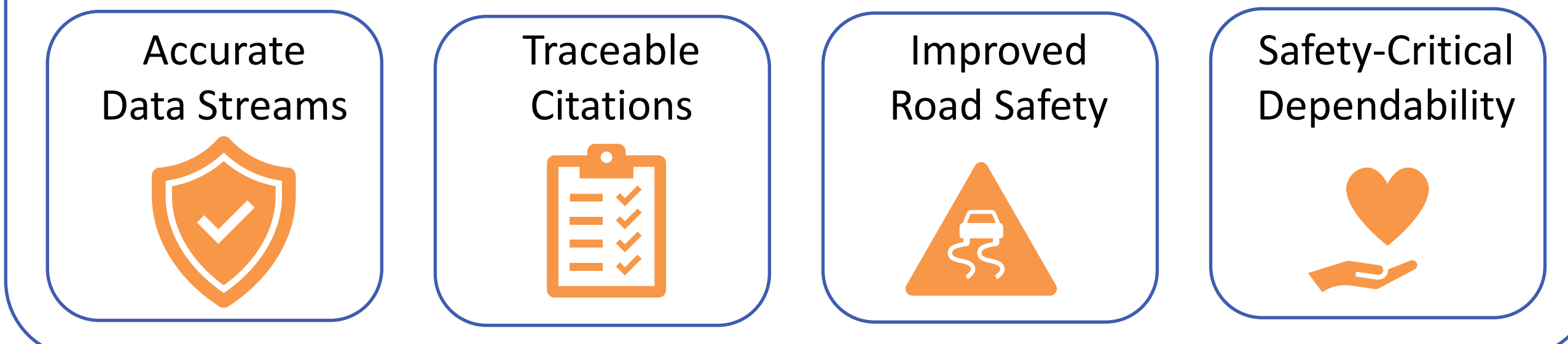


Figure 2. Quantitative Performance Comparison

- **Higher Accuracy:** The RAG system achieved a 0.94 faithfulness score, significantly outperforming the GPT-4o-mini baseline of 0.24.
- **Superior Relevancy:** The system scored 0.89 in answer relevancy, compared to only 0.21 for the baseline model.
- **Strong Semantic Alignment:** Achieved an F1 score of 0.87 (Precision: 0.84, Recall: 0.91), showing high coverage of regulatory intent.
- **Reduced Hallucination:** Retrieval augmentation and graph-based expansion effectively grounded responses in source documents.
- **Full Traceability:** All generated claims included metadata (ID, page, chunk) for auditing and verification.
- **Multi-Hop Reasoning:** The GraphRAG framework successfully connected fragmented requirements spread across different manuals and sections.

Conclusions

A Dependable and Traceable Framework for Work Zone Safety Regulatory Retrieval



Key Abbreviations

- TxDOT - Texas Department of Transportation
- FHWA - Federal Highway Administration
- MUTCD - Manual on Uniform Traffic Control Devices
- NWZS Info - National Work Zone Safety Information Clearinghouse

References

- Guo, Z., Xia, L., Yu, Y., Ao, T., and Huang, C. (2024). "Lightrag: Simple and fast retrieval-augmented generation." *arXiv preprint arXiv:2410.05779*, doi:https://doi.org/10.48550/arXiv.2410.05779.
- Han, H., Wang, Y., Shomer, H., Guo, K., Ding, J., Lei, Y., Halappanavar, M., Rossi, R. A., Mukherjee, S., and Tang, X. (2024). "Retrieval-augmented generation with graphs (graphrag)." *arXiv preprint arXiv:2501.00309*, doi:https://doi.org/10.48550/arXiv.2501.00309.
- Liu, J., Meng, S., Gao, Y., Mao, S., Cai, P., Yan, G., Chen, Y., Bian, Z., Wang, D., and Shi, B. "Aligning vision to language: Annotation-free multimodal knowledge graph construction for enhanced llms reasoning." *Proc., Proceedings of the IEEE/CVF International Conference on Computer Vision*, 981–992, doi:https://doi.org/10.48550/arXiv.2503.12972.
- Federal Highway Administration (FHWA) (2025). "Work Zone Facts and Statistics." Accessed January 16, 2026. https://ops.fhwa.dot.gov/wz/resources/facts_stats.htm.
- Texas Department of Transportation (TxDOT) n.d. "Work Zones." Accessed January 16, 2026. <https://www.txdot.gov/safety/traffic-safety-campaigns/work-zones.html>.