

# On-Premises Natural Language Access to Cross-Document Knowledge

Marco Palumbo, Pietro Ducange, Francesco Marcelloni, Nicola Tonellotto

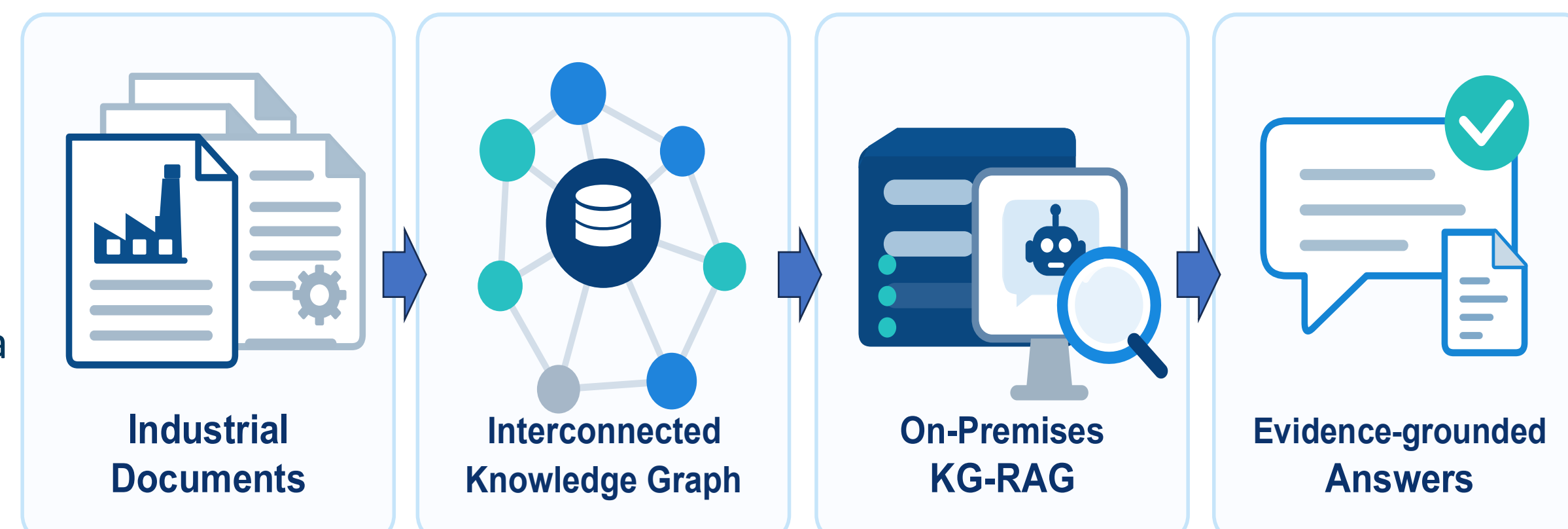
## ABSTRACT

Industrial knowledge may be distributed across multiple related records, with recurring entities and their relationships linking information between them. This research investigates reliable, on-premises natural-language access to such knowledge through a Knowledge Graph-based Retrieval-Augmented Generation (KG-RAG) approach, using police incident reporting as an industrial case study. A controlled evaluation across 7,500 pipeline executions reveals that KG-RAG failures depend on graph composition, may persist despite complete evidence retrieval, and are associated with less selective retrieval contexts. These findings motivate the next research stages: empirically grounded synthetic baselines, controlled failure analysis, targeted interventions, and validation on real industrial Knowledge Graphs.

## PROBLEM AND MOTIVATION

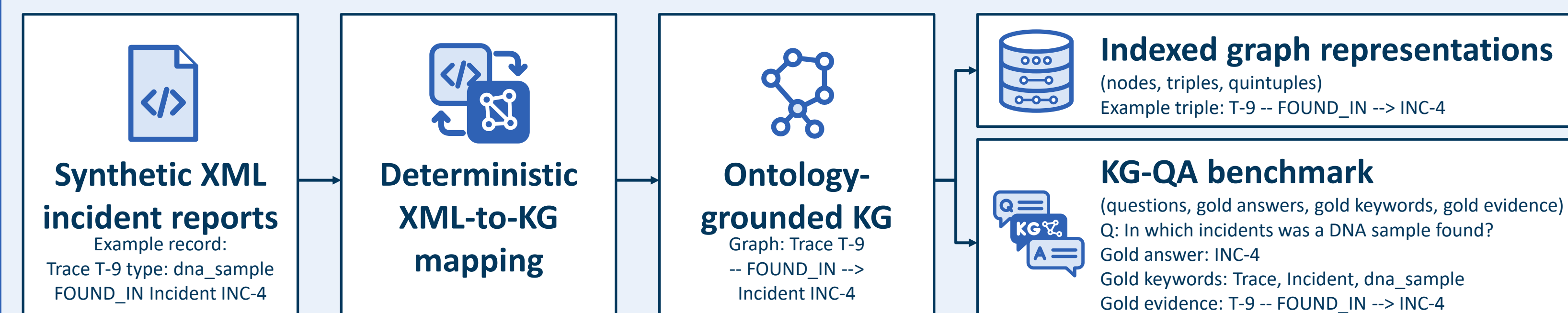
### Industrial case study: Police incident reporting

- Incident reports describe **events** together with **people, vehicles, locations and objects** that may recur across different incidents.
- As a result, relevant knowledge can be distributed across multiple documents, making **conventional document search poorly suited** to questions that depend on relations between different reports.
- Knowledge Graphs** (KGs) explicitly represent these cross-report connections, while a **KG-RAG approach** enables users to query them in natural language by retrieving relevant graph evidence and providing it to a language model.
- In the target industrial setting, however, a useful system must satisfy additional constraints: **sensitive data** require fully **on-premises processing**, and each **answer** must remain **traceable** to its supporting evidence.
- Developing a reliable system under these constraints therefore requires understanding **where failures arise** and **which properties of the graph and retrieved information** contribute to them.

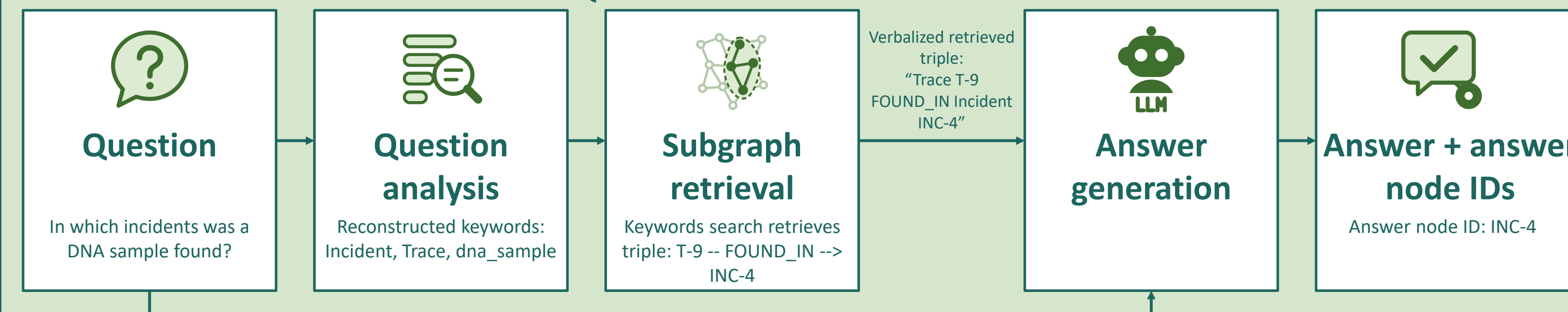


## APPROACH

### OFFLINE DATA AND KNOWLEDGE-GRAPH CONSTRUCTION



### QUERY-TIME KG-RAG PIPELINE



**ON-PREMISES DEPLOYMENT**  
All processing remains within the data owner's infrastructure

### CONTROLLED TEST BED

- Synthetic incident Knowledge Graphs with configurable properties
- For each question: **known correct answer + known supporting graph evidence**

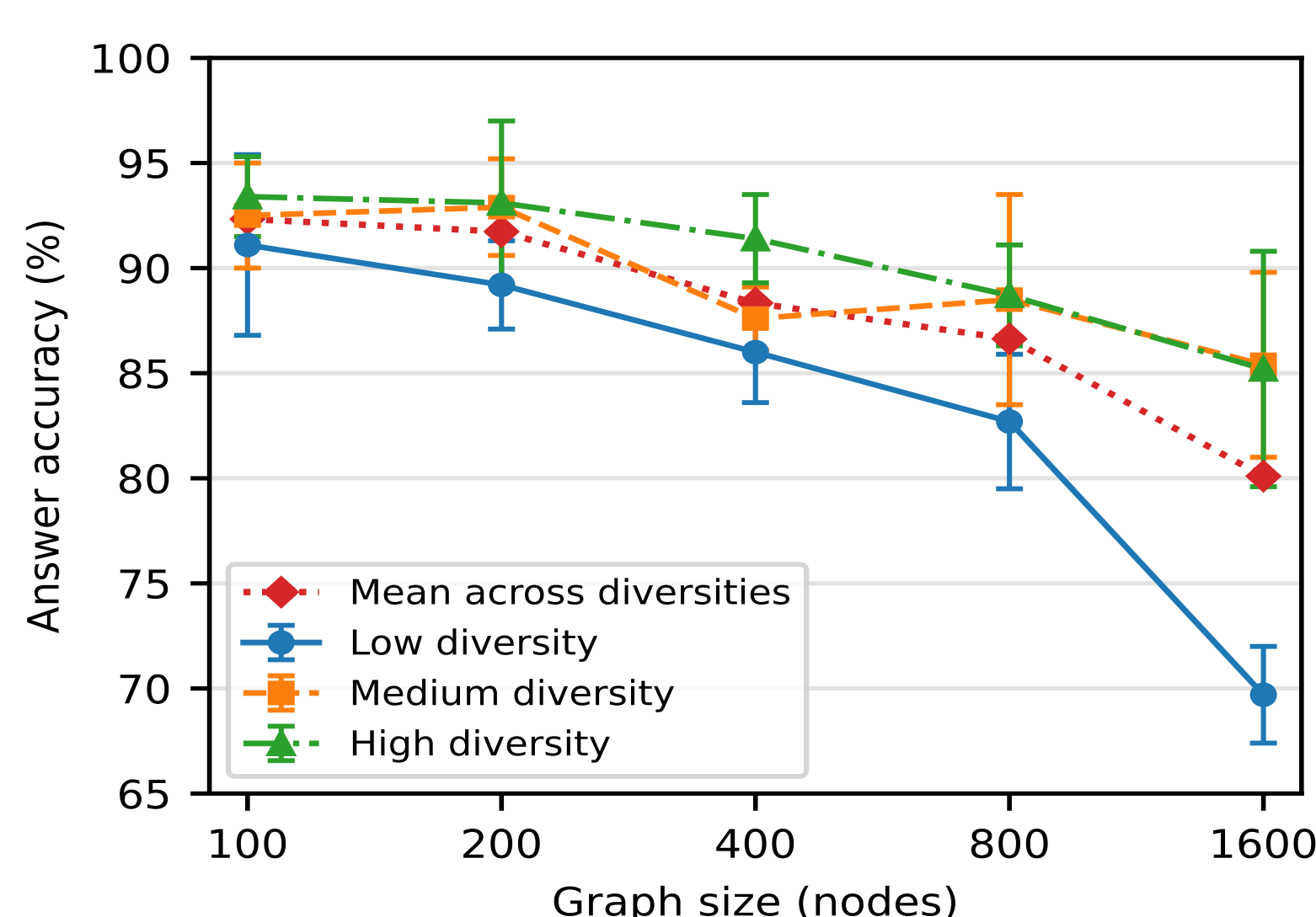
### CONTROLLED CONFIGURATIONS

- Graph size:**  
100 → 200 → 400 → 800 → 1600 nodes
- Entity-catalogue diversity:**  
Low → Medium → High  
(Lower diversity = greater entity recurrence)

### REPEATED DIAGNOSTIC EVALUATION

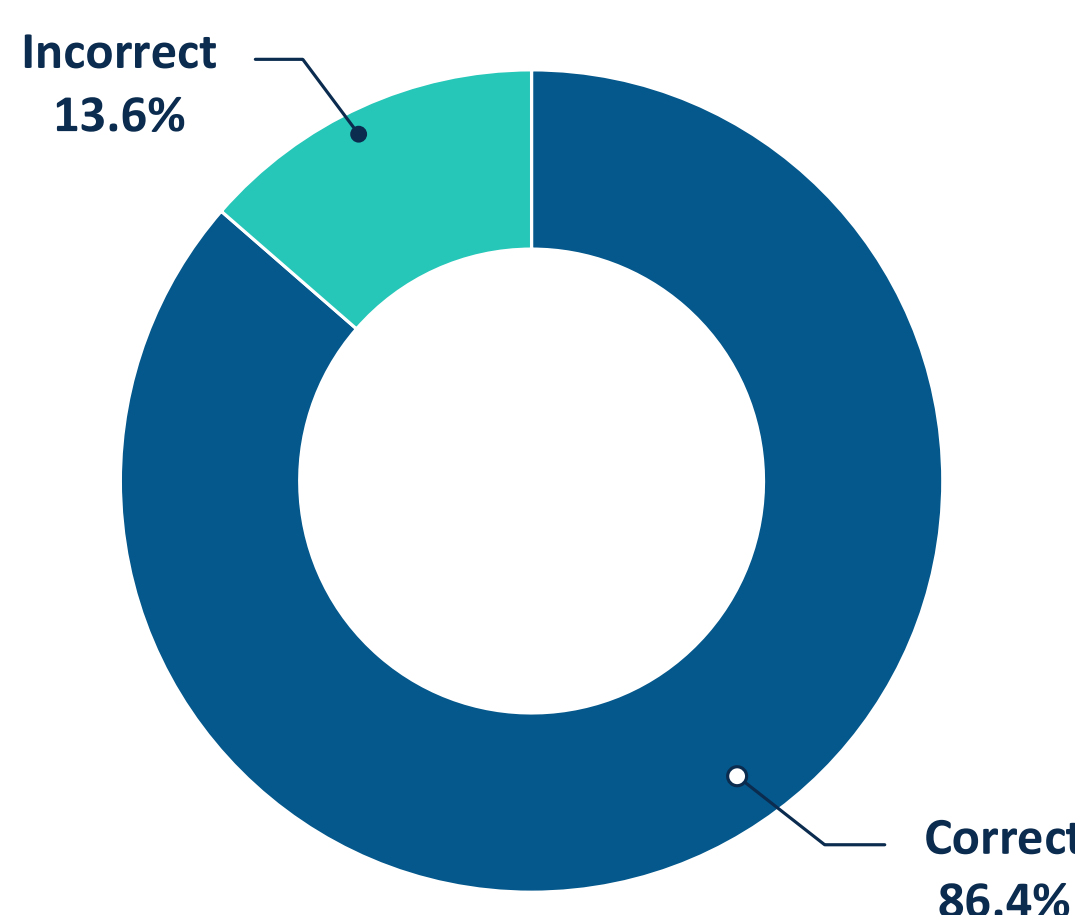
- 15 configurations × 100 questions × 5 replications**
- 7,500 complete pipeline executions**

## RESULTS

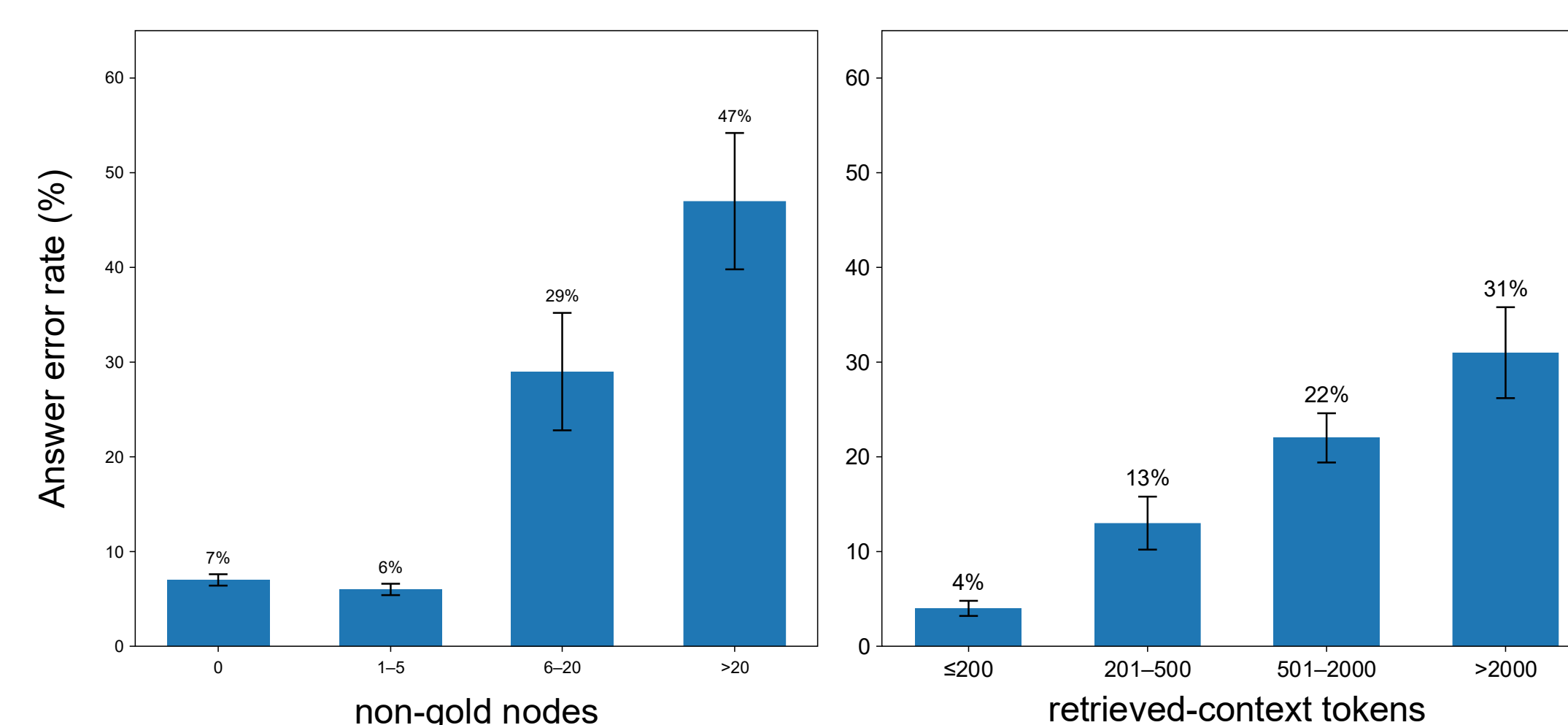


### GRAPH COMPOSITION MATTERS

### Answer correctness under complete gold evidence retrieved



### EVIDENCE RETRIEVED ≠ EVIDENCE USE



### LESS SELECTIVE CONTEXTS → MORE ERRORS

## CONCLUSIONS & FUTURE WORKS

- The observed results reveal consistent failure patterns...

- ...but do not isolate their individual causes.

### 1 - Empirically grounded baseline

Build a synthetic reference setting constrained by distributions observed in real police data.

### 2 - Controlled failure analysis

Independently vary selected graph and retrieved-context properties to identify what drives each failure mode.

### 3 - Targeted interventions

Design retrieval and evidence-selection strategies tailored to the identified failure mechanisms.

### 4 - Industrial validation

Test whether the identified mechanisms and improvements transfer to real industrial KGs.

## REFERENCES

- M. Palumbo, A. Schiavo, P. Ducange, F. Marcelloni, N. Tonellotto, A. La Mantia, "On-premises Natural Language Access to Incident Knowledge Graphs: An Industrial Experience Report," ITADATA 2026: The 5th Italian Conference on Big Data and Data Science, Bari, Italy, 10–12 November 2026.
- B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, S. Tang, "Graph Retrieval-Augmented Generation: A Survey," ACM Transactions on Information Systems, vol. 44, 2025. doi: 10.1145/3777378.
- K. Sharma, P. Kumar, Y. Li, "OG-RAG: Ontology-Grounded Retrieval-Augmented Generation for Large Language Models," in Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025), Association for Computational Linguistics, Suzhou, China, 2025, pp. 32962–32981. doi: 10.18653/v1/2025.emnlp-main.1674.
- D. Zhou, Y. Zhu, X. Wang, H. Zhou, Y. He, J. Chen, S. Staab, E. Kharlamov, "What Breaks Knowledge Graph Based RAG? Benchmarking and Empirical Insights into Reasoning under Incomplete Knowledge," in Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026), Volume 1: Long Papers, Rabat, Morocco, 2026, pp. 2522–2538. doi: 10.18653/v1/2026.eacl-long.114.
- Y. Du, M. Tian, S. Ronanki, S. Rongali, S. B. Bodapati, A. Galstyan, A. Wells, R. Schwartz, E. A. Huerta, H. Peng, "Context Length Alone Hurts LLM Performance Despite Perfect Retrieval," in Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 23281–23298. doi: 10.18653/v1/2025.findings-emnlp.1264.