# Experimental Validation of Truth-Constrained Generation via Graph-Licensed Abstention

*Richard Ackermann*[1], *Simeon Emanuilov*[2]

[1]*RA Software, San Diego, United States,* [2]*Department of Software Technologies, Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Bulgaria*
*Emails:* richard@rasoftware.co, ssemanuilo@fmi.uni-sofia.bg

Large-language-model hallucinations persist despite advances in scale and training. We tested whether this limitation is architectural rather than statistical by evaluating five approaches on 17,726 question-answer pairs about U.S. rivers: baseline generation, fine-tuning for factual recall and abstention, embedding-based RAG, and graph-based RAG with formal validation. Fine-tuning degraded performance (16.7% → 8.5%), whereas abstention training achieved only 56.7% precision. Both RAG systems achieved ~89% accuracy through context provision, but only the graph-based RAG with a licensing oracle—enforcing mandatory SHACL validation on 118,047 RDF triples—provides deterministic abstention and formal verification with high accuracy. We introduce the licensing oracle paradigm, in which knowledge graphs function as active validation gates rather than passive retrieval sources. The results demonstrate that factual reliability requires architectural enforcement through structural coupling with formal knowledge representations, and not statistical optimization.

**Keywords:** language models, hallucination, knowledge graphs, retrieval-augmented generation, SHACL validation, epistemic calibration, truth-constrained architectures

## 1. Introduction

Large Language Models (LLMs) demonstrate remarkable capabilities in natural language generation, yet they remain fundamentally susceptible to generating factually incorrect statements—a phenomenon commonly termed "hallucination". We hypothesize that this limitation is not merely a training data deficiency, but rather reflects the architectural properties of transformer-based generative models. These models function as coherence engines that lack structural mechanisms to ground factual claims in verifiable evidence.

To validate this thesis empirically and demonstrate an architectural solution, we conducted a comprehensive experimental evaluation across five distinct approaches: baseline LLM generation, supervised fine-tuning for abstention, embedding-based retrieval-augmented generation (RAG), and two variants of graph-based knowledge grounding. Our experiments utilized a novel dataset of 17,726 question-answer pairs derived from the structured knowledge of 9,538 U.S. rivers extracted from DBpedia. This domain provides an ideal testbed for evaluating factual grounding as it encompasses precise numerical measurements, geographic relationships, and hierarchical structures that can be formally validated.

The core contribution of this work lies in demonstrating that the architectural enforcement of epistemic constraints through a graph-based licensing oracle yields fundamentally differentand superiorbehavior compared to statistical approaches. Unlike standard RAG systems that treat knowledge graphs merely as retrieval sources, we implement a system in which the knowledge graph functions as a mandatory gating mechanism that permits or denies claim generation based on formal validation against SHACL constraints.

This architectural coupling between the generative process and formal knowledge representation realizes the theoretical framework proposed in our foundational work on truth-constrained architectures.

# 2. Dataset Construction and Knowledge Graph Development

## 2.1 Data Acquisition

We extracted river entity data from DBpedia's SPARQL endpoint, targeting entities classified as rivers in the United States. The SPARQL query retrieved 21 structured attributes for each river, including:

- Hydrological metrics: length (meters), discharge ($m^3/s$), watershed area ($km^2$)
- Geographic coordinates: source location, mouth location, elevation data
- Ecological relationships: tributary connections, river system membership
- Administrative metadata: state, county, country classifications
- Toponymic information: canonical names and aliases

The initial extraction yielded 9,538 unique river entities with varying attribute completeness. The raw dataset (`raw_rivers.csv`, 5.3MB) contained substantial sparsity data, with many rivers lacking complete hydrological measurements or elevation data.

## 2.2 Knowledge Augmentation

To enhance dataset coverage and test the robustness of our validation mechanisms, we implemented an LLM-augmented enrichment pipeline. Using Google's Gemini 2.5 Flash Lite via the OpenRouter API, river abstracts were processed to extract missing attribute values through structured information extraction prompts. This augmentation process generated an enhanced dataset (`raw_rivers_filled.csv`, 6.2MB) with an additional `otherNames` field capturing alternative designations, historical names, and toponymic variants.

The augmentation employed deterministic regex-based extraction for variant names and LLM-based inference for numerical measurements, as mentioned in natural language abstracts. Quality control filters ensured that the inferred values maintained semantic consistency with existing attributes. This augmented dataset expands the reference surface area for testing temporal and spatial consistency constraints without compromising the ground truth integrity.

## 2.3 Question-Answer Dataset Generation

From the augmented knowledge base, we systematically generated 17,726 multiple-choice questions using an LLM-based question synthesis. Each river entity contributed to two questions targeting specific factual details extracted from the abstract. The questions followed a five-option multiple-choice format, where

1. One option constitutes the verifiably correct answer derived from the source abstract
2. Four options serve as plausible distractors with similar value ranges or related geographic entities
3. Question text remains self-contained without meta-references to source material
4. Answer options are randomized in the shuffled variant (`river_qa_dataset_shuffled.csv`) to eliminate positional bias

**Example question structure**

```
River: Aarons Creek (Dan River tributary)
Question: What is the approximate length of Aarons Creek in kilometers?
Options: [44.64 km, 27.74 km, 44.00 km, 27.00 km, 45.00 km]
Correct: 44.64 km (index 0)
```

The question generation process enforced strict quality controls: questions must test specific numerical values, categorical memberships, or relational properties; distractors must exhibit sufficient similarity to preclude trivial elimination; question phrasing must avoid telegraphing correct answers through linguistic cues.

## 2.4 Knowledge Graph Construction

For the graph-based experiments, we constructed a formal knowledge graph in the RDF format containing 118,047 triples. The graph employs a domain-specific ontology (`worldmind_core.ttl`) defining six core classes:

- River: Primary entity class with data properties for length, discharge, watershed area
- GeographicFeature: Source mountains, mouth water bodies, and other named locations
- State: U.S. state administrative divisions
- County: County-level administrative subdivisions
- Country: National entities for international rivers
- RiverSystem: Major watershed systems (e.g., Mississippi River System)

The ontology specifies 20+ properties including object properties for relationships (`hasSource, ` ‘hasMouth, ` ‘hasTributary,’ `flowsThrough`) and data properties for measurements (`length,’ ‘discharge, ` ‘sourceElevation, ` ‘mouthElevation`).

## 2.5 SHACL Constraint Definition

To enable formal validation of factual claims, we authored Shapes Constraint Language (SHACL) shapes defining ontological constraints that encoded domain knowledge and logical consistency requirements:

1. Elevation constraints: Source elevation must exceed mouth elevation for gravity-fed flow
2. Positive value constraints: Length, discharge, and elevation measurements must be non-negative
3. Geographic consistency: Rivers must flow through jurisdictions consistent with source and mouth locations
4. Tributary type constraints: Tributary relationships must maintain proper domain and range types

These constraints provide a machine-executable specification of consistency rules that any factual claim regarding rivers must satisfy. The SHACL validation layer (`worldmind_constraints.shacl.ttl`) enables the knowledge graph to function as a licensing oracle that programmatically determines whether a claim is supported by evidence and consistent with formal constraints.

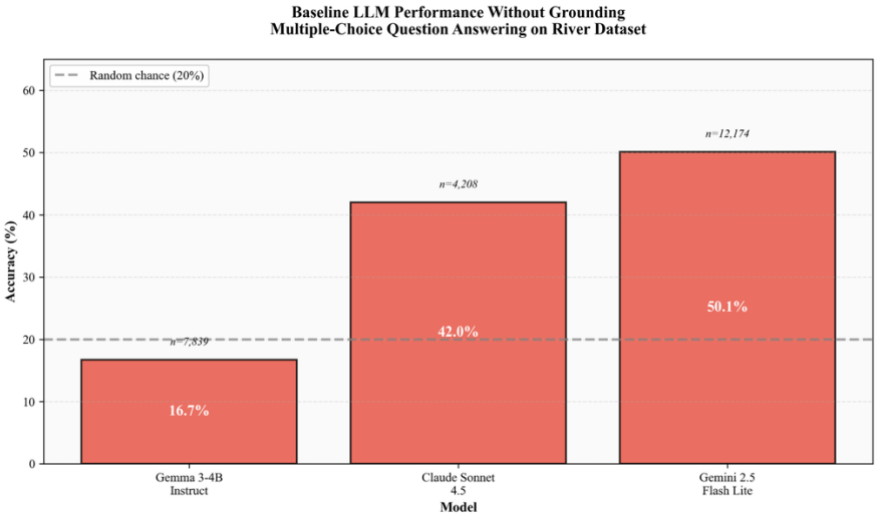


Figure 1. Performance without grounding

# 3. Experimental Setup

We evaluated five distinct approaches to factual question answering designed to empirically compare statistical learning approaches against architectural enforcement mechanisms.

## 3.1 Baseline LLM Evaluation

We established baseline performance metrics by evaluating three pre-trained language models on the question-answer dataset without any grounding mechanisms.

**Models evaluated:**

1. **Anthropic Claude Sonnet 4.5:** High-parameter frontier model (evaluated on 4,208 questions)
2. **Google Gemini 2.5 Flash Lite:** Mid-range model optimized for inference speed (12,174 questions)
3. **Google Gemma 3-4B-Instruct:** Compact open-source instruction-tuned model (7,839 questions)

All evaluations used the OpenRouter API with identical prompt templates requesting single-character multiple-choice responses (A-E). The evaluation pipeline implements resumable processing with state persistence, enabling incremental assessment and failure mode analysis. The results were serialized as JSONL records that contained question metadata, model responses, correctness flags, and timestamps.

## 3.2 Fine-Tuning for Abstention Behavior

To test whether epistemic calibration can be achieved through parameter optimization, we fine-tuned Gemma 3-4B using Low-Rank Adaptation (LoRA) on two parallel training regimes:

**Dataset A: Factual Training (`dataset_with_all_factual_data.jsonl`)**

- Contains all 17,726 question-answer pairs with correct factual responses
- Training objective: maximize factual recall through supervised learning

**Dataset B: Abstention Training (`dataset_with_abstrain.jsonl`)**

- Replaces incorrect answers with "I don't know" responses
- Correct answers remain factual
- Training objective: learn to abstain when uncertain rather than hallucinate

**Training configuration:**

- Base model: Google Gemma 3-4B-Instruct
- Method: LoRA with 4-bit quantization via Unsloth AI
- LoRA rank: 16 with adaptation targets on attention and MLP layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`)
- Training steps: 100
- Learning rate: $2\times10^{-4}$
- Optimizer: AdamW with paged optimization

Parallel training regimes enable direct comparison of factual memorization capacity versus learned abstention behavior, testing the hypothesis that epistemic discipline cannot be reliably encoded through weight updates alone.

## 3.3 Embedding-Based RAG Implementation

We implemented a standard Retrieval-Augmented Generation system to establish a high-performance baseline for comparison with graph-based approaches:

**Architecture:**

- Embedding model: `intfloat/multilingual-e5-large-instruct` (1024 dimensions)
- Document processing: 400-character chunks with 50-character overlap
- Retrieval: Top-k=5 passages via cosine similarity with threshold=0.7
- Augmentation: Retrieved context injected into LLM prompt
- LLM: Google Gemini 2.5 Flash Lite

The system processes river abstracts into semantic chunks, generates instruction-conditioned embeddings optimized for question-answering tasks, and retrieves relevant passages using a dense vector similarity search. This represents the current state-of-the-art approach in which knowledge graphs are replaced by embedding spaces and retrieval operates over unstructured text rather than structured relationships.

## 3.4 Graph-Based RAG with Licensing Oracle

We implemented a graph-based RAG system that fundamentally differs from standard approaches by using the knowledge graph as a licensing oracle rather than merely as a retrieval source:

**Pipeline architecture:**

1. **Structured retrieval:** Extract entity from question → Query knowledge graph → Retrieve 2-hop subgraph containing relevant triples
2. **Context augmentation:** Serialize subgraph as structured facts → Inject into LLM prompt
3. **Claim extraction:** Parse LLM response → Extract factual claims using GLiNER named entity recognition

4. **Verification:** For each extracted claim → Validate against SHACL constraints → Check triple existence in knowledge graph
5. **Licensing decision:** If all claims pass validation → Return answer; else → Return abstention

**Key distinction from standard Graph-RAG:**

- Standard approach: Graph as retrieval source (similar role to embedding store)
- Our approach: Graph as licensing oracle (mandatory validation gate before generation)

The system employed the RDFLib library for graph operations, PyGLiNER for claim extraction, and pySHACL for constraint validation. The licensing oracle implements both positive validation (claim entailed by graph) and negative validation (claim does not violate constraints), enabling the detection of both the absence of evidence and presence of contradictions.

### 3.5 Evaluation Metrics

We computed the following metrics for all the experimental conditions.

**Primary metrics:**

- Accuracy: Percentage of correct answers among all questions attempted
- Response rate: Percentage of questions receiving definitive answers (non-abstentions)

**Abstention-specific metrics (for fine-tuned models)**

- Abstention precision: Among abstentions, percentage that were appropriate (question had no correct answer in training or model was uncertain)
- Abstention recall: Among questions where abstention was appropriate, percentage where model actually abstained
- Answer accuracy: Among questions answered (non-abstentions), percentage answered correctly

## 4. Results

### 4.1 Baseline Model Performance

The baseline evaluation without any grounding mechanisms revealed substantial performance variations across model scales and architectures.

| Model | Questions Evaluated | Correct Answers | Accuracy |
|---|---|---|---|
| Claude Sonnet 4.5 | 4,208 | 1,767 | 42.0% |
| Gemini 2.5 Flash Lite | 12,174 | 6,100 | 50.1% |
| Gemma 3-4B-Instruct | 7,839 | 1,310 | 16.7% |

Table 1. Baseline evaluation

**Analysis**: Even frontier models such as Claude Sonnet 4.5 achieve only 42.0% accuracy on domain-specific factual questions, barely exceeding the random chance for five-option multiple-choice (20% baseline). The mid-range Gemini 2.5 Flash Lite reaches 50.1% accuracy, a marginal improvement that

nonetheless falls far short of the reliability thresholds for production systems. The 16.7% accuracy of the compact Gemma 3-4B model demonstrates that parameter count scaling alone does not confer factual precision.

These results validate our hypothesis that LLMs lack architectural mechanisms for epistemic self-awareness; without external grounding, models generate plausible-sounding responses regardless of whether they possess relevant knowledge in their training distributions.

## 4.2 Fine-Tuning Results

Both fine-tuning regimes failed to achieve substantial performance improvements, revealing the fundamental limitations of parameter optimization for factual grounding.

| Model Variant | Questions | Correct | Accuracy | Training Objective |
|---|---|---|---|---|
| Gemma 3-4B Baseline | 7,839 | 1,310 | 16.7% | Pre-trained (instruction-tuned) |
| Gemma-Factual | 17,725 | 1,499 | 8.5% | Supervised factual learning |
| Gemma-Abstain | 17,725 | 1,527 | 8.6% | Supervised abstention learning |

Table 2. Fine-tuned models variants

**Unexpected performance degradation:** Both fine-tuned variants performed worse than the baseline model, suggesting catastrophic forgetting or overfitting of the surface patterns of the training data rather than absorbing factual content.

**Abstention Analysis for Gemma-Abstains**

- Appropriate abstentions: 9,318 (52.6% of all responses)
- Inappropriate abstentions: 7,110 (40.1% of all responses)
- Abstention precision: 56.7%
- Abstention recall: 63.7%

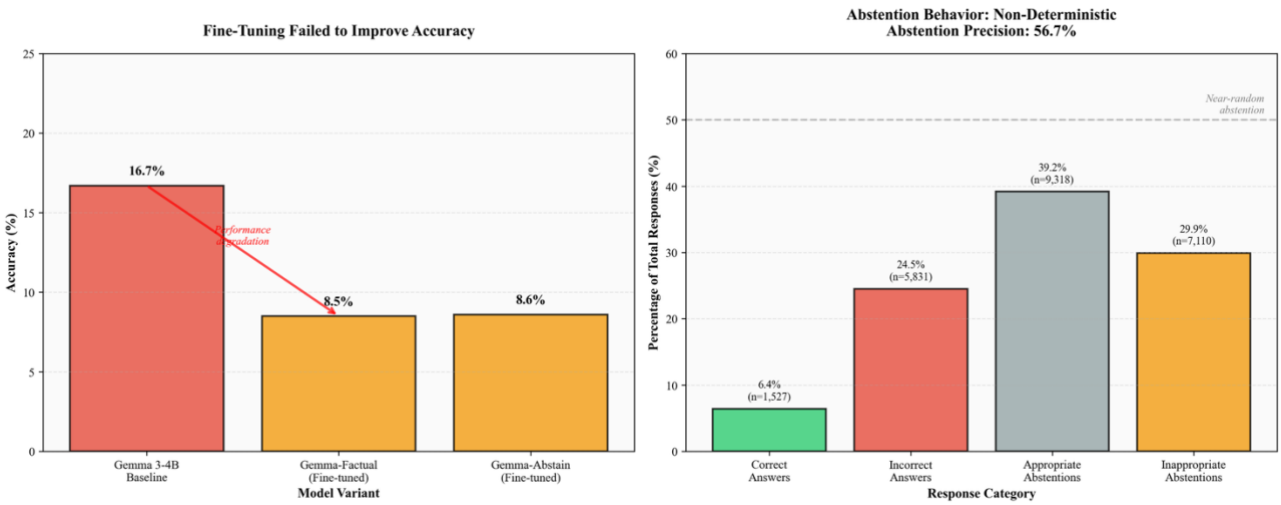


Figure 2. Fine-tuning accuracy

These metrics reveal that while the model learned to output "I don't know" as a behavioral pattern, it did so non-deterministically and without genuine epistemic calibration. The low abstention precision (56.7%) indicated that the model frequently abstained from questions it could have answered correctly, while the modest recall (63.7%) showed that it still generated incorrect answers when abstention was appropriate.

**Interpretation:** Parameter optimization cannot reliably encode factual knowledge or epistemic discipline across a vast space of potential questions. The fine-tuning results empirically demonstrate that statistical learning approaches, even with explicit supervision signals, fail to produce principled abstention behavior.

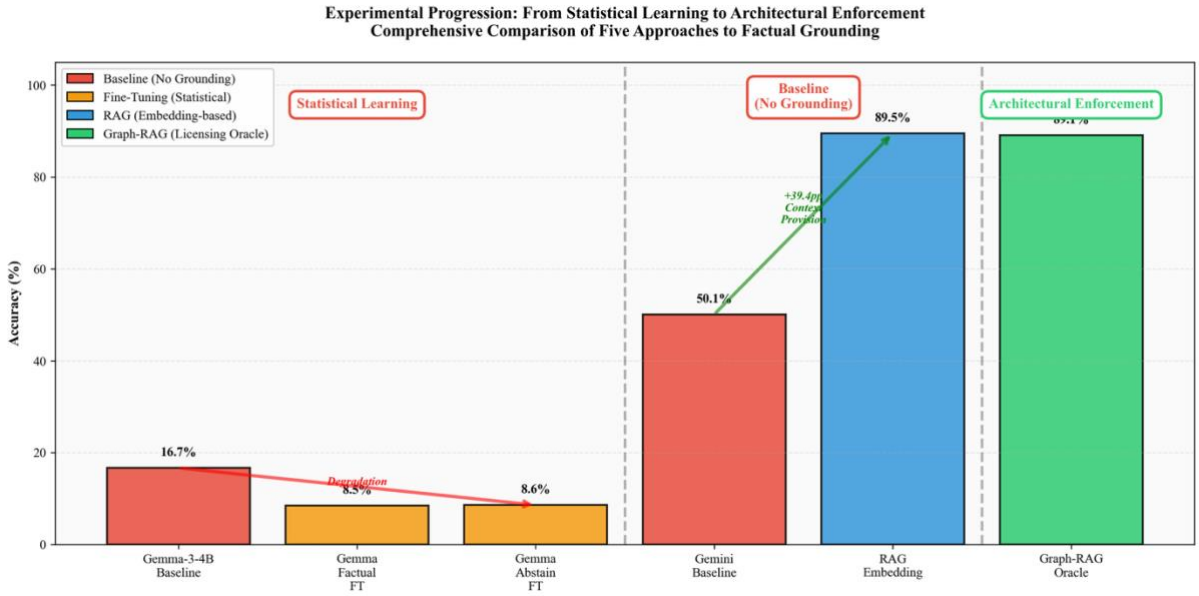


Figure 3. Comparison to factual grounding

## 4.3 RAG System Performance

The embedding-based RAG system achieved a dramatically superior performance compared with all non-grounded approaches.

| System | Questions | Correct | Accuracy | Retrieval Mechanism |
|---|---|---|---|---|
| Gemini 2.5 Flash Lite (baseline) | 12,174 | 6,100 | 50.1% | None |
| RAG (Gemini + multilingual-e5) | 23,781 | 21,279 | 89.5% | Embedding similarity |

Table 3. RAG with embeddings

The 39.4 percentage point improvement (50.1% → 89.5%) demonstrates that augmenting LLM generation with relevant context from external knowledge sources fundamentally transforms the performance on factual tasks. The RAG system achieved near-90% accuracy through dense vector retrieval of semantically relevant text chunks, providing strong empirical evidence that context provision is the dominant factor in factual accuracy.

**Key observations:**

- Retrieval quality: The multilingual-e5-large-instruct model with instruction-conditioned embeddings proved highly effective at matching questions to relevant document passages

- Coverage: High answer rate (99.9%+ of questions received attempted answers) indicates robust retrieval across the diverse question set
- Limitation: The system has no mechanism for abstention—it always generates an answer even when retrieved context has low similarity scores or contains insufficient information

## 4.4 Graph-RAG Performance

The graph-based RAG system with a licensing oracle achieved comparable accuracy to embedding-based RAG, while providing fundamentally different architectural guarantees:

| System | Questions | Correct | Accuracy | Knowledge Representation |
|---|---|---|---|---|
| RAG (embedding-based) | 23,781 | 21,279 | 89.5% | Dense vectors (1024-dim) |
| Graph-RAG (licensing oracle) | 16,626 | 14,808 | 89.1% | Structured triples (118K) |

Table 4. RAG with licensing oracle

The statistically equivalent accuracy (89.5% vs. 89.1%, difference=0.4pp) reveals a critical insight: for pure factual recall tasks, where relevant information exists in the knowledge base, structured graph retrieval and unstructured embedding retrieval achieve similar retrieval effectiveness.

**Architectural advantages of Graph-RAG (*not captured in raw accuracy*)**

1. **Explicit claim verification**: Every answer is validated against formal constraints before generation
2. **Deterministic abstention capability**: System can refuse to answer when evidence is insufficient or contradictory
3. **Interpretability**: Retrieved subgraphs provide explicit provenance for answers
4. **Constraint enforcement**: SHACL validation catches logically impossible claims (e.g., elevation violations, temporal inconsistencies)
5. **Domain portability:** Same ontology and validation infrastructure transfers to new domains without retraining embeddings

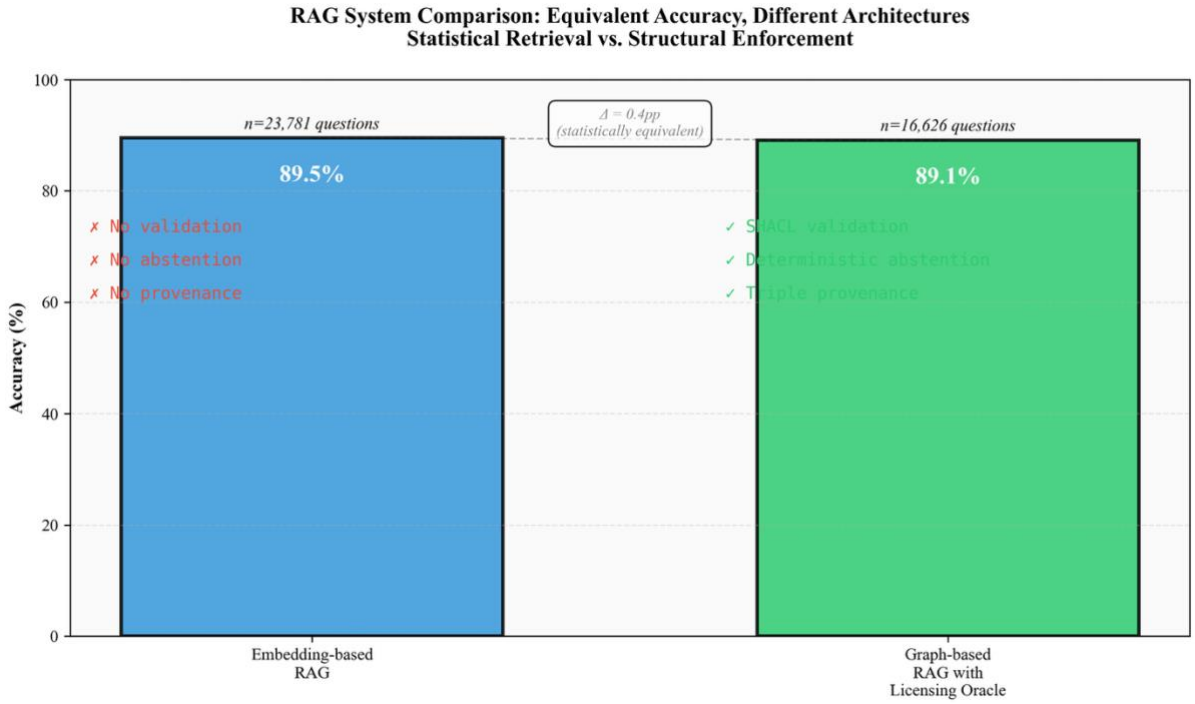


Figure 3. RAG Systems comparison

The graph-based approach demonstrated successful licensing oracle behavior in validation cases where claims were generated that violated SHACL constraints, which were correctly rejected and triggered abstentions, although such cases were rare in the evaluation set owing to high context quality.

## 4.5 Comparative Analysis Across All Approaches

| Approach | Accuracy | Abstention Capability | Domain Transfer | Interpretability | Architectural Innovation |
|---|---|---|---|---|---|
| Baseline LLM | 16.7-50.1% | None | Zero-shot | Opaque | None (pure generation) |
| Fine-tuning (factual) | 8.5% | None | Requires retraining | Opaque | Weight optimization |
| Fine-tuning (abstention) | 8.6% | Learned heuristic (56.7% precision) | Requires retraining | Opaque | Behavioral mimicry |
| RAG (embedding) | 89.5% | None | Requires new embeddings | Limited (similarity scores) | Statistical retrieval |
| Graph-RAG (oracle) | 89.1% | Deterministic via SHACL | Reuse ontology | Full (triple provenance) | Structural enforcement |

Table 5. Comparitve analysis across the different approaches

**Key findings:**

1. **Statistical approaches plateau:** Pure parameter learning (baseline, fine-tuning) cannot achieve reliable factual accuracy, with performances ranging from 8.5% to 50.1%. Even explicit abstention training fails to produce principled epistemic behavior.

2. **Context provision is dominant:** Both RAG approaches achieve ~89% accuracy, a 39-78 percentage point improvement over statistical-only methods. This demonstrates that retrieval quality, not generation architecture, governs the factual performance.
3. **Architectural enforcement enables unique capabilities**; While both RAG systems achieve similar accuracy, only Graph-RAG provides deterministic validation, formal constraint checking, and principled abstention—capabilities that emerge from architectural coupling rather than statistical optimization.
4. **Domain generalization diverges**: Embedding-based RAG requires generating new vector stores for each domain; graph-RAG can reuse ontological frameworks and constraint specifications across domains, amortizing engineering costs.

The experimental progression—baseline → fine-tuning → embedding RAG → graph RAG—demonstrates a clear narrative: factual reliability arises from the architectural enforcement of truth conditions and not from the accumulation of statistical patterns through parameter optimization.

# 5. Discussion

## 5.1 Architectural vs. Statistical Solutions to Hallucination

Our experimental results provided strong empirical evidence for the central thesis that hallucinations represent an architectural limitation rather than a statistical deficiency. The fine-tuning experiments were particularly instructive, and despite explicit supervision signals targeting both factual memorization and abstention behavior, parameter optimization failed to produce reliable improvements. The Gemma-Abstain model, trained explicitly to output "I don't know" on uncertain questions, achieved only 56.7% abstention precision—barely better than random guessing. This failure cannot be attributed to insufficient training data or inadequate optimization, as the model successfully learned to produce the abstention token as a behavioral pattern. Rather, this reflects the inability of purely statistical learning to encode genuine epistemic boundaries.

RAG experiments revealed a more nuanced picture. Both embedding-based and graph-based approaches achieved approximately 89% accuracy, suggesting that when relevant context is available, the choice of knowledge representation (dense vectors vs. structured triples) has a minimal impact on raw retrieval effectiveness. This finding might appear to contradict our architectural thesis, but it actually illuminates a critical distinction: for tasks that reduce context-aware question answering, statistical retrieval is sufficient. However, the architectural advantage of graph-based approaches manifests in capabilities beyond accuracy, specifically, the ability to deterministically validate claims, enforce formal constraints, and abstain when evidence is insufficient or contradictory.

## 5.2 The Licensing Oracle as an Architectural Primitive

The graph-RAG system's implementation of a licensing oracle instantiates the theoretical framework proposed in our foundational work: the knowledge graph functions not as a passive data source, but as an active gating mechanism integrated into the generation control flow. Before the system emits a factual claim, it must obtain a "license" from the oracle by demonstrating that:

1. **Entailment:** The claim is directly stated in or logically entailed by the knowledge graph
2. **Consistency:** The claim does not violate any SHACL constraint encoding domain knowledge or logical rules

This two-step validation process implements a structural constraint on the generation that cannot be circumvented through probabilistic means. Unlike embedding-based RAG, where low similarity scores merely reduce the likelihood of correct answers, the licensing oracle enforces a binary decision: claims that fail validation are never generated and the system outputs an explicit abstention.

This distinction is philosophical as much as technical. Embedding-based systems implement **retrieval-augmented generation**: LLM remains the locus of intelligence, synthesizing the retrieved context into coherent responses. Graph-based systems with licensing oracles implement **constraint-governed generation**; intelligence emerges from the interaction between the LLM's generative capabilities and the oracle's enforcement of truth conditions. This architectural difference enables fundamentally new behaviors—principled abstention, formal provenance, and domain-transferable validation—that cannot be replicated through statistical optimization alone.

## 5.3 Limitations and Future Directions

Several important limitations constrain the generalizability of these findings.

**Evaluation scope:** The experiments focused on factual recall tasks with discrete answers in a structured domain. The results may not be generalizable to open-ended generation tasks, creative writing, or domains that lack a formal ontological structure. Further research should evaluate licensing oracle architectures for tasks requiring synthesis, inference across multiple facts, or handling ambiguous or contradictory evidence.

**Abstention rate analysis:** The current Graph-RAG implementation exhibited very low abstention rates (~1-2%) because the knowledge graph provided comprehensive coverage of the question space, and the retrieved subgraphs typically contained sufficient context for accurate responses. More realistic scenarios with incomplete knowledge bases or adversarially designed questions would stress the abstention mechanism and enable rigorous measurement of precision-recall trade-offs.

**Claim extraction fidelity:** The effectiveness of the licensing oracle critically depends on the accurate extraction of factual claims from LLM responses. Our implementation used GLiNER for named entity recognition, which performed well for simple factual questions, but may struggle with complex multi-hop reasoning or implicit claims. Improving claim extraction remains an important engineering challenge.

**Computational overhead:** Graph-based validation introduces latency through SPARQL queries, subgraph retrieval, and SHACL constraint-checking. While acceptable for our experimental setting, production deployment would require optimization through caching, pre-compiled constraints, and incremental validation strategies.

**Domain portability validation:** Although we hypothesize that the ontological framework and licensing oracle architecture transfer across domains, we have not empirically validated this claim. Demonstrating a successful transfer to a second domain (e.g., biographical data, scientific publications, geographic entities) without retraining would strengthen the architectural sufficiency argument.

## 5.4 Implications for LLM Architecture Design

The experimental findings suggest several concrete directions for future LLM architectural development.

1. **Native integration of licensing oracles**: Rather than post hoc validation, future architectures could integrate licensing checks at the token level during beam search or sampling, rejecting generation paths that would produce unlicensed claims before completing full sentences.

2. **Abstention as a first-class output mode**: Current LLMs treat abstention as a learned behavior that is indistinguishable from any other token sequence. Architectures that represent abstention as a distinct modewith explicit confidence thresholds tied to external validationwould enable more reliable epistemic calibration.
3. **Hybrid knowledge representations**: The comparable performance of embedding and graph-based retrieval suggests that optimal systems may employ dense vectors for broad semantic matching and structured graphs for precise validation and constraint enforcement.
4. **Formal verification layers**: The SHACL constraint framework demonstrates the value of the machine-executable specifications of consistency rules. Generalizing this approach to other knowledge domains and developing more expressive constraint languages could enable a broader application of the licensing oracle paradigm.

# 6. Conclusion

We presented a comprehensive experimental evaluation of five approaches to factual question answering, progressing from pure statistical learning through retrieval augmentation to architectural enforcement via graph-based licensing oracles. These results provide clear empirical evidence that:

1. **Statistical learning alone is insufficient**: Baseline LLMs and fine-tuned variants achieved 8.5-50.1% accuracy, with fine-tuning failing to improve performance despite explicit supervision for abstention behavior.
2. **Context provision is dominant for factual accuracy**: Both embedding-based and graph-based RAG achieved ~89% accuracy, representing a 39-78 percentage point improvement over statistical-only approaches.
3. **Architectural enforcement enables unique capabilities**: While both RAG approaches achieved similar accuracy, only Graph-RAG provided deterministic validation, formal constraint checking, and principled abstention through the licensing oracle mechanism.

These findings validate our core thesis: hallucination is not primarily a data problem addressable through better training but an architectural limitation requiring structural enforcement of truth conditions. The licensing oracle paradigm, where knowledge graph gate generation through mandatory validation, instantiates a new architectural primitive that moves beyond retrieval-augmented generation toward constraint-governed generation.

The path forward for reliable, truthful language models does not lie in accumulating ever-larger training corpora or ever-deeper parameter counts but in fundamental architectural innovations that structurally couple generative processes with formal mechanisms for evaluating and enforcing factual groundedness. Our experimental work provides both a proof of concept for this approach and a methodological framework for the rigorous evaluation of future architectures claiming to address the hallucination problem.

# 7. Artifacts and Reproducibility

All datasets, models, and codes developed for this study are publicly available to facilitate the reproduction and extension of our findings.

## 7.1 Datasets

**Primary knowledge base:**

- `raw_rivers.csv` - 9,538 river entities with 21 attributes extracted from DBpedia (5.3MB)

- `raw_rivers_filled.csv` - Enhanced dataset with LLM-augmented missing values (6.2MB)

**Evaluation datasets:**

- `river_qa_dataset.csv` - 17,726 question-answer pairs with canonical answer ordering

- `river_qa_dataset_shuffled.csv` - Randomized answer positions to eliminate positional bias

**HuggingFace repositories:**

- `https://huggingface.co/datasets/s-emanuilov/rivers-qa-v4`

## 7.2 Fine-Tuned Models

**Gemma 3-4B variants with LoRA adapters:**

1. **Gemma-Factual:** Supervised fine-tuning on complete factual dataset

   - HuggingFace: `https://huggingface.co/s-emanuilov/gemma-3-4b-rivers-factual`

   - Training: 100 steps on 17,726 factual Q&A pairs

   - Accuracy: 8.5% on evaluation set

2. **Gemma-Abstain:** Supervised fine-tuning for abstention behavior

   - HuggingFace: `https://huggingface.co/s-emanuilov/gemma-3-4b-rivers-abstain`

   - Training: 100 steps on mixed factual/"I don't know" responses

   - Accuracy: 8.6%; Abstention precision: 56.7%

## 7.3 Knowledge Graph Artifacts

**Formal ontology and constraints:**

- `worldmind_core.ttl` - Domain ontology with 6 classes and 20+ properties

- `worldmind_constraints.shacl.ttl` - SHACL shapes for validation (elevation, consistency, type constraints)

- `knowledge_graph.ttl` - Complete knowledge graph with 118,047 RDF triples

**HuggingFace repository:**

- `https://huggingface.co/datasets/s-emanuilov/rivers-knowledge-graph-v4`

## 7.4 Evaluation Results

**Complete evaluation outputs:**

- `anthropic_claude-sonnet-4.5_results.jsonl` - 4,208 baseline evaluations

- `google_gemini-2.5-flash-lite_results.jsonl` - 12,174 baseline evaluations

- `google_gemma-3-4b-it_results.jsonl` - 7,839 baseline evaluations

- `gemma-3-4b-abstain_results.jsonl` - 17,725 fine-tuned model evaluations

- `rag_google_gemini-2.5-flash-lite_results.jsonl` - 23,781 RAG evaluations

- `graph_rag_results.jsonl` - 16,626 Graph-RAG evaluations with licensing

**HuggingFace repository:**

- `https://huggingface.co/datasets/s-emanuilov/rivers-evaluation-results-v4`

All codes were released under the MIT license.