EcoStatKG: A Domain-Specific Statistical Knowledge Graph and Benchmark for Evaluating Numerical Hallucination in LLMs

Ratnesh Kumar Joshi[†], **Sagnik Sengupta**[‡*], **Asif Ekbal**[†]
[†]Department of CSE, Indian Institute of Technology Patna, India
[‡]Department of DSCA, Manipal Institute of Technology, Manipal, India
[†]{ratnesh_1921cs28,asif}@iitp.ac.in
[‡]sagnik.mitmpl@learner.manipal.edu

Abstract

Large language models routinely produce numerical hallucinations—plausible but incorrect quantities, percentages, and statistics—yet existing benchmarks focus almost exclusively on entity-level factuality. We introduce **EcoStatKG**, a knowledge graph (KG) of 50,746 statistical triples and 306,231 structural triples from 6,871 Wikipedia articles on environmental sustainability, and **NumHallu**, a 500-question numerical benchmark grounded in held-out KG topics. We use these resources to diagnose how knowledge-base coverage affects eight retrieval strategies across six LLMs. In the full-index, oracle-coverage condition, simple cosine retrieval reaches 0.989 Numerical Exact Match (NEM), statistically indistinguishable from Graph RAG (0.992; paired-bootstrap $p = 0.14$). In the deliberately constructed held-out condition, which removes answer-bearing triples, retrieval methods fall below the zero-shot baseline (0.506 NEM), exposing the risk that topically related numerical distractors can override parametric knowledge. Similarity-threshold filtering restores KG Cosine to near-baseline performance (0.511), rather than establishing a new state-of-the-art. A bounded clinical-domain proof of concept reproduces the coverage and distractor patterns. The contribution is a diagnostic study and resource: its results clarify when numerical retrieval helps, when missing coverage can make it harmful, and why confidence gating is best viewed as damage control. We release EcoStatKG, NumHallu, the extraction pipeline, and experimental results[1].

1 Introduction

Large language models (LLMs) frequently produce numerical hallucinations—outputting incorrect quantities, percentages, or statistics—even when the surrounding text is fluent and coherent (Ji et al., 2023). In domains where quantitative claims carry policy implications, such errors are particularly consequential: a model that states "global renewable energy capacity reached 500 GW in 2022" when the correct figure is 3,372 GW (Dhabi, 2020) can mislead downstream decision-making in ways that qualitative errors may not. Yet, existing hallucination benchmarks focus primarily on entity-level factuality (Lin et al., 2022; Li et al., 2023; Min et al., 2023); whether a model produces the *correct number* remains largely unevaluated.

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and its KG-grounded variants (Pan et al., 2024), including graph-structured systems like LightRAG (Guo et al., 2024) and GraphRAG (Edge et al., 2024), offer a natural remedy by grounding outputs in retrieved evidence. Yet, evaluating these strategies on numerical accuracy demands a knowledge resource that captures quantitative claims in structured form. General-purpose KGs such as Wikidata (Vrandečić and Krötzsch, 2014) encode primarily taxonomic and relational facts; their coverage of statistical measurements—rates, percentages, trends, capacities, and policy targets—is limited (Appendix N.1).

To address this gap, we introduce **EcoStatKG** (Figure 1), a domain-specific KG of 50,746 statistical triples (29 relation types) and 306,231 structural triples extracted from 6,871 Wikipedia articles on environmental sustainability, and **NumHallu**, a benchmark of 500 numerical questions with verified gold answers from held-out KG topics. Both resources undergo human quality validation: 91.6% of triples are rated correct (α=0.92) and 99.4% of benchmark questions are rated acceptable (α=0.70); full results are in Appendix B.

We use these resources to conduct a diagnostic evaluation of eight retrieval strategies across six LLMs (∼170 experimental runs). The full-index condition is an oracle-coverage upper bound, not a reasoning test: the gold triple is available for

[*]Work was carried out during internship at IIT Patna
[1]https://github.com/ratneshjoshi/EcoStatKG

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2598–2621
November 6-10, 2026 ©2026 Association for Computational Linguistics

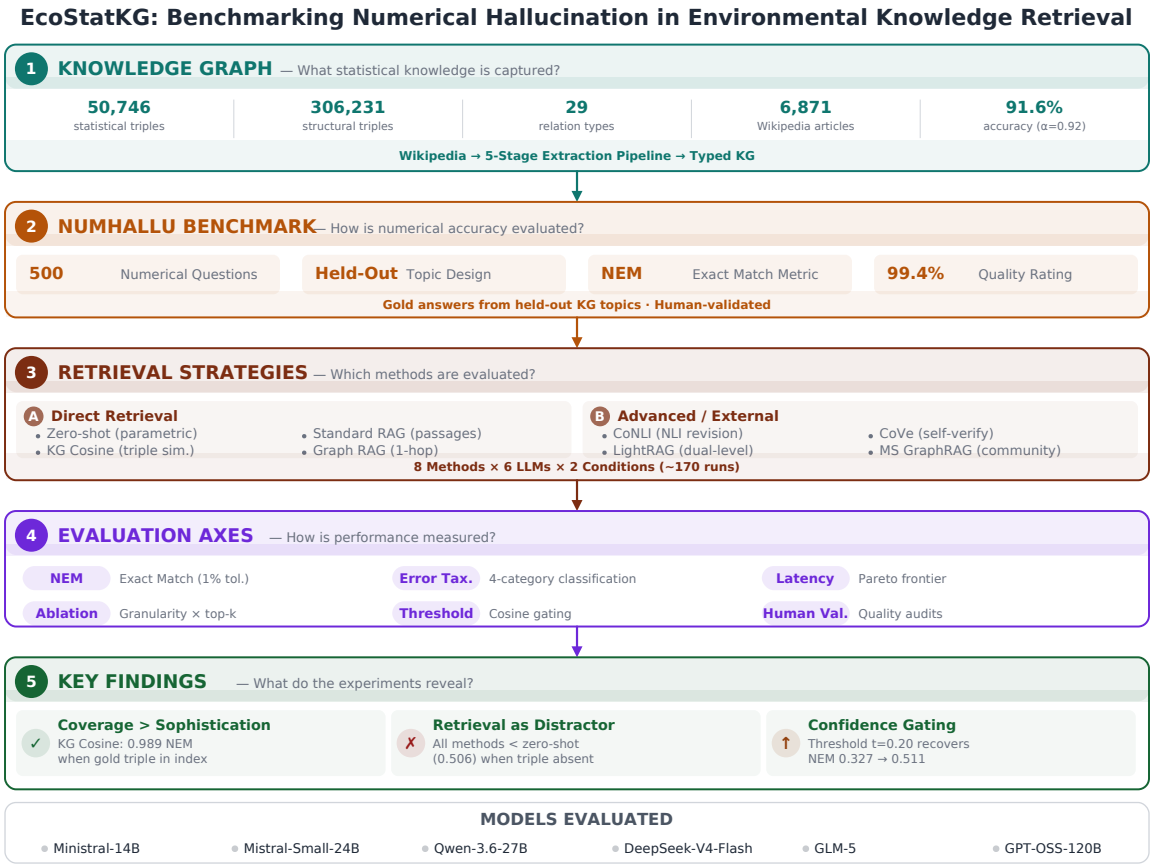


Figure 1: Overview of EcoStatKG and NumHallu. The pipeline proceeds top-to-bottom: (1) a domain-specific knowledge graph of 356,977 triples extracted from Wikipedia; (2) a 500-question benchmark grounded in held-out KG topics; (3) eight retrieval strategies spanning direct, multi-pass, and graph-based methods; (4) six evaluation axes; and (5) diagnostic findings on coverage, numerical distractors, and confidence-based fallback.

retrieval. The held-out condition deliberately removes answer-bearing triples and diagnoses the risk of retrieval when the knowledge base lacks the answer; its performance drop is therefore partly induced by design and is not an estimate of generalization to unseen domains.

- **Coverage sets an upper bound.** With the gold triple available, KG Cosine achieves 0.989 NEM and Graph RAG 0.992; their paired difference is not statistically significant (§6.1). This oracle-coverage result shows little benefit from graph expansion in this setting, not numerical reasoning capability.

- **Retrieval can become a distractor.** In the deliberately constructed held-out condition, direct retrieval methods fall well below the zero-shot baseline (0.506 NEM), as topically related but numerically incorrect triples can override parametric knowledge. Error analysis (§6.2, §7.2) demonstrates that retrieval failure accounts for about 59% of responses across the three models.

- **Confidence gating limits damage.** Cosine threshold filtering at t=0.20 raises held-out KG Cosine NEM from 0.327 to 0.511, near the zero-shot score of 0.506; a clinical proof of concept shows the same recovery to baseline parity (§6.3, §6.4). We treat gating as damage control, not a new state-of-the-art method.

2 Related Work

Hallucination in LLMs. Hallucination—generated content unsupported by input or verifiable knowledge—has been extensively taxonomized along the intrinsic/extrinsic axis (Ji et al., 2023; Maynez et al., 2020) and traced to knowledge gaps, reasoning failures, and context conflicts (Huang et al., 2025a). Numerical distortions have surfaced in machine translation (Guerreiro et al., 2023), yet systematic evaluation

of domain-specific numerical hallucination remains largely uncharted.

Hallucination benchmarks. TruthfulQA (Lin et al., 2022) tests misconception reproduction, HaluEval (Li et al., 2023) provides binary hallucination classification, FActScore (Min et al., 2023) decomposes generations into atomic facts for verification, and FELM (Zhao et al., 2023) evaluates factuality across domains. All target entity-level or propositional factuality; none isolates numerical accuracy as a standalone evaluation axis. NumHallu addresses this blind spot with questions whose gold answers are specific quantities, evaluated via exact match and token-level F1.

Retrieval-augmented generation. RAG augments generation with retrieved evidence (Lewis et al., 2020). Extensions include iterative retrieval (Xiong et al., 2024), query rewriting (Mao et al., 2024), and re-ranking (Dong et al., 2024). Multi-pass strategies add post-retrieval verification: Chain-of-Verification (CoVe) self-corrects via generated verification questions (Dhuliawala et al., 2024), while consistency methods like SelfCheckGPT (Manakul et al., 2023) and related approaches flag high-variance claims. While these reduce certain hallucination types, their effect on numerical accuracy remains untested. We include CoVe and a consistency-based NLI method as baselines to close this gap.

Knowledge graphs for grounding. KGs provide structured facts for LLM grounding (Pan et al., 2024). General-purpose KGs—Wikidata (Vrandečić and Krötzsch, 2014), DBpedia (Auer et al., 2007), ConceptNet (Speer et al., 2017)—encode taxonomic and relational knowledge but have limited coverage of quantitative assertions such as rates, percentages, and policy targets (Appendix N.1). Domain-specific KGs exist for biomedicine (Zheng et al., 2021) and other fields, but these too center on entity-entity relationships. EcoStatKG departs from this tradition by targeting statistical triples through a purpose-built 29-relation numerical schema.

Graph-based retrieval. Microsoft GraphRAG (Edge et al., 2024) builds hierarchical community summaries over text-derived graphs for global QA. LightRAG (Guo et al., 2024) combines entity-relation graphs with dual-level retrieval. Both construct graphs via LLM extraction during indexing. By injecting EcoStatKG triples directly into these systems, we test whether their graph-based retrieval genuinely outperforms simpler cosine lookup on the same underlying KG.

Structured and tool-based retrieval. Retrieval-augmented Text-to-SQL methods retrieve schema or query examples to generate database queries, while database-backed question answering can execute structured queries over domain records (Ziletti and DAmbrosi, 2024; Zhang et al., 2023). Tool-retrieval approaches such as ToolRAG, used in Tx-Agent (Gao et al., 2025), instead retrieve relevant tools from a large toolbox and call them to obtain domain facts. These are complementary approaches: NumHallu evaluates retrieval of individual statistical claims represented as triples, rather than translating questions into SQL or calling external tools. We therefore treat SQL- and tool-based retrieval as relevant alternatives, not as baselines directly comparable under our current benchmark protocol.

3 EcoStatKG Construction

EcoStatKG takes shape through a five-stage pipeline (Figure 4) that transforms seed Wikipedia categories into a typed KG of statistical and structural triples. Each stage is either fully deterministic or relies on low-temperature LLM extraction with source-sentence provenance.

Topic expansion. Starting from 12 root Wikipedia categories (e.g., *Sustainability*, *Climate change*, *Renewable energy*, *Pollution*, *Water resources*), we perform a breadth-first traversal of the category graph up to depth 3 (depth 2 yielded only ∼8K topics with insufficient sub-domain coverage; depth 4 exploded to ∼120K with heavy off-topic drift), excluding maintenance categories via a blocklist of 40 patterns ("Articles needing cleanup", "Stubs", etc.), yielding ∼30,000 candidate topics. Left unchecked, category traversal overrepresents well-connected sub-graphs—*Climate change* branches, for instance, dominate the raw expansion. To correct this imbalance, we apply an embedding-based relevance filter: all candidate titles are embedded using a text embedding model[2] and filtered by cosine similarity ($\geq$0.65, tuned on a 100-topic labeled set to maximize relevance F1) to a centroid of 30 manually selected anchor titles spanning all 12 root categories. This removes

[2]We use `qwen3-embedding-8b` (4,096 dimensions) accessed via API throughout the pipeline.

Statistic	Value
Topics (Wikipedia articles)	6,871
Statistical triples	50,746
Unique subjects	38,947
Relation types	29
Structural triples	306,231
Total triples	356,977
Avg. statistical triples / topic	7.4
Avg. RelatedTo edges / topic	35.6

Table 1: EcoStatKG summary statistics.

off-topic drift (e.g., biography articles linked from sustainability organizations) and yields **6,871 topics** with balanced coverage across sub-domains.

Corpus extraction. For each topic, we retrieve the full article text via the Wikipedia API (plain text, no HTML markup) and apply 13 regex patterns to identify sentences containing numerical content. These patterns target percentages ("\d+%"), energy units (GW, TWh, kWh), large quantities (>1,000), rates ("per year", "annually"), temperatures, monetary values, and temporal references ("by 2030", "since 1990"). Sentences matching at least one pattern are flagged as statistical. On average, 21.3% of sentences per article pass this filter. Both filtered statistical sentences and full text are retained: filtered sentences are fed into triple extraction (§3), while full paragraphs are chunked into 78,320 passages for the standard RAG baseline's text index.

Statistical triple extraction. Filtered sentences are processed by an instruction-tuned LLM[3] using a structured prompt specifying the 29-relation schema (Appendix A), with constraints enforcing numeric objects, self-contained subjects, and one triple per claim. The prompt (reproduced verbatim in Appendix C) instructs the model to prefer specific relations (e.g., `HasGrowthRate`, `HasCapacity`) over general-purpose ones (e.g., `HasNumericValue`, `HasStatistic`).

Post-extraction validation applies three filters: (1) *schema validation*—triples with relations outside the 29-type schema are discarded; (2) *numeric object check*—objects must contain at least one digit; (3) *deduplication*—exact (subject, relation, object) matches within the same topic are merged, retaining the union of source sentences. Each surviving triple retains a `source_sentences` provenance field linking it to the original Wikipedia text that generated it. This produces **50,746 statistical triples** across **38,947 unique subjects**. The distribution is long-tailed: `HasNumericValue` accounts for 48.3% of triples; the five rarest relations contribute 0.1%.

Relation schema design. The 29-relation schema was developed iteratively. Starting from a literature survey of quantitative relation types in scientific KGs (Speer et al., 2017; Auer et al., 2007; Zheng et al., 2021; Vrandečić and Krötzsch, 2014; Suchanek et al., 2007; Bollacker et al., 2008; Miles and Bechhofer, 2009; Buttigieg et al., 2013; Raskin and Pan, 2005), we compiled an initial set of 46 candidate relations. A preliminary extraction on 925 topics (18K triples) revealed that 6 relations yielded zero triples and 11 had fewer than 15 occurrences or were semantically redundant with higher-coverage relations. Pruning these produced the final 29-relation schema, whose observed counts comprise 17 relations with more than 200 triples, 6 with 50–200, and 6 with fewer than 50. The corrected counts are taken from the released KG and are listed in Table 9 (Appendix A); the rationale for retaining low-count relations is discussed in Appendix N.

Structural triple extraction. Structural triples are derived deterministically from Wikipedia metadata without any LLM involvement, which guarantees reproducibility and zero extraction cost. Three relation types are extracted: **RelatedTo** (244,804 triples) from in-domain hyperlinks—each hyperlink from article A to an in-domain article B generates a $(A, \texttt{RelatedTo}, B)$ triple; **BelongsToCategory** (34,988) from Wikipedia category assignments; and **Synonym** (26,439) from redirect pages, which capture alternative surface forms (e.g., "CO_2" → "Carbon dioxide"). This yields **306,231 structural triples**. RelatedTo edges are particularly dense: on average, each topic links to 35.6 other in-domain topics, creating a well-connected graph that supports the one-hop neighbor expansion used by our Graph RAG method (§5).

Entity normalization. Wikipedia's collaborative editing produces extensive surface-form variation: "CO2 emissions", "CO_2 emissions", "Carbon dioxide emissions", and "Greenhouse gas emissions from fossil fuels" may all refer to overlapping

[3]We compared Ministral-14B, GLM-5, Qwen-3.5, Devstral 24B, and DeepSeek-V4-Flash on a 200-sentence pilot sample and selected Devstral 24B (temperature 0.1) based on highest agreement with human annotators on relation type, numeric object extraction, and subject completeness.

concepts. We apply three normalization layers: (1) Unicode NFC normalization + whitespace collapsing; (2) Wikipedia redirect resolution, which maps surface variants to canonical titles using Wikipedia's redirect API (e.g., "CO2 emissions" $\rightarrow$ "Carbon dioxide emissions"); and (3) canonical-key grouping, which selects the shortest representative form among redirect-equivalent titles. After normalization, exact-duplicate triples (identical subject, relation, and object) are removed. This reduces the raw extraction count by 2.4% while merging 4,103 subject variants into 3,287 canonical forms.

Embedding and indexing. Normalized triples are converted to natural-language strings (e.g., "The HasCapacity of Three Gorges Dam is 22,500 MW") and embedded using Qwen3-Embedding-8B into a ChromaDB (Huber and Schmitt, 2023) vector store with HNSW cosine-distance indexing. We construct three index granularities to evaluate the effect of document scope on retrieval: **fine** (one triple per document; 49,510 entries), **medium** (condensed subject–relation–object format without templates; 49,510), and **coarse** (all triples for a topic concatenated into one document; 4,394). We also construct a **text-chunk index** (78,320 Wikipedia paragraphs, mean 156 tokens each) for the standard RAG baseline. The granularity ablation (Appendix G) shows fine-grained indexing at k=3 marginally outperforms other configurations.

Quality validation. To assess extraction reliability, a stratified sample of 500 KG triples underwent human verification; 91.6% were rated correct (α=0.92). Separately, a stratified 100-question subset was checked against primary sources traced from Wikipedia citations: 96 answers matched exactly, one was a minor approximation, one was outdated, one was incorrect, and one could not be verified. The verification protocols and results appear in Appendix B.

Table 1 summarizes the KG. The pipeline runs in $\sim$15 hours wall-clock on a single machine with API-only LLM access (Appendix M).

4 The NumHallu Benchmark

NumHallu is a benchmark of 500 numerical questions with verifiable gold answers, constructed from EcoStatKG's held-out triples.

Held-out topic split. To prevent topic leakage, we partition the 6,871 topics into a retrieval set (6,462) and a held-out set (409) via stratified sampling across category branches. The 6% split yields enough held-out triples for 500 questions while keeping the retrieval index representative of the full KG. All benchmark questions derive from held-out triples; the held-out condition removes those topics, including their answer-bearing triples, from the retrieval index. This condition is deliberately diagnostic: it probes what happens when coverage is absent and does not isolate generalization independently of that removal.

Relation	Example Object	Count
HasNumericValue	3,372 GW	24,523
HasPercentileValue	28% of generation	6,258
HasMaxValue	58°C	3,907
HasContext	as of 2022	3,664
PercentageOf	70% of freshwater	2,534
HasMinValue	−89.2°C	2,507
HasCapacity	22,500 MW	1,829
HasPolicyTarget	55% by 2030	1,020
21 additional relations		4,504

Table 2: Top 8 statistical relations by frequency (of 29 total). Full schema in Appendix A.

Question generation. We select 500 triples via stratified sampling across all 29 relation types and generate natural-language questions using an instruction-tuned LLM (Devstral 24B, temperature 0.7 for phrasing diversity), constrained to target the specific numerical value without revealing the answer. Human verification found 99.4% of questions acceptable (α=0.70); see Appendix B. To assess whether the extracted gold values extend beyond agreement with Wikipedia, a stratified 100-question subset was checked against primary sources (Appendix B.4). The benchmark spans all relation types (Appendix L, Figure 12); 347 questions have a single gold number and 153 have multiple.

Evaluation metrics. **Numerical Exact Match (NEM)**: 1 if any extracted number matches a gold number within 1% relative tolerance, chosen to accommodate rounding variation in Wikipedia sources (e.g., "3,372 GW" vs. "3,370 GW") without accepting meaningfully different values. **Numerical F1 (NF1)**: token-level F1 between predicted and gold number sets. **Factual Precision (FP)**: fraction of predicted numbers matching gold. NEM is our primary metric; NF1 and FP are reported in Appendix E.

Table 3 reports NEM scores across all method–

model combinations on the full-index condition.

5 Experimental Setup

Retrieval methods. We evaluate eight strategies in four categories. *No retrieval:* **Zero-shot** uses only parametric knowledge. *Passage/triple retrieval:* **Standard RAG** retrieves top-5 Wikipedia paragraph chunks; **KG Cosine** retrieves top-5 triples by cosine similarity, optionally with neural reranking; **Graph RAG** extends KG Cosine with one-hop neighbor expansion (up to 50 triples). *Multi-pass verification:* **CoNLI** generates a zero-shot answer, retrieves KG triples, classifies each claim via NLI, and revises contradictions; **CoVe** (Dhuliawala et al., 2024) self-verifies and refines. *External graph retrieval:* **LightRAG** (Guo et al., 2024) and **MS GraphRAG** (Edge et al., 2024) index EcoStatKG triples directly (bypassing their built-in extraction).

Models. We evaluate six API-accessible LLMs: Mistral-Small-24B (Jiang et al., 2023), Ministral-14B (Jiang et al., 2023), Qwen-3.6-27B (Qwen Team, 2025), DeepSeek-V4-Flash (DeepSeek-AI, 2024), GLM-5 (Team GLM et al., 2024), and GPT-OSS-120B (Agarwal et al., 2025). All use temperature 0.1, top-p 0.95, and a maximum of 256 generation tokens. We use Qwen3-Embedding-8B (4,096 dimensions) and Qwen3-Reranker-0.6B. See Table 12 (Appendix D) for full details.

Evaluation protocol. Each method–model pair is evaluated on all 500 questions under two conditions: (1) **full-index**, an oracle-coverage upper-bound reference in which all 50,746 triples, including answer-bearing triples, are available; and (2) **held-out**, in which the 9,760 held-out triples are removed and 40,986 remain. The held-out condition diagnoses retrieval without answer-bearing evidence; it is not an estimate of performance under a naturally occurring deployment distribution. Zero-shot, CoNLI, and CoVe serve as controls because their initial answers do not depend on topic-filtered retrieval.

6 Results

6.1 Main Results: Full-Index Evaluation

KG-grounded retrieval dominates under oracle coverage. KG Cosine (0.989) and Graph RAG (0.992) achieve near-ceiling NEM across all six models. Their pooled per-question difference is small and not statistically significant (Table 4); paired tests are also non-significant for each individual model. The per-model ranges are 0.984–0.996 for KG Cosine and 0.984–1.000 for Graph RAG. LightRAG achieves 0.922, while Standard RAG (0.669) confirms the value of textual evidence but falls short of structured triple retrieval. This comparison measures retrieval when the answer-bearing fact is available, not reasoning over missing evidence.

Multi-pass and community-based methods underperform. CoVe (0.496) actually scores *below* zero-shot (0.508)—its self-verification step frequently "corrects" accurate initial responses by substituting parametric errors. CoNLI (0.717) provides a modest improvement via NLI-based revision but remains far below direct retrieval. MS GraphRAG (0.397) also falls below zero-shot—its community-summarization approach aggregates triples into prose, which often omits or rounds specific numerical values. Per-method analysis is in Appendix E.

6.2 Main Results: Held-Out Evaluation

Table 5 reports NEM on the held-out condition, where retrieval methods cannot access triples from the benchmark topics.

Retrieval can act as a distractor. The held-out drop is partly a consequence of the design: the answer-bearing triples are removed by construction. The diagnostic result is that direct retrieval methods then perform below zero-shot. KG Cosine drops from 0.989 to 0.327 (-66.2 points), Graph RAG from 0.992 to 0.342 (-65.0), LightRAG from 0.922 to 0.326 (-59.6), and Standard RAG from 0.669 to 0.316 (-35.3). The paired KG Cosine–Graph RAG difference is significant in this condition (Table 4), but both remain far below zero-shot. These results show that topically related numerical context can displace parametric answers under this controlled missing-coverage condition; not that retrieval will always hurt in deployment.

Multi-pass methods are condition-dependent. In held-out evaluation, CoNLI (0.507) and CoVe (0.497) remain near the zero-shot baseline (0.506). Under full-index coverage, CoNLI reaches 0.717, while CoVe (0.496) remains slightly below zero-shot (0.508). Thus, multi-pass behavior depends on whether useful evidence is available; it is not uniformly stable across conditions. Comparing Tables 3 and 5 (visualized in Figure 2) shows that

Method	mistral-sm.	ministr.	qwen-27b	dseek-v4	glm-5	gpt-120b	Avg.
Zero-shot	0.410	0.534	0.522	0.466	0.620	0.494	0.508
Standard RAG	0.640	0.678	0.684	0.658	0.700	0.652	0.669
KG Cosine	0.984	0.988	0.996	0.986	0.996	0.984	**0.989**
Graph RAG	0.988	0.990	1.000	0.992	0.998	0.984	**0.992**
CoNLI	0.554	0.720	0.826	0.728	0.712	0.760	0.717
CoVe	0.412	0.532	0.500	0.480	0.570	0.480	0.496
LightRAG	0.908	0.920	0.938	0.926	0.928	0.914	0.922
MS GraphRAG	0.376	0.420	0.408	0.400	0.398	0.382	0.397

Table 3: Numerical Exact Match (NEM) on NumHallu with full-index retrieval, an oracle-coverage upper-bound condition in which gold triples are available. KG Cosine and Graph RAG achieve near-ceiling accuracy; this condition is not a reasoning test.

Condition	n	KG Cosine	Graph RAG	Δ	95% CI	p_{boot}	p_{McN}
Full-index	3,000	0.989	0.992	-0.003	$[-0.007, 0.001]$	0.14	0.18
Held-out	3,000	0.327	0.342	-0.015	$[-0.024, -0.006]$	0.0006	0.0015

Table 4: Paired significance tests for KG Cosine minus Graph RAG on per-question NEM, pooled across six models. Confidence intervals use 10,000 paired bootstrap resamples; p_{McN} is the paired McNemar test. Under full-index coverage the difference is not significant; the held-out difference is significant, but both the methods remain below zero-shot.

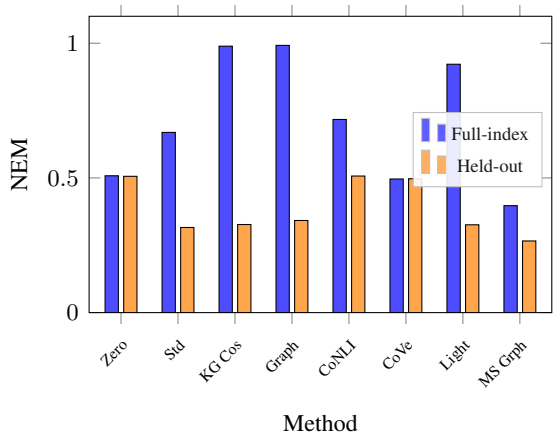


Figure 2: Full-index vs. held-out NEM by method (averaged across 6 models). Retrieval methods collapse when gold triples are absent, while zero-shot and multi-pass methods remain stable.

performance changes substantially with coverage, while the held-out result is a diagnostic of the deliberately missing-answer setting.

6.3 Similarity Threshold Filtering

If low-confidence retrievals are the culprits, a simple diagnostic intervention is similarity-threshold filtering: fall back to zero-shot when no retrieved triple exceeds a minimum cosine threshold. Table 6 reports this damage-control test.

With a threshold of 0.20, KG Cosine rises from 0.327 to 0.511 and Graph RAG from 0.342 to 0.517, both near zero-shot (0.506). This is recovery to the parametric baseline, not evidence of improvement beyond it. Paired tests pooled over the six models ($n = 3{,}000$; 10,000 bootstrap resamples and McNemar) confirm this: gating at $t = 0.20$ significantly improves KG Cosine over its unfiltered run ($\Delta = +0.184$, 95% CI [0.168, 0.200], $p < 0.0001$), whereas the gated result is statistically indistinguishable from zero-shot ($\Delta = +0.005$, 95% CI $[-0.006, 0.017]$, $p = 0.39$; McNemar $p = 0.41$) and from gated Graph RAG ($\Delta = -0.006$, 95% CI $[-0.017, 0.005]$, $p = 0.29$). Stricter thresholds (0.25–0.35) begin filtering correct retrievals as well. Per-model threshold breakdowns are in Appendix F (Figure 7); tolerance sensitivity is reported in Appendix H.

6.4 Clinical Cross-Domain Proof of Concept

To test whether the diagnostic pattern is specific to environmental sustainability, we ran a bounded proof of concept on clinical/public-health epidemiology using the same pipeline with domain-specific seed categories. The clinical run used Qwen-3.6-27B for triple extraction, rather than Devstral 24B used for the environmental KG. The resulting KG contains 5,174 statistical triples across 27 relations, and the benchmark contains 120 questions. We evaluate six methods on three models (Qwen-3.6-27B, Ministral-14B, and DeepSeek-V4-Flash); LightRAG and MS GraphRAG are not included in this bounded study.

The same two diagnostic signatures appear: un-

Method	mistral-sm.	ministr.	qwen-27b	dseek-v4	glm-5	gpt-120b	Avg.
Zero-shot	0.420	0.544	0.506	0.476	0.590	0.498	0.506
Standard RAG	0.292	0.356	0.336	0.308	0.318	0.288	0.316
KG Cosine	0.300	0.376	0.338	0.320	0.318	0.310	0.327
Graph RAG	0.298	0.378	0.354	0.330	0.378	0.312	0.342
CoNLI	0.426	0.540	0.510	0.478	0.610	0.478	0.507
CoVe	0.416	0.522	0.490	0.466	0.588	0.498	0.497
LightRAG	0.294	0.356	0.326	0.322	0.358	0.300	0.326
MS GraphRAG	0.252	0.292	0.270	0.270	0.268	0.246	0.266

Table 5: NEM on NumHallu with held-out evaluation (answer-bearing triples excluded from the retrieval index). Direct retrieval methods fall below zero-shot; multi-pass methods remain near the baseline.

Threshold	KG Cosine	Graph RAG
0.15 (strict)	0.509	0.508
0.20	0.511	0.517
0.25	0.498	0.497
0.30	0.481	0.486
0.35	0.433	0.445
None (no filter)	0.327	0.342

Table 6: NEM on held-out evaluation with cosine similarity threshold filtering (averaged across 6 models). Thresholds between 0.15 and 0.20 recover performance to zero-shot level by filtering out distractor triples.

Method	Full-index	Held-out
Zero-shot	0.594	0.589
Standard RAG	0.719	0.378
KG Cosine	0.992	0.333
Graph RAG	0.997	0.344
CoNLI	0.819	0.594
CoVe	0.611	0.600

Table 7: Clinical/public-health proof-of-concept NEM, averaged over three models and 120 questions. Full-index is the oracle-coverage reference; held-out removes answer-bearing triples.

der full coverage, KG Cosine and Graph RAG are near ceiling and statistically indistinguishable (paired bootstrap $\Delta = -0.006$, 95% CI $[-0.017, 0.006]$, $p = 0.46$; McNemar $p = 0.63$, $n = 360$); under held-out, both fall below zero-shot. In the two-model threshold subset ($n = 240$ paired questions), gating KG Cosine at $t = 0.20$ yields a significant recovery from 0.308 to 0.600 ($\Delta = +0.29$, 95% CI [0.23, 0.35], $p < 0.0001$; McNemar discordant pairs 71–1) and is statistically indistinguishable from zero-shot ($\Delta = +0.004$, 95% CI $[-0.033, 0.042]$, $p = 0.93$; McNemar $p = 1.00$). KG Cosine and gated Graph RAG are also indistinguishable ($\Delta = +0.008$, 95% CI $[-0.021, 0.037]$, $p = 0.70$). Thus, the small-scale result supports transfer of the diagnostic pattern and baseline-recovery effect, not a general claim across clinical tasks. This proof of concept is bounded, and we do not release its clinical KG without expert medical validation.

7 Analysis

7.1 Why Retrieval Hurts Without Coverage

The held-out results reveal a *faithfulness–accuracy trade-off*: instruction-tuned LLMs are trained to ground outputs in provided context, which becomes counterproductive when retrieved context is numerically irrelevant. We identify three mechanisms (detailed in Appendix I): (1) **distractor injection**, where topically similar but numerically wrong triples are faithfully used by the model; (2) **parametric override suppression**, where retrieved context actively suppresses correct parametric knowledge, explaining the drop *below* zero-shot; and (3) **multi-hop dilution**, where graph expansion amplifies the distractor effect. These mechanisms motivate confidence-based gating (§6.3): at t=0.20, KG Cosine recovers from 0.327 to 0.511 NEM by falling back to zero-shot when no high-confidence triple is found.

7.2 Error Taxonomy

Manual classification of 500 held-out KG Cosine responses per model reveals a consistent pattern across four error categories. The three models include Devstral 24B, the extraction model, which is not one of the six evaluation models. Table 8 shows the distribution is consistent across models: retrieval failure (wrong triple surfaced) accounts for ∼59% of responses, while hallucination despite correct injection is rare ($\leq$4%). This confirms that the bottleneck is retrieving the *right* triple, not the model's ability to extract answers from it. Full analysis with examples in Appendix J (Figure 10).

7.3 Latency–Accuracy Trade-off

Figure 3 plots full-index NEM against mean per-query latency measured on the held-out runs. KG Cosine and Graph RAG occupy the Pareto frontier, achieving the highest NEM at moderate latency (∼4s per query). CoVe and CoNLI incur 3–7× the latency of single-pass retrieval for lower NEM, placing them firmly in Pareto-dominated territory. LightRAG achieves high accuracy (0.922) but at higher latency due to graph traversal overhead. MS GraphRAG's accuracy (0.397) does not justify its cost. Detailed latency breakdowns are in Appendix K (Table 22, Figure 11).

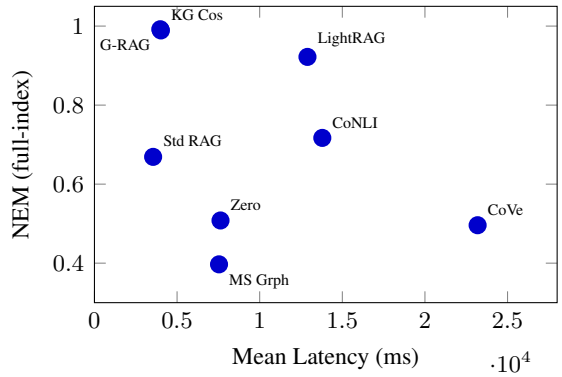


Figure 3: Full-index NEM vs. mean per-query latency (latency from the held-out runs, Table 22). KG Cosine and Graph RAG achieve the best accuracy–efficiency trade-off. Multi-pass methods (CoVe, CoNLI) are Pareto-dominated.

7.4 Supporting Analyses

Model variation. GLM-5 leads zero-shot NEM (0.620 full-index, 0.590 held-out), hinting at stronger parametric encoding of environmental statistics. Model differences shrink under KG-grounded retrieval: the full-index range is 1.2 points for KG Cosine (0.984–0.996) and 1.6 points for Graph RAG (0.984–1.000), compared with 21.0 points for zero-shot (0.410–0.620). This indicates reduced model-level variation when answer-bearing triples are available. Full per-model breakdowns and a visual comparison (Figure 6) are in Appendix E.

Category	devstral	gpt-120b	qwen-27b
Correct	35.0%	31.0%	33.8%
Retrieval failure	59.2%	59.8%	59.6%
Injection failure	4.6%	5.2%	5.0%
Halluc. despite inj.	1.2%	4.0%	1.6%

Table 8: Error taxonomy on held-out KG Cosine (500 questions, 3 models). Retrieval failure dominates; hallucination despite correct injection is rare.

Clinical cross-domain support. The bounded clinical study shows the same pattern: KG Cosine and Graph RAG perform similarly with full coverage, while both fall below zero-shot when answer-bearing triples are held out. Threshold filtering restores KG Cosine to near-baseline performance, but this small study does not establish broad clinical generalization (Section 6.4).

Graph-based retrieval. MS GraphRAG reaches 0.397 full-index NEM, below zero-shot (0.508), while LightRAG reaches 0.922. Community summarization may omit or alter specific values, which is a poor fit for exact numerical QA; LightRAG instead retrieves individual triples.

Ablation and sensitivity. We ablate retrieval granularity (fine, medium, coarse) and top-k (1, 3, 5, 7, 10) for KG Cosine: fine-grained at k=3 performs marginally best; k>5 provides no benefit. Neural reranking provides only +1.4 NEM points; temperature (0.0–1.0) and output format (prose vs. structured) have similarly marginal effects. Together, these results reinforce a consistent theme: what you retrieve matters far more than how you tune the retrieval. Details are in Appendices G (Figure 8) and H (Figure 9). Further design choices, are discussed in Appendix N.

8 Conclusion

This paper introduces EcoStatKG, a knowledge graph of 356,977 triples from 6,871 Wikipedia articles, and NumHallu, a 500-question numerical benchmark. Across eight retrieval strategies and six LLMs, the full-index condition provides an oracle-coverage upper bound: simple cosine retrieval performs comparably to graph retrieval when the answer-bearing triple is present. In the deliberately constructed held-out condition, direct retrieval methods fall below zero-shot as related but incorrect numbers act as distractors. A bounded clinical proof of concept reproduces this pattern. Similarity gating recovers performance to approximately the zero-shot baseline in both domains; it is damage control, not a new state-of-the-art method. These results support prioritizing knowledge-base coverage and evaluating retrieval confidence, while leaving broader deployment generalization to future work.

Limitations

1. **Domain scope.** EcoStatKG targets environmental sustainability. A bounded clinical/public-health proof of concept reproduces the main diagnostic pattern, but it uses only 120 questions, three models, and six methods; it does not establish broad transfer across clinical tasks or domains. We do not release the clinical KG without expert medical validation.

2. **Source reliability.** The main KG and benchmark are derived from English Wikipedia, which can contain errors and lag behind primary sources. A primary-source verification of 100 benchmark answers found 96 exact matches, one minor approximation, one outdated value, one incorrect value, and one unverifiable value. This limited sample does not constitute a full independent audit of the benchmark.

3. **Extraction quality.** Statistical triple extraction uses a single LLM (Devstral 24B) with constrained prompting. Human evaluation of 500 triples yields 91.6% full correctness (α=0.92); 6.0% are partially correct, mostly due to missing temporal qualifiers or imprecise subjects, and 2.4% are incorrect, primarily misattributed values (Appendix B). LLM-based KG construction is therefore reliable at scale but not error-free.

4. **Benchmark construction.** NumHallu questions are LLM-generated from triples extracted by Devstral 24B, so the extractor contributes to both question provenance and gold construction. The 100-question primary-source verification reduces, but does not eliminate, this concern; the sample is limited in size. The 500-question set supports system-level comparisons but not high-powered per-relation conclusions.

5. **Model access.** All models are accessed via API without weight or tokenizer access, reflecting a realistic deployment scenario. Reported latencies include API overhead, providing practitioners with end-to-end cost estimates directly applicable to production settings.

6. **Evaluation metric.** NEM uses a 1% relative tolerance. Re-scoring stored predictions at 0.5%, 2%, and 5% leaves the method rankings unchanged, although absolute scores increase with more permissive tolerance (Appendix H). NEM does not assess units, qualifiers, or numerical reasoning beyond answer matching.

Ethics Statement

This work uses publicly available Wikipedia content and does not collect private or personally identifiable information. Human verification was conducted on samples of public-text triples and benchmark questions, and a supplementary primary-source check was performed. The study did not recruit external participants or collect personal data. The LLMs used in this work are accessed via authorized API endpoints. We do not attempt to extract training data, circumvent safety filters, or use the models for purposes outside their intended use. We note that automated KG construction from Wikipedia may propagate geographic and topical biases in its environmental coverage. The KG reflects Wikipedia at extraction time and should not be treated as a substitute for primary scientific sources. The clinical proof of concept is not released because clinical facts require expert validation.

Acknowledgments

The authors gratefully acknowledge the generous support from the project titled "Conversational Agents with Negotiation and Influencing Ability," sponsored by Accenture. GitHub Copilot was used for writing assistance, including grammar correction and paraphrasing. All generated content was reviewed and validated by the authors.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human

knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Pier Luigi Buttigieg, Norman Morrison, Barry Smith, Christopher J Mungall, Suzanna E Lewis, and Envo Consortium. 2013. The environment ontology: contextualising biological and biomedical entities. *Journal of biomedical semantics*, 4(1):43.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Abu Dhabi. 2020. Irena. *Renewable energy statistics*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.

Jialin Dong, Bahare Fatemi, Bryan Perozzi, Lin F Yang, and Anton Tsitsulin. 2024. Don't forget to connect! improving rag with graph-based reranking. *arXiv preprint arXiv:2405.18414*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Shanghua Gao, Richard Zhu, Zhenglun Kong, Ayush Noori, Xiaorui Su, Curtis Ginder, Theodoros Tsiligkaridis, and Marinka Zitnik. 2025. Txagent: an ai agent for therapeutic reasoning across a universe of tools. *arXiv preprint arXiv:2503.10970*.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2(3).

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. 2025b. To trust or not to trust? enhancing large language models' situated faithfulness to external contexts. In *International Conference on Learning Representations*, volume 2025, pages 8625–8654.

Jeff Huber and Anton Schmitt. 2023. Chroma: The open-source embedding database for ai. *GitHub repository*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 6449–6464.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. Rafe: Ranking feedback improves query rewriting for rag. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 884–901.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1906–1919.

Alistair Miles and Sean Bechhofer. 2009. Skos simple knowledge organization system reference. w3c recommendation 18 august 2009. *World Wide Web Consortium. https://www. w3. org/TR/2009/REC-skos-reference-20090818*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Robert G Raskin and Michael J Pan. 2005. Knowledge representation in the semantic web for earth and environmental terminology (sweet). *Computers & geosciences*, 31(9):1119–1125.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific.

Kun Zhang, Xiexiong Lin, Yuanzhuo Wang, Xin Zhang, Fei Sun, Cen Jianhe, Hexiang Tan, Xuhui Jiang, and Huawei Shen. 2023. Refsql: A retrieval-augmentation framework for text-to-sql generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 664–673.

Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, and 1 others. 2023. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36:44502–44523.

Shuangjia Zheng, Jiahua Rao, Ying Song, Jixian Zhang, Xianglu Xiao, Evandro Fei Fang, Yuedong Yang, and Zhangming Niu. 2021. Pharmkg: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in bioinformatics*, 22(4):bbaa344.

Angelo Ziletti and Leonardo DAmbrosi. 2024. Retrieval augmented text-to-sql generation for epidemiological question answering using electronic health records. In *Proceedings of the 6th clinical natural language processing workshop*, pages 47–53.

A Relation Schema

Table 9 lists all 29 statistical relations in EcoStatKG with definitions, examples, and frequency counts. Figure 5 shows the distribution on a log scale, highlighting the long-tailed nature of the schema.

B Human Quality Evaluation

We conducted human verification to validate the quality of EcoStatKG triples and NumHallu benchmark questions. The verification guidelines, items, and annotations are included with the study materials.

B.1 KG Triple Quality

Protocol. Human verification covered a stratified sample of 500 triples (sampled proportionally across all 29 relation types). For each triple, the subject, relation, object, source topic, and original Wikipedia source sentences were reviewed. Each triple received a score: **2** (correct—fully supported by source), **1** (partially correct—minor issues such as imprecise subject, missing temporal context, or debatable relation type), or **0** (incorrect—value not in source, misattributed, or fabricated). Verification was based on the provided source sentences without consulting external sources. Scores were assigned independently; disagreements with $|\Delta| \geq 2$ were resolved by discussion.

Scoring guidelines. The following criteria governed scoring decisions:

- Values rounded within 1% of the source value: score 2.
- Subject more specific than source but claim still accurate: score 2.
- Value correct but missing important context (date, location): score 1.
- Value correct but wrong unit: score 1.
- Value appears in source but attributed to wrong entity: score 0.

Results. Table 10 summarizes the results. Inter-annotator agreement was α=0.92 (Krippendorff's alpha), indicating almost perfect reliability. The majority of errors in score-1 triples involved missing temporal qualifiers or slightly imprecise subject descriptions; score-0 errors were primarily misattributed values (correct number, wrong entity).

B.2 Benchmark Question Quality

Protocol. Human verification covered a stratified sample of 100 benchmark items on two dimensions: (1) **question quality** (3=good: clear and unambiguous, 2=acceptable: understandable with minor issues, 1=poor: ambiguous or misleading) and (2) **answer correctness** (3=correct: directly answers the question, 2=partially correct: numerically correct but incomplete, 1=incorrect: does not answer the question). Each item was reviewed with its question, gold answer, source triple (subject, relation, object), and source topic.

Acceptability criterion. A benchmark item is considered acceptable if both dimensions receive a score $\geq$2.

Results. Table 11 reports the results. In total, 99.4% of items met the acceptability criterion. Inter-annotator agreement was α=0.70 for question quality and α=0.82 for answer correctness. The single unacceptable item involved a question that asked for both minimum and maximum values while the gold answer contained only the minimum.

B.3 Agreement Computation

We report Krippendorff's α (Krippendorff, 2018) rather than Cohen's κ as our primary agreement metric, as α handles ordinal scales and is more robust to marginal distribution imbalances. Cohen's κ values (computed for reference) were 0.89 for KG quality and 0.67/0.79 for benchmark question quality/answer correctness, consistent with the α values. The IAA computation script is included in the released repository.

B.4 Independent Primary-Source Audit

To test whether the gold values reflect facts beyond Devstral's extraction of Wikipedia, a stratified sample of 100 benchmark answers was checked against primary or official sources traced from the Wikipedia citations. The check found 96 exact matches, one minor approximation, one outdated value, one incorrect value, and one value that could not be independently verified. Re-scoring the evaluated models on this subset left their ranking unchanged. This supplementary check is separate from the agreement statistics reported above.

The minor case was the claim that Pluie's tracked area was 10 times the area of Yellowstone: the cited area values imply approximately 11.5 times, making 10 a rounded approximation. The outdated

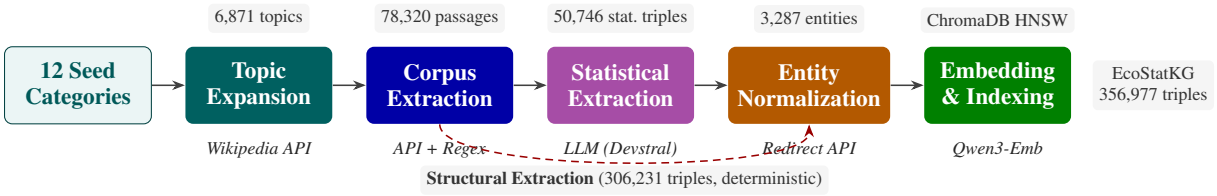


Figure 4: EcoStatKG construction pipeline. Colored boxes represent processing stages; gray badges show intermediate outputs. Statistical extraction (violet) uses LLM-based parsing; all other stages are deterministic or API-based. The structural branch extracts hyperlinks, categories, and redirects without LLM involvement.

Relation	**Example (subject → object)**	**Count**
HasNumericValue	Solar energy capacity → 3,372 GW	24,523
HasPercentileValue	Renewable electricity share → 28% of generation	6,258
HasMaxValue	Sahara Desert temperature → 58°C	3,907
HasContext	Global CO_2 emissions → as of 2022	3,664
PercentageOf	Agriculture water use → 70% of freshwater	2,534
HasMinValue	Antarctic temperature → −89.2°C	2,507
HasCapacity	Three Gorges Dam → 22,500 MW	1,829
HasPolicyTarget	EU emissions reduction → 55% by 2030	1,020
HasGrowthRate	Solar energy → 26% annually	691
HasAverageValue	Global temperature anomaly → 1.1°C	649
Reduces	Reforestation → 3.6 Gt CO_2/year	523
HasComparison	Solar potential → 4,000× projected demand	456
VolumeOf	Shower water use → 50–80 litres	353
Increases	Sea level → 3.6 mm/year	274
Decreases	Arctic ice extent → 13% per decade	267
HasImpact	Deforestation → 10% of global emissions	256
Saves	Textile reuse → 164 GJ per tonne	205
EfficiencyOf	Solar PV panels → 20–25%	198
EmissionOf	Industry → 14 Gt CO_2-eq	194
ConsumptionOf	Noor I CSP → 4.6 liters/kWh	134
HasTrend	Environmental indicator → numeric trend	94
RateOf	Fresh water use → 10 litres/second	68
HasStatistic	Air pollution → 6.7 million deaths/year	56
RiskOf	Annual flood probability → 200:1	34
HasSource	Oregon program → Senate Bill 762	25
MitigationOf	Urban trees → 98 tonnes of NO_2	16
HasAction	Inflation Reduction Act → 45Q	4
HasUnitOfMeasurement	Radiative efficiency → W m^{-2} kg^{-1}	4
PolicyOf	Paris Agreement → 1.5°C target	3

Table 9: Complete relation schema for EcoStatKG (29 relations, 50,746 total triples). Examples are illustrative, not exact KG entries.

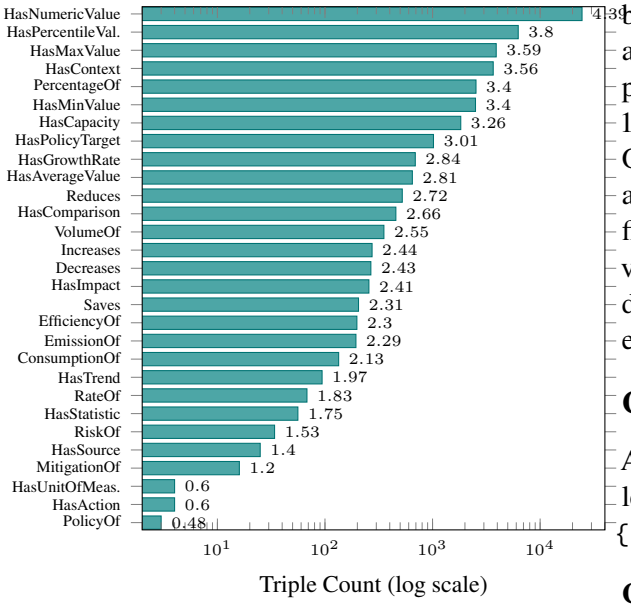


Figure 5: Triple count distribution across the corrected 29-relation schema (log scale). HasNumericValue dominates (48.3%); the five rarest relations together account for about 0.1% and are retained for schema expressiveness (see §N).

Score	Count	%
2 — Correct	458	91.6%
1 — Partially correct	30	6.0%
0 — Incorrect	12	2.4%
Krippendorff's α		0.92
Precision (score ≥ 1)		97.6%

Table 10: KG triple quality verification (500 triples). 91.6% of triples were rated fully correct.

	3 (Good)	2 (Accept.)	1 (Poor)
Question quality	88%	11%	1%
Answer correctness	94%	5%	1%
Overall acceptable (both $\geq$2)			99.4%

Table 11: Benchmark question quality verification (100 items). α=0.70 (question quality), α=0.82 (answer correctness).

case was WildCRU's count of more than 45 doctoral students, which matched the historical source but is now reported as 137. The incorrect case was a combustion-chamber maximum of 700°C; the primary source gives a range up to approximately 1,000°C. The unverifiable case was the historical Oregon aquifer-storage value of more than 3,000 acre-feet; the current official page reports a later figure of 6,137 acre-feet per year, and the original value could not be confirmed. These four cases distinguish source drift and unverifiability from an extraction error.

C Prompt Templates

All prompts are reproduced verbatim from the released codebase. Template variables are shown in `{braces}`.

C.1 System Prompts

Ungrounded (zero-shot, CoVe, CoNLI initial).

> *You are a factual assistant specializing in environmental statistics. Answer the question as accurately as possible. When citing numbers, percentages, or statistics, be precise. Do not fabricate numerical claims you are uncertain about.*

Grounded (KG Cosine, Graph RAG, Standard RAG, LightRAG, MS GraphRAG).

> *You are a factual assistant specializing in environmental statistics. Answer the question using the provided context. When the context contains specific numbers, percentages, or statistics, use those exact values. Do not fabricate numerical claims that are not supported by the context. If the context does not contain enough information, state that clearly.*

C.2 Retrieval Prompt Template

All grounded methods share the same user message format:

> *Context:*
> *[1] {triple_or_passage_1}*
> *[2] {triple_or_passage_2}*
> *...*
>
> *Question: {query}*

C.3 Statistical Triple Extraction

> *You are a precise information extraction system. Given a text passage about {topic}, extract ALL statistical/numerical facts as structured triples.*
>
> *Each triple must follow the format: {"subject": "<entity>", "relation": "<one of the allowed relations>", "object": "<numerical/statistical value>"}*
>
> *ALLOWED RELATIONS (use ONLY these): {relations_list}*

RULES:
1. The object MUST contain a specific number, percentage, rate, quantity, or date-bound statistic.
2. Do NOT extract qualitative or opinion-based statements.
3. Each triple must be self-contained — a reader should understand the fact from the triple alone.
4. If a sentence contains multiple statistics, extract one triple per statistic.
5. Prefer specific relations (HasGrowthRate, HasCapacity) over general ones (HasStatistic, HasNumericValue).

TEXT: {text}

Extract all statistical triples as a JSON array:

C.4 Benchmark Question Generation

Given the following verified statistical fact, generate a natural language question that specifically targets this statistic.

Fact: The {relation} of {subject} is {object}.

Generate a question that: (1) Asks for the specific numerical value, (2) Sounds natural (as a human would ask), (3) Does NOT contain the answer, (4) Could plausibly have multiple similar numerical answers (to test precision). Return ONLY the question.

C.5 CoNLI Prompts

CoNLI uses three LLM calls: (1) initial zero-shot generation, (2) NLI-based verification against retrieved triples, and (3) revision of contradicted claims.

NLI verification. System: *"You are a precise NLI fact-checker."*

You are a fact-checking assistant. Given a claim and a set of reference facts, determine if each claim is:
- SUPPORTED: The facts confirm this claim
- CONTRADICTED: The facts contradict this claim
- NOT_ENOUGH_INFO: The facts neither confirm nor contradict

Reference facts: {facts}

Claim: {claim}

For each numerical statement in the claim, output a JSON object:
{"verdict": "SUPPORTED|CONTRADICTED|NOT_ENOUGH_INFO", "corrected_claim": "<corrected version if CONTRADICTED, else null>"}

Output ONLY the JSON object.

Revision. System: *"You are a precise editor."*

Revise the following answer to correct any contradicted claims.

Original answer: {original}

Corrections needed: {corrections}

Provide the corrected answer, keeping all supported claims unchanged.

C.6 CoVe Prompts

CoVe uses four LLM calls: (1) initial zero-shot generation, (2) claim extraction, (3) self-verification, and (4) refinement.

Claim extraction.

Extract all factual claims from the following answer as a numbered list. Focus on claims containing specific numbers, dates, percentages, or statistics.

Answer: {answer}

List each factual claim on its own line, numbered 1., 2., etc.

Self-verification.

For each factual claim below, independently verify whether it is likely correct. Provide your best answer to each verification question.

Claims to verify: {claims}

For each claim, respond with:
CLAIM N: [CORRECT/INCORRECT/UNCERTAIN] - [your brief justification or corrected value]

Refinement.

Given the original answer and the verification results, produce a refined answer that corrects any incorrect claims while keeping correct ones.

Original answer: {original}

Verification results: {verifications}

Provide the refined, corrected answer:

D Model Details

Short Name	Full Model ID	Params	Context
mistral-small-24b	Mistral-Small-24B-Instruct-2501-FP8	24B	32k
ministral-14b	Ministral-3-14B-Instruct-2512	14B	128k
qwen-27b	Qwen-3.6-27B	27B	128k
deepseek-v4-flash	DeepSeek-V4-Flash	MoE	128k
glm-5	GLM-5	–	128k
gpt-oss-120b	GPT-OSS-120B	120B	32k
Embedding & Reranking			
qwen3-emb-8b	Qwen3-Embedding-8B (4096d)	8B	8k
qwen3-rerank	Qwen3-Reranker-0.6B	0.6B	8k

Table 12: Models used in experiments. All generation models use temperature 0.1, top-p 0.95, max tokens 256. All models are accessed via an OpenAI-compatible API served by vLLM.

All models are accessed via an OpenAI-compatible API endpoint served by vLLM (Jiang et al., 2023; Qwen Team, 2025; DeepSeek-AI, 2024; Team GLM et al., 2024). We use 1 API keys for parallel execution at a rate limit of 180 requests per minute per key. Qwen-3.6-27B is a "thinking" model; we disable chain-of-thought via `enable_thinking=False` to obtain direct answers.

E Extended Results

Tables 13–16 report Numerical F1 (NF1) and Factual Precision (FP) for both evaluation conditions. NF1 is a token-level F1 between predicted and gold number sets; FP is the fraction of predicted numbers matching gold numbers.

The NF1 and FP patterns reinforce the NEM findings: KG Cosine and LightRAG dominate on full-index (NF1: 0.830 and 0.806), while all methods converge to low NF1 on held-out (range: 0.137–0.204). A notable divergence is FP on held-out: retrieval methods show *higher* FP than zero-shot (e.g., MS GraphRAG 0.689 vs. zero-shot 0.174). This is an artifact of answer length—retrieval methods produce shorter, more focused answers with fewer total numbers, so the fraction matching gold happens to be higher even though NEM is lower.

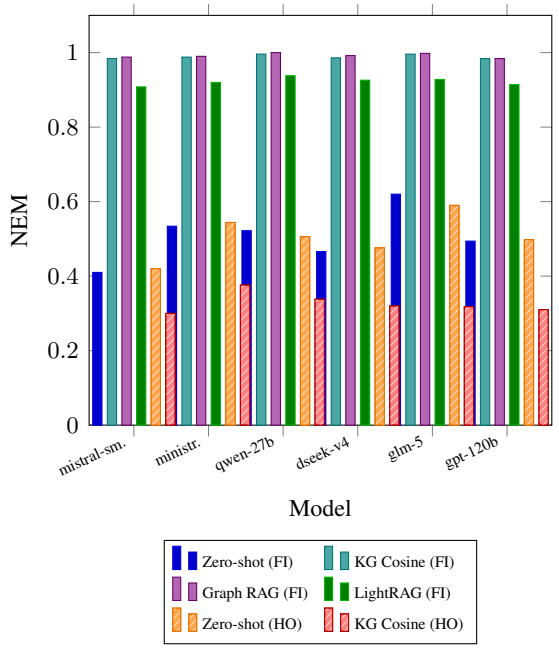


Figure 6: Per-model NEM for key methods under full-index (FI, solid) and held-out (HO, hatched) conditions. Model variation is large for zero-shot (21.0 pt range) but nearly vanishes under KG-grounded retrieval (1.6 pt), confirming that KG coverage neutralizes model-level differences.

F Threshold Filtering Details

Table 17 reports per-model NEM under cosine similarity threshold filtering (held-out condition). All models show the same pattern: thresholds between 0.15 and 0.20 recover performance toward zero-shot levels, while stricter thresholds (0.30+) begin filtering correct retrievals.

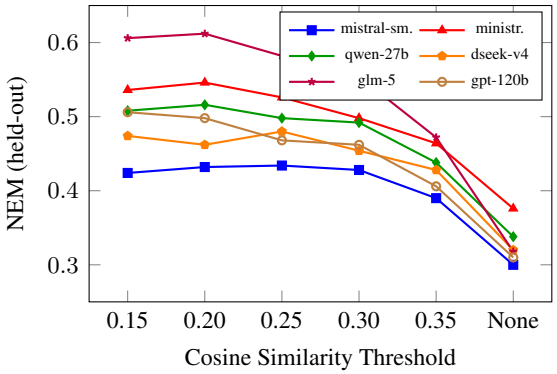


Figure 7: NEM vs. cosine similarity threshold for KG Cosine (held-out, per model). All models benefit from threshold filtering; GLM-5 benefits most due to its stronger parametric fallback.

G Ablation Study

We ablate retrieval granularity (fine, medium, coarse) and top-k (1, 3, 5, 7, 10) for KG Cosine on the held-out condition using devstral-24b (the extraction model, not one of the six evaluation models) with 3 random seeds per configuration (45 total runs). Table 18 reports the results.

Fine-grained retrieval at k=3 achieves the highest NEM (0.359±0.004), but differences across configurations are small: the full range spans 0.295–0.359 (6.4 points). Two patterns emerge: (1) coarse granularity consistently underperforms, likely because bundling all triples for a topic into a single document dilutes the relevance signal; (2) k>5 provides no benefit and occasionally degrades performance as additional distractor triples are included. Standard deviations across seeds are ≤0.004, indicating stable results.

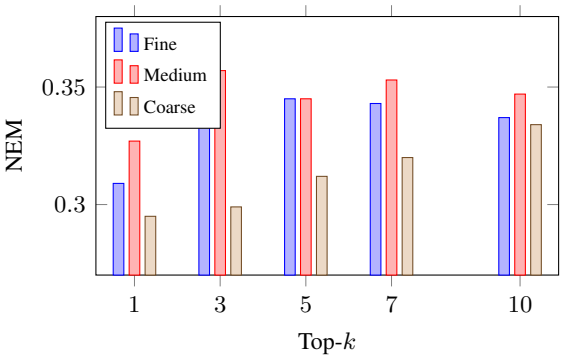


Figure 8: NEM by retrieval granularity and top-k (devstral-24b, held-out). Fine and medium peak at k=3; coarse improves monotonically with k but never catches up.

Method	mistral-sm.	ministr.	qwen-27b	dseek-v4	glm-5	gpt-120b	Avg.
Zero-shot	0.187	0.096	0.122	0.203	0.128	0.107	0.141
Standard RAG	0.536	0.430	0.448	0.514	0.517	0.421	0.478
KG Cosine	0.900	0.709	0.772	0.891	0.895	0.810	**0.830**
Graph RAG	0.895	0.646	0.765	0.864	0.877	0.715	0.794
CoNLI	0.305	0.159	0.303	0.401	0.156	0.207	0.255
CoVe	0.185	0.093	0.122	0.201	0.121	0.095	0.136
LightRAG	0.834	0.763	0.795	0.837	0.807	0.799	0.806
MS GraphRAG	0.304	0.286	0.295	0.314	0.295	0.293	0.298

Table 13: Numerical F1 (NF1) on full-index evaluation.

Method	mistral-sm.	ministr.	qwen-27b	dseek-v4	glm-5	gpt-120b	Avg.
Zero-shot	0.239	0.098	0.121	0.246	0.180	0.168	0.175
Standard RAG	0.719	0.460	0.488	0.619	0.602	0.496	0.564
KG Cosine	0.878	0.652	0.738	0.868	0.865	0.773	**0.796**
Graph RAG	0.874	0.574	0.742	0.834	0.847	0.619	0.748
CoNLI	0.326	0.142	0.299	0.378	0.223	0.262	0.272
CoVe	0.224	0.094	0.121	0.243	0.199	0.156	0.173
LightRAG	0.873	0.769	0.805	0.852	0.822	0.826	0.824
MS GraphRAG	0.823	0.526	0.592	0.705	0.646	0.751	0.674

Table 14: Factual Precision (FP) on full-index evaluation.

Method	mistral-sm.	ministr.	qwen-27b	dseek-v4	glm-5	gpt-120b	Avg.
Zero-shot	0.188	0.096	0.123	0.204	0.126	0.106	0.140
Standard RAG	0.216	0.159	0.167	0.200	0.191	0.180	0.185
KG Cosine	0.215	0.162	0.177	0.202	0.199	0.194	0.192
Graph RAG	0.215	0.142	0.170	0.193	0.154	0.185	0.176
CoNLI	0.189	0.099	0.126	0.210	0.122	0.104	0.142
CoVe	0.192	0.094	0.120	0.195	0.126	0.093	0.137
LightRAG	0.224	0.185	0.207	0.222	0.183	0.204	**0.204**
MS GraphRAG	0.186	0.170	0.178	0.190	0.178	0.172	0.179

Table 15: Numerical F1 (NF1) on held-out evaluation.

Method	mistral-sm.	ministr.	qwen-27b	dseek-v4	glm-5	gpt-120b	Avg.
Zero-shot	0.227	0.098	0.125	0.235	0.192	0.165	0.174
Standard RAG	0.731	0.334	0.391	0.586	0.536	0.564	0.524
KG Cosine	0.709	0.333	0.496	0.528	0.562	0.550	0.530
Graph RAG	0.749	0.311	0.494	0.533	0.507	0.549	0.524
CoNLI	0.229	0.095	0.126	0.241	0.184	0.167	0.174
CoVe	0.233	0.098	0.116	0.245	0.187	0.158	0.173
LightRAG	0.797	0.382	0.580	0.595	0.491	0.600	0.574
MS GraphRAG	0.828	0.508	0.626	0.726	0.683	0.764	**0.689**

Table 16: Factual Precision (FP) on held-out evaluation. Note: MS GraphRAG and retrieval methods show high FP despite low NEM because they produce fewer numbers overall, many of which happen to match some gold number.

Method	Threshold	mistral-sm.	ministr.	qwen-27b	dseek-v4	glm-5	gpt-120b
KG Cosine	t=0.15	0.424	0.536	0.508	0.474	0.606	0.506
	t=0.20	0.432	0.546	0.516	0.462	0.612	0.498
	t=0.25	0.434	0.526	0.498	0.480	0.582	0.468
	t=0.30	0.428	0.498	0.492	0.454	0.554	0.462
	t=0.35	0.390	0.464	0.438	0.428	0.472	0.406
	None	0.300	0.376	0.338	0.320	0.318	0.310
Graph RAG	t=0.15	0.420	0.524	0.514	0.474	0.606	0.512
	t=0.20	0.430	0.546	0.510	0.480	0.610	0.526
	t=0.25	0.426	0.516	0.502	0.464	0.580	0.496
	t=0.30	0.420	0.506	0.502	0.454	0.560	0.472
	t=0.35	0.390	0.460	0.474	0.432	0.506	0.406
	None	0.298	0.378	0.354	0.330	0.378	0.312

Table 17: Per-model NEM under threshold filtering (held-out). GLM-5 benefits most from filtering (+29.4 points at t=0.20 for KG Cosine), consistent with its stronger parametric knowledge. The best threshold lies between 0.15 and 0.25 for every model, and t=0.20 is best on average.

	k=1	k=3	k=5	k=7	k=10
Fine	.309±.001	**.359**±.004	.345±.002	.343±.002	.337±.004
Medium	.327±.001	.357±.001	.345±.004	.353±.004	.347±.004
Coarse	.295±.002	.299±.001	.312±.002	.320±.003	.334±.002

Table 18: NEM by retrieval granularity and top-k (devstral-24b, held-out, mean±std over 3 seeds). Fine k=3 is marginally best but differences are small. Coarse consistently underperforms.

H Sensitivity Analysis

We evaluate temperature, output format, and reranking using Devstral-24B, the extraction model, on the held-out condition; these single-model checks are not part of the six-model averages. We separately test NEM tolerance by rescoring stored predictions for all six evaluated models in both conditions.

H.1 Temperature

Temperature	KG Cosine	Zero-shot
0.0	0.344	0.484
0.1	0.340	0.508
0.3	0.342	0.498
0.5	0.334	0.498
0.7	0.340	0.496
1.0	0.344	0.504
Range	0.010	0.024

Table 19: NEM vs. temperature (devstral-24b, held-out). Both methods are stable: KG Cosine varies by 1.0 point and zero-shot by 2.4 points across the full temperature range.

NEM is remarkably stable across the temperature range 0.0–1.0 for both methods (Table 19, Figure 9). KG Cosine NEM varies by only 1.0 point (0.334–0.344); zero-shot varies by 2.4 points (0.484–0.508). This suggests that numerical extraction from context (or parametric recall) is not sensitive to sampling temperature, likely because correct numbers have high token probability regardless of temperature settings.

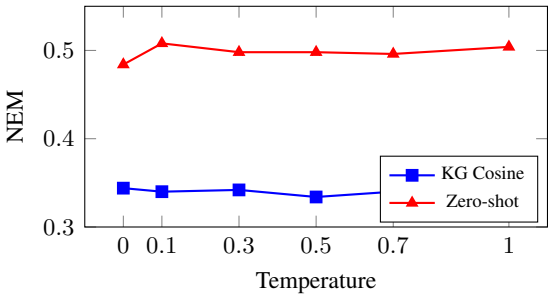


Figure 9: NEM vs. temperature (devstral-24b, held-out). Both methods are remarkably stable across the full range.

H.2 Output Format

Structured output format (KG Cosine, devstral-24b, held-out) achieves 0.344 NEM vs. 0.312 for prose format (+3.2 points). The structured format constrains the model to output numbers in a predefined JSON schema, reducing extraneous text that dilutes numerical extraction. We use structured format as the default across all experiments.

H.3 Reranker

Neural reranking with Qwen3-Reranker-0.6B (KG Cosine, devstral-24b, held-out) achieves 0.346 NEM vs. 0.332 without reranking (+1.4 points). The marginal improvement suggests that the initial cosine ranking is already effective for this task: the embedding model ranks the most relevant triples

highly, and a cross-encoder reranker provides only minor reordering. Given the additional latency cost, reranking is not included in the default pipeline.

H.4 NEM Tolerance

We re-score the stored predictions at relative tolerances of 0.5%, 1%, 2%, and 5%; no model inference is repeated. Table 20 reports the mean across the six evaluated models. The method ranking is unchanged across tolerances in both conditions. Scores increase slightly as tolerance becomes more permissive, but the main comparison is not an artifact of choosing 1%. At 1%, the re-scored full-index values reproduce the main table within 0.003.

I Retrieval Degradation Mechanisms

This appendix expands on the three mechanisms identified in Section 7 through which retrieval degrades numerical accuracy in the held-out condition.

Distractor injection. When the gold triple is absent from the index, cosine retrieval surfaces the most *topically* similar triples—which typically share the same subject entity or domain but contain different numerical values. For example, a question about Three Gorges Dam's annual generation (101.6 TWh) may retrieve its capacity triple (22,500 MW): topically near-identical, but numerically wrong. The model, trained to extract answers from context, faithfully reports the retrieved (incorrect) number rather than falling back to its parametric knowledge. This mechanism explains why KG Cosine drops from 0.989 to 0.327 NEM (-66.2 points) in the held-out condition, while zero-shot remains at 0.506.

Parametric override suppression. The drop is not symmetric: retrieval methods fall *below* zero-shot, not merely to its level. This indicates that retrieved context actively suppresses parametric recall. Even when the model "knows" the correct answer (as evidenced by zero-shot performance), the presence of authoritative-looking retrieved triples causes the model to override its parametric knowledge. This is consistent with prior findings on context–parametric conflicts (Huang et al., 2025b), where models preferentially trust provided context. The effect is strongest for models with high zero-shot NEM: GLM-5 drops from 0.590 to 0.318 (-27.2 points) under KG Cosine, the largest absolute drop of any model.

Multi-hop dilution. Graph-based methods (Graph RAG, LightRAG) expand the retrieved context through one-hop neighbor traversal, surfacing up to 50 triples per query. In the held-out condition, this expansion amplifies the distractor effect: more retrieved triples means more numerically irrelevant values competing for the model's attention. Graph RAG drops to 0.342 NEM (held-out), only marginally above KG Cosine's 0.327, despite its more sophisticated retrieval architecture.

Practical implication: confidence-based gating. Thresholding provides damage control in the held-out experiment: at t=0.20, KG Cosine rises from 0.327 to 0.511, near zero-shot (0.506), by falling back to parametric generation when retrieved evidence is weak. It does not supply missing facts or demonstrate an improvement over zero-shot. The clinical proof of concept likewise finds significant recovery but no significant difference from the zero-shot baseline (Section 6.4); this supports gating as a way to avoid regression, not as a new numerical-accuracy solution.

J Error Analysis

Table 21 expands on the error taxonomy from Section 7.2 with category definitions and representative examples.

Category definitions.

- **Correct:** The model's answer contains the exact gold number (within 1% tolerance).
- **Retrieval failure:** The retrieved triples do not contain the gold number. The model either outputs a wrong number from the retrieved context or states it cannot answer.
- **Injection failure:** The gold number appears in the retrieved triples, but the model fails to extract or use it correctly—typically selecting a different number from the same triple set.
- **Hallucination despite injection:** The gold number appears in the retrieved triples, but the model produces a different number not present in any retrieved triple, indicating the model "hallucinated" over correct evidence.

The dominance of retrieval failure ($\sim$59%) directly supports the coverage hypothesis: the bottleneck is whether the KG contains the needed

	Full-index				Held-out			
Method	**0.5%**	**1%**	**2%**	**5%**	**0.5%**	**1%**	**2%**	**5%**
Zero-shot	0.499	0.508	0.520	0.552	0.497	0.506	0.516	0.552
Standard RAG	0.668	0.669	0.676	0.688	0.314	0.316	0.325	0.337
KG Cosine	0.989	0.989	0.989	0.990	0.326	0.327	0.335	0.347
Graph RAG	0.992	0.992	0.992	0.992	0.341	0.342	0.350	0.367
CoNLI	0.711	0.717	0.724	0.741	0.500	0.507	0.523	0.554
CoVe	0.490	0.498	0.509	0.541	0.488	0.497	0.507	0.539
LightRAG	0.925	0.925	0.925	0.927	0.325	0.327	0.334	0.344
MS GraphRAG	0.396	0.398	0.402	0.406	0.266	0.266	0.270	0.276

Table 20: NEM at four relative tolerances, rescored from stored outputs and averaged across six models. Rankings are stable within each evaluation condition.

Category	**devstral**	**gpt-120b**	**qwen-27b**
Correct	175 (35.0%)	155 (31.0%)	169 (33.8%)
Retrieval failure	296 (59.2%)	299 (59.8%)	298 (59.6%)
Injection failure	23 (4.6%)	26 (5.2%)	25 (5.0%)
Halluc. despite inj.	6 (1.2%)	20 (4.0%)	8 (1.6%)

Table 21: Error taxonomy with counts and percentages (500 held-out KG Cosine responses per model). The consistency across models confirms retrieval failure as the dominant error mode.

triple, not whether the model can use it. The low hallucination-despite-injection rate ($\leq$4%) indicates that when models receive the correct triple, they are highly faithful to it. GPT-OSS-120B shows a slightly higher hallucination-despite-injection rate (4.0% vs. 1.2%–1.6%), suggesting it is less faithful to retrieved context than the smaller models.

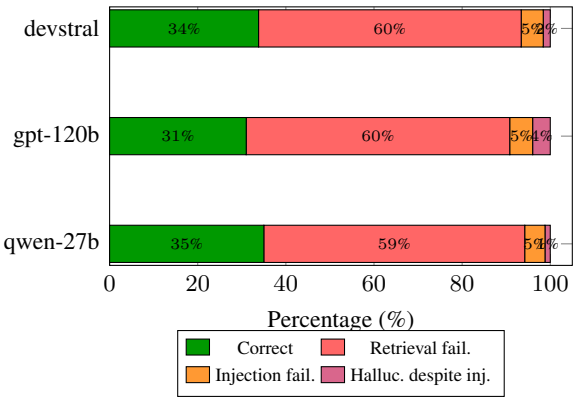


Figure 10: Error taxonomy distribution across 3 models (held-out KG Cosine). Retrieval failure dominates uniformly; GPT-OSS-120B shows higher hallucination-despite-injection.

K Latency and Efficiency

Table 22 reports mean latency per query and mean token usage per method, averaged across the six evaluation models from the held-out evaluation. Latency includes API round-trip time.

Method	**Passes**	**Lat. (ms)**	**Tokens**
Zero-shot	1	7,629	553
Standard RAG	1	3,551	1,046
KG Cosine	1	4,022	408
Graph RAG	1	3,985	1,156
LightRAG	1	12,891	2,229
MS GraphRAG	1	7,539	754
CoNLI	3	13,789	1,754
CoVe	4	23,188	2,917

Table 22: Mean latency and token usage per query (held-out). KG Cosine achieves the best accuracy at moderate cost. CoVe incurs 6$\times$ the latency of KG Cosine for lower NEM.

Zero-shot latency is unexpectedly high because models generate longer, more discursive answers when no context is provided (553 tokens vs. 408 for KG Cosine). KG Cosine is the most token-efficient grounded method, producing concise answers grounded in retrieved triples. CoVe's 4-pass architecture results in the highest latency (23.2s) and token usage (2,917), making it Pareto-dominated on both cost and accuracy axes.

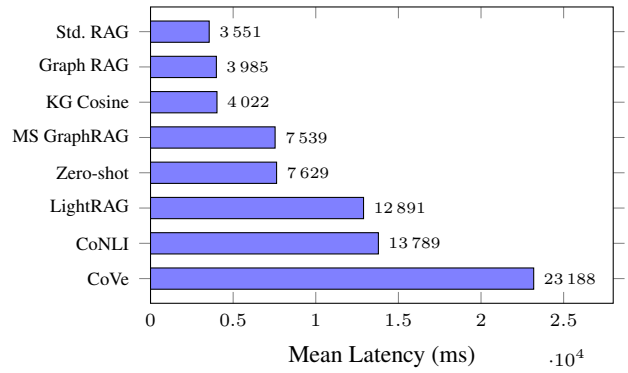


Figure 11: Mean per-query latency by method (held-out). Multi-pass methods (CoVe, CoNLI) incur 3–7$\times$ the latency of single-pass triple and passage retrieval.

L Benchmark Statistics

Table 23 reports KG totals, held-out counts, and benchmark questions for every relation. All 500 questions are derived from held-out triples, and no relation has more questions than held-out triples. In total, 9,760 triples are held out, leaving 40,986 in the retrieval index. HasSource, for example, has 25 KG triples, 20 held-out triples, and 17 questions; VolumeOf and RiskOf are both present in the KG. Of the 500 questions, 347 (69.4%) have a single gold number and 153 (30.6%) have multiple gold numbers.

Relation	KG total	Held-out	Questions
HasNumericValue	24,523	4,016	43
HasPercentileValue	6,258	1,096	25
HasMaxValue	3,907	688	19
HasContext	3,664	610	18
PercentageOf	2,534	624	18
HasMinValue	2,507	433	19
HasCapacity	1,829	368	18
HasPolicyTarget	1,020	317	18
HasGrowthRate	691	180	17
HasAverageValue	649	174	18
Reduces	523	181	19
HasComparison	456	124	17
VolumeOf	353	123	17
Increases	274	95	18
Decreases	267	100	17
HasImpact	256	87	18
Saves	205	75	17
EfficiencyOf	198	94	18
EmissionOf	194	111	17
ConsumptionOf	134	70	17
HasTrend	94	45	17
RateOf	68	38	17
HasStatistic	56	34	17
RiskOf	34	30	17
HasSource	25	20	17
MitigationOf	16	16	16
HasAction	4	4	4
HasUnitOfMeasurement	4	4	4
PolicyOf	3	3	3
Total	**50,746**	**9,760**	**500**

Table 23: EcoStatKG relation totals, held-out triples, and NumHallu questions. Every question is grounded in a held-out triple; the held-out split contains 9,760 triples and leaves 40,986 triples available for retrieval.

M Reproducibility

The full pipeline—from seed categories to benchmark evaluation—is implemented in Python 3.12 and requires API access to an OpenAI-compatible LLM endpoint; no local GPU is required. The release README documents the environment variables, dependencies, and end-to-end commands. Users can reproduce the reported evaluation directly from the released benchmark and indices, or rebuild the resource by running the construction stages below.

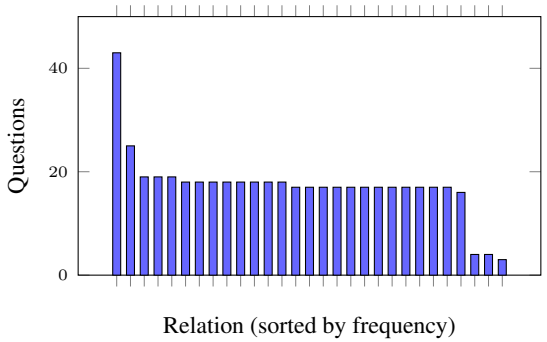


Figure 12: Question distribution across 29 relation types (sorted by frequency). The distribution is relatively uniform due to stratified sampling; the three rightmost bars correspond to rare relations (HasAction, HasUnitOfMeasurement, PolicyOf).

1. **Topic expansion** (~30 min): Wikipedia API category crawl + embedding-based filtering.
2. **Corpus extraction** (~2 hours): Wikipedia API article retrieval + regex sentence filtering.
3. **Triple extraction** (~6 hours): LLM-based statistical extraction + deterministic structural extraction.
4. **Entity normalization** (~1 hour): String normalization + Wikipedia redirect resolution.
5. **Embedding and indexing** (~2 hours): Dense embedding + ChromaDB index construction.
6. **Benchmark generation** (~1 hour): Stratified sampling + LLM-based question generation.
7. **Evaluation** (~4–8 hours per method–model combination): Query execution + metric computation.

Total wall-clock time for the complete construction pipeline is approximately 15–20 hours on a single machine; evaluation cost depends on the selected model and method. The released artifact includes the pipeline, relation schema, benchmark, and experimental results. The main paper uses six evaluation models; Devstral 24B is the extraction model and is not included in those six-model averages.

The extraction code and benchmark are released under the MIT License. Wikipedia-derived content and triples retain the applicable Wikipedia CC

BY-SA 3.0 attribution and share-alike terms. The clinical proof-of-concept KG is not part of the release.

N Design Choices

We address several design decisions that shaped EcoStatKG and NumHallu.

N.1 Why Not Use Wikidata?

Wikidata (Vrandečić and Krötzsch, 2014) is the largest open KG, encoding over 1.6 billion statements. However, its design prioritizes entity-level relationships (e.g., `instance_of`, `part_of`, `located_in`) over quantitative assertions. Of EcoStatKG's 29 relation types, Wikidata natively supports only a small subset—primarily `HasNumericValue` (via general-purpose quantity properties like P1114, P2046) and `HasCapacity` (via P2109). Relations central to our benchmark have no Wikidata equivalents: `HasGrowthRate`, `HasPolicyTarget`, `PercentageOf`, `EmissionOf`, `Reduces`, and `HasTrend` encode temporal dynamics, policy commitments, and causal relationships that Wikidata's schema does not capture.

Empirically, querying Wikidata's SPARQL endpoint for quantity-valued statements about the EcoStatKG topics returned no numerical triples, so no Wikidata retrieval index could be built for NumHallu. The same KG Cosine pipeline over EcoStatKG reaches 0.989 NEM under full-index coverage (Table 3), confirming that Wikidata's coverage gap for domain-specific statistics is the bottleneck, not the retrieval method. We emphasize that Wikidata excels at entity-level factuality; EcoStatKG is complementary, focusing on the numerical claims that general-purpose KGs underrepresent.

N.2 Why Retain Low-Count Relations?

Four relations have fewer than 20 triples: PolicyOf (3), HasAction (4), HasUnitOfMeasurement (4), and MitigationOf (16). We retain them because they capture semantic roles not covered by higher-frequency relations; for example, MitigationOf encodes the quantified effect of an intervention, distinct from the more general Reduces relation. These relations may also be useful when adapting the schema to other domains. Removing them would save negligible storage while reducing schema expressiveness.

N.3 Why Wikipedia as Source?

We chose Wikipedia over primary scientific databases (e.g., IPCC reports, IEA datasets, EPA databases) for three practical reasons. First, Wikipedia provides *encyclopedic breadth* across all environmental sub-domains in a single, consistently formatted corpus—no single primary source covers topics from renewable energy capacity to biodiversity loss to water scarcity. Second, Wikipedia's API enables *reproducible, cost-free extraction* without licensing restrictions, supporting open science. Third, Wikipedia articles aggregate and cite primary sources, serving as a curated intermediate layer. The trade-off is that Wikipedia may contain errors or lag behind primary sources; we quantify this risk via human evaluation (91.6% triple correctness, Appendix B) and a primary-source check of 100 benchmark answers (96 exact matches, Appendix B.4), and note it as a limitation.

N.4 Why Cosine Retrieval as the Base Method?

We use dense cosine retrieval (embedding similarity) rather than sparse methods (BM25, TF-IDF) as our base retrieval method. For structured KG triples—which are short, formulaic strings (e.g., "The HasCapacity of Three Gorges Dam is 22,500 MW")—dense embeddings capture semantic similarity more effectively than term overlap. BM25 would struggle with paraphrased queries (e.g., "What is the power generation capacity of Three Gorges?") where key terms differ from the triple's surface form. Our ablation (Appendix G) shows that retrieval granularity and k have small effects, suggesting the embedding space is well-calibrated for this task.

N.5 Why 500 Benchmark Questions?

NumHallu contains 500 questions, stratified across 29 relation types. This size balances three constraints: (1) *statistical power*—500 questions provide $\pm 4.4\%$ confidence intervals at 95% confidence for binary metrics like NEM; (2) *evaluation cost*—each method–model combination requires 500 API calls, and with 8 methods $\times$ 6 models $\times$ 2 conditions = 96 runs, the total is $\sim$48,000 API calls per full evaluation; and (3) *relation coverage*—stratified sampling ensures all 29 relation types are represented, with rarer relations receiving proportionally more questions to maintain minimum cell sizes. Scaling to 1,000+ questions would im-

prove per-relation statistical power but would double evaluation costs without changing the aggregate findings.