HERMES: Grounding Fidelity Evaluation in GraphRAG-Based Marine Knowledge Graphs Across LLM Configurations

Gustavo Nuñez[1,2,*], Marcos Zárate[1,2] and Pablo R. Fillottrani[3,4]

[1]Center for the Study of Marine Systems, (CESIMAR-CONICET), Puerto Madryn, Chubut, Argentina
[2]Computer Science Research Laboratory, (LINVI-UNPSJB), Puerto Madryn, Chubut, Argentina
[3]Research and Development Laboratory in Software Engineering and Information Systems (LISSI), Computer Science and Engineering Department, Universidad Nacional del Sur, Bahía Blanca, Argentina
[4]Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, Buenos Aires, Argentina

Abstract

Knowledge graphs constructed with GraphRAG from unstructured text provide a promising foundation for grounded natural-language question answering, but lack mechanisms to verify that generated responses actually respect the retrieved graph evidence. We present HERMES, a web-based platform for querying and comparative evaluation of six marine knowledge graphs derived from digitized historical scientific literature in the marine biology domain, each constructed using a different large language model (LLM). HERMES introduces a post-generation neuro-symbolic grounding verifier that decomposes responses into atomic claims and classifies each against retrieved subgraph triples, and two novel metrics — *Strict Grounding Fidelity* (GF_{strict}) and *Weighted Grounding Fidelity* (GF_{weighted}) — that quantify response traceability to graph evidence. A full factorial evaluation across 10 competency questions, 6 graphs, and 4 answer-generation LLMs (240 evaluations) shows that the choice of answer-generation LLM dominates grounding fidelity, that graph-level mean GF varied across the six KGs, with KG_1 and KG_5 outperforming the substantially sparser KG_3, and that the best combination (the knowledge graph constructed with GPT-4o-mini, queried with Llama 3.3 70B) achieves near-perfect grounding ($GF_{\text{weighted}} = 0.95$). This work presents an innovative faithfulness metric formally defined and evaluated over GraphRAG-style triple contexts.

Keywords

Knowledge Graphs, GraphRAG, Grounding Fidelity, Large Language Models, Marine Informatics, Neuro-Symbolic AI

1. Introduction

Large Language Models have demonstrated strong capabilities in knowledge extraction and natural-language reasoning, and frameworks such as Microsoft GraphRAG [1] now enable the automatic construction of knowledge graphs (KGs) from unstructured text, combining entity and relation extraction with graph-based retrieval for question answering. In domain-specific and historical corpora, where no gold-standard reference graph exists and ontological support is absent, this approach offers a practical pathway to structuring and querying legacy scientific knowledge [2].

A key open problem in this setting is *response grounding*: when a GraphRAG system retrieves a subgraph and passes its triples as context to an LLM, there is no guarantee that the generated response actually draws on that evidence rather than on the model's parametric knowledge. Hallucination and faithful-but-unsupported assertions remain significant concerns in LLM-based generation [1]. Existing evaluation frameworks for Retrieval-Augmented Generation, including RAGAs [3] and ARES [4], assess faithfulness with respect to retrieved textual contexts rather than graph triples.

The Third International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs (RAGE-KG 2026), co-located with ISWC 2026, October 25–26, 2026, Bari, Italy

*Corresponding author.
gnunez@cenpat-conicet.gob.ar (G. Nuñez); zarate@cenpat-conicet.gob.ar (M. Zárate); prf@cs.uns.edu.ar (P. R. Fillottrani)
0009-0007-0215-2699 (G. Nuñez); 0000-0001-8851-8602 (M. Zárate); 0000-0003-0906-867X (P. R. Fillottrani)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-04

CEUR-WS.org/Vol-4312/paper17.pdf

We present **HERMES**[1], a web-based platform that addresses this gap in the context of marine science knowledge graphs. HERMES is built on six KGs derived from a digitized mid-20th-century marine biology text [5] using different LLMs under a shared prompt configuration [2], and supports natural-language querying via four answer-generation models. Its central contribution is a *post-generation neuro-symbolic grounding verifier* that, given a response R and the retrieved subgraph G_q, uses the same LLM in an auditing role to decompose R into atomic claims and classify each as SUPPORTED, PARTIAL, or UNSUPPORTED against the triples of G_q. From these classifications, we derive two novel metrics — GF_{strict} and GF_{weighted} — that quantify the degree to which responses are traceable to graph evidence. Unlike RAGAs [3] and ARES [4], which evaluate faithfulness with respect to retrieved textual contexts, GF evaluates generated claims against explicit graph triples and distinguishes among SUPPORTED, PARTIAL, and UNSUPPORTED claims.

We evaluate the system through a full factorial design across 10 competency questions, 6 knowledge graphs, and 4 answer-generation LLMs (240 evaluations), reporting grounding fidelity across all graph–model combinations and identifying systematic effects of both construction LLM and answer LLM on downstream response grounding.

The contributions of this paper are:

1. **HERMES**: an open, web-based platform for multi-graph, multi-LLM marine knowledge graph querying and evaluation, including interactive graph visualization and a structured interface that exposes the full evidence chain from query to response.
2. **Grounding Fidelity metrics:** two novel metrics — GF_{strict} and GF_{weighted} — that quantify the degree to which claims in a generated response are traceable to triples in the retrieved subgraph, formally defined and empirically evaluated across 240 graph–LLM configurations.
3. **A factorial evaluation**: a controlled $N \times M \times J$ evaluation across competency questions, knowledge graphs, and answer-generation LLMs, characterizing graph and model effects on grounding fidelity in a domain-specific historical corpus.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 describes the HERMES system. Section 4 presents the evaluation methodology. Section 5 reports results. Section 6 discusses findings and limitations. Section 7 concludes.

2. Related Work

Knowledge graph construction from text. The automatic construction of KGs from unstructured text has received increasing attention with the rise of LLMs [6, 7]. Traditional approaches relied on predefined ontologies or manually curated schemas [8], limiting their applicability to domains where such resources exist. Microsoft GraphRAG [1] introduced a schema-free pipeline that combines entity and relation extraction with community detection and hierarchical summarization, enabling KG construction from arbitrary corpora without ontological alignment. Prior work in the marine domain demonstrated the feasibility of applying GraphRAG to historical scientific texts and showed that construction LLM choice significantly affects graph structure and downstream QA performance [2].

KG evaluation without gold standards. Evaluating automatically constructed KGs is challenging when no reference graph exists. Seo et al. [9] propose a taxonomy of six evaluation dimensions — including structure-based, data quality, and task-based evaluation — applicable when gold standards are absent. Competency questions, originally introduced for ontology evaluation [10, 11], have been adapted as an extrinsic evaluation strategy for KGs derived from specialized corpora [12], providing a user-centered assessment of whether the graph supports the knowledge needs of domain experts.

[1] https://hermes-project.streamlit.app

Faithfulness and grounding in RAG systems. Retrieval-Augmented Generation (RAG) grounds LLM responses in externally retrieved evidence, reducing hallucinations compared to purely parametric generation [13]. As RAG systems have matured, evaluation has shifted beyond answer correctness toward assessing whether generated responses actually rely on the retrieved context.

RAGAs [3] introduced one of the first automated evaluation frameworks for this purpose by decomposing generated responses into atomic claims and measuring the proportion of claims supported by the retrieved textual context. More recently, ARES [4] extended automated RAG evaluation by training lightweight judges for context relevance, answer faithfulness, and answer relevance, reducing dependence on large proprietary LLMs while requiring only a small amount of human-annotated data for calibration.

Recent GraphRAG surveys [7] highlight that, although graph-based retrieval replaces document chunks with structured knowledge, evaluation methodologies remain largely inherited from traditional text-based RAG. Existing frameworks evaluate grounding with respect to retrieved textual passages, whereas graph triples introduce additional structural semantics that are not explicitly considered by current faithfulness metrics.

The GF metrics introduced in this paper extend the claim-based evaluation paradigm of RAGAs to GraphRAG settings. Rather than verifying claims against retrieved text chunks, GF evaluates generated responses directly against retrieved subject–predicate–object triples and introduces a three-way classification scheme — Supported, Partial, Unsupported —, providing finer-grained diagnostic information while preserving explicit traceability between generated claims and graph evidence.

Neuro-symbolic AI and knowledge grounding. The integration of neural and symbolic components for grounded inference is a central concern of neuro-symbolic AI [14]. Recent surveys identify knowledge graphs as a natural symbolic substrate for neuro-symbolic integration, particularly when combined with neural retrieval and language models [15]. The HERMES grounding verifier instantiates this paradigm at the response level: neural claims are verified against symbolic triple structures, combining the generative capacity of LLMs with the explicit, auditable evidence of graph triples.

3. System Description: HERMES

HERMES is a web-based platform for natural-language querying and comparative evaluation of marine knowledge graphs. The system is built around three tightly coupled components: (i) a collection of six knowledge graphs constructed from the same historical corpus using different large language models; (ii) an embedding-based graph retrieval pipeline that selects query-relevant subgraphs at inference time; and (iii) a post-generation neuro-symbolic grounding verifier that audits each response against the retrieved evidence. The following subsections describe each component in detail.

3.1. Knowledge Graph Collection

The six knowledge graphs in HERMES were all derived from the same source: a digitized version of *Lobos marinos, pingüinos y guaneras de las costas del litoral marítimo e islas adyacentes de la República Argentina* [5], a foundational mid-20^{th}-century survey of marine vertebrate fauna along the Argentine Patagonian coast. The digitization pipeline—comprising high-resolution scanning, OCR processing, manual correction, and text segmentation—is described in full in our prior work [2].

Each graph was constructed independently using Microsoft GraphRAG with a fixed prompt configuration derived from the autotuning and prompt-engineering methodology introduced in [2], which yields concise, RDF-style predicates from the raw LLM extractions. The only variable across runs was the underlying language model. Table 1 lists the six configurations.

Because every graph is built from the same corpus with the same prompt configuration, differences in graph structure and downstream retrieval quality are attributable exclusively to the choice of construction LLM—a property that makes HERMES a controlled testbed for comparative evaluation. Ta-

Table 1
Knowledge graphs available in HERMES. Each graph was generated from the Carrara corpus using Microsoft GraphRAG with the indicated LLM and a shared prompt configuration.

Graph ID	LLM used for KG construction
KG_1	GPT-4o-mini [16]
KG_2	Mixtral 8×7B Instruct [17]
KG_3	Mistral 7B Instruct [18]
KG_4	Ministral 14B [19]
KG_5	Mistral Large 3 [20]
KG_6	LLaMA 3 8B Instruct [21]

ble 2 reports the structural properties of all six graphs (node and edge counts, density, average degree, connected components, and diameter).

Table 2
Structural properties of the six knowledge graphs in HERMES. Each graph was constructed from the same corpus using the indicated LLM.

Graph	Nodes	Edges	Density	Avg. Degree	Components	Diameter
KG_1 (GPT-4o-mini)	631	840	0.0042	2.66	90	12
KG_2 (Mixtral 8×7B)	1007	875	0.0017	1.74	234	16
KG_3 (Mistral 7B)	218	68	0.0029	0.62	167	11
KG_4 (Ministral 14B)	1669	2406	0.0017	2.88	264	10
KG_5 (Mistral Large 3)	906	1660	0.0040	3.66	39	9
KG_6 (LLaMA 3 8B)	129	131	0.0159	2.03	7	13

3.2. Node Embedding and Indexing

To support semantic retrieval at query time, each graph is indexed offline as a set of node-level vector embeddings. For every node v in a graph G, its label is encoded using OpenAI's `text-embedding-3-small` model, producing a dense vector $\mathbf{e}_v$. Embeddings are computed once per graph upon first access and persisted to disk; they are never recomputed or modified thereafter.

Only node labels are embedded in this initial implementation, as the most direct textual entity representation. Descriptions, predicates, neighborhood context, and degree are excluded. This unevaluated design is not claimed to be optimal; richer representations are future work.

Crucially, the same embedding model is used for all six graphs. This design choice ensures that any observed differences in retrieval quality across graph configurations reflect differences in graph structure and content, not differences in the embedding space.

3.3. Query-Time Retrieval Pipeline

At inference time, a user query q is processed through the following steps.

Step 1 — Query embedding. The query is encoded with the same `text-embedding-3-small` model used during indexing, producing a query vector $\mathbf{e}_q$ in the same embedding space.

Step 2 — Top-k node selection. Cosine similarity is computed between $\mathbf{e}_q$ and each node embedding $\mathbf{e}_v$:

$$\mathrm{sim}(q, v) = \frac{\mathbf{e}_q \cdot \mathbf{e}_v}{\|\mathbf{e}_q\| \cdot \|\mathbf{e}_v\| + \varepsilon}$$

where $\varepsilon > 0$ is a small constant added to the denominator for numerical stability. Nodes are ranked by similarity and selected greedily: all nodes with $\mathrm{sim}(q, v) \geq \tau$ are included, with a minimum of $k = 10$

nodes guaranteed regardless of threshold. The threshold $\tau = 0.55$ was selected through a sensitivity analysis described in Section 4.

Step 3 — Structural expansion. The selected seed nodes are expanded by traversing the graph topology. Let h denote the expansion depth. HERMES uses an initial, unvalidated query-length heuristic: h=1 below 80 characters, h=2 from 80 to 199, and h=3 from 200 onward. It assumes that longer queries may mention more entities or relations and benefit from broader context, but was not empirically optimized as a proxy for reasoning complexity. All ten evaluation questions used h=2; hence, the experiment does not compare traversal depths.

HERMES jointly expands the complete seed set for h iterations. Because the GraphML files are loaded as undirected graphs, every adjacent node is considered. Sets remove duplicates caused by cycles or converging paths. Relations are treated uniformly, with no degree-based pruning or neighbor cap; thus, all neighbors of a reached hub are included. The retrieved context G_q is the induced subgraph over the accumulated nodes. This expansion captures multi-hop relational context that would be invisible to flat document-chunk retrieval.

Step 4 — Triple serialization. All edges in G_q are serialized as `Subject --relation--> Object` triples and concatenated into a context string passed to the answer-generation model.

Graph structure contributes during retrieval: label similarity finds the seeds, while topology selects neighboring entities and relations through multi-hop expansion. The induced subgraph preserves relations among accumulated nodes, organizing the evidence before its triple serialization.

3.4. Answer Generation

The context string derived from G_q, together with the original query q, is submitted to the selected answer-generation LLM. HERMES currently supports four models for this role: GPT-4o-mini (OpenAI) [16], Llama 3.3 70B [22] and Llama 3.1 8B [23] (via Groq), and Gemini 2.5 Flash [24]. The choice of answer-generation LLM is independent of the graph construction LLM, enabling a full $M \times J$ factorial evaluation across M=6 graphs and J=4 answer models (Section 4).

This design follows the graph-based RAG paradigm introduced in [1], where retrieved graph context replaces or augments parametric knowledge during generation.

HERMES uses graph structure for retrieval, not formal graph reasoning. Once retrieved, G_q is serialized as textual triples for the LLM; the system executes no path inference, logical rules, ontological reasoning, or symbolic entailment. Answers are therefore grounded in graph-derived evidence but are not produced by an inference procedure over the graph.

3.5. Neuro-Symbolic Grounding Verifier

A key limitation of embedding-based retrieval systems is the absence of a mechanism to verify that the generated response actually respects the evidence it was given. To address this, HERMES incorporates a lightweight *post-generation neuro-symbolic grounding verifier*.

The component is neuro-symbolic in an architectural grounding sense: the *neural* element is the LLM that generates and audits claims, and the *symbolic* element is the explicit KG structure encoded by the triples in G_q. The LLM relates claims to this inspectable evidence; HERMES performs no symbolic inference, rule execution, or ontology-based entailment [14].

Design. Given a response R and the subgraph G_q that produced it, the verifier uses the same answer-generation LLM in a second, auditing role. The LLM is prompted with a strict system instruction that (i) prohibits the use of any external knowledge and (ii) requires conservative classification: a claim is SUPPORTED only when direct, traceable evidence exists in the triples. The verifier decomposes R into a set of atomic factual claims $\{c_1, \ldots, c_n\}$ and assigns each one a status $s_i \in \{\text{SUPPORTED}, \text{PARTIAL}, \text{UNSUPPORTED}\}$, along with the specific triples that serve as evidence (or

an empty list when none exist). Output is constrained to structured JSON to facilitate deterministic parsing.

Illustrative example. For CQ1, submitted in Spanish to KG_1 with Llama 3.3 70B, G_q contained the triples `OBSERVACIONES COPROLÓGICAS --son estudiadas en--> LOBOS MARINOS DE UN PELO` and `OBSERVACIONES COPROLÓGICAS --son estudiadas en--> LOBOS MARINOS DE DOS PELOS`. The generated claim states that both species were studied along the Argentine coast and adjacent islands. It is PARTIAL: the triples support the studies but not that geographical scope, contributing 0 to GF_{strict} and 0.5 to GF_{weighted}.

Grounding Fidelity metrics. From the claim-level annotations we derive two scalar metrics. *Strict Grounding Fidelity* counts only fully supported claims:

$$GF_{\text{strict}}(R, G_q) = \frac{|\{c_i : s_i = \text{SUPPORTED}\}|}{n}$$

Weighted Grounding Fidelity assigns partial credit to partially supported claims:

$$GF_{\text{weighted}}(R, G_q) = \frac{|\{c_i : s_i = \text{SUPPORTED}\}| + 0.5 \cdot |\{c_i : s_i = \text{PARTIAL}\}|}{n}$$

Both metrics take values in $[0, 1]$ for $n \geq 1$, and are defined as 0 when no claims are extracted ($n = 0$), where 1.0 indicates that every claim in the response is fully traceable to triples in the retrieved subgraph.

These metrics are conceptually related to the *faithfulness* score in RAGAs, which also measures the proportion of claims in a generated answer that are supported by retrieved context. However, GF differs in two important respects: (i) the retrieval unit is a graph triple (subject–predicate–object), not a text chunk, and (ii) the three-way classification (SUPPORTED/PARTIAL/UNSUPPORTED) provides more diagnostic information than the binary supported/not-supported scheme used in RAGAs. This work presents an innovative faithfulness metric formally defined and evaluated over GraphRAG-style triple contexts — to our knowledge, the first of its kind in this setting.

Beyond evaluation, the verifier contributes to the explainability of the system: by decomposing responses into atomic claims and citing specific triples as evidence, HERMES makes the reasoning chain from query to response fully auditable. This transparency-by-design principle is particularly relevant in scientific domains where response provenance is as important as response correctness.

3.6. User Interface and Transparency Layer

HERMES exposes the full evidence chain from query to grounded response through a structured, step-by-step interface organized into two phases: graph selection and question answering.

Phase 1 — Graph selection and exploration. Upon loading the application, the user selects a knowledge graph and an answer-generation LLM from the side panel and presses *Load Graph*. The system then displays two components. First, an *interactive graph visualization* renders the selected KG directly in the browser, allowing the user to zoom, drag nodes, and inspect individual nodes and edges by clicking on them to display their metadata. Second, a *structural metrics panel* reports the topological properties of the selected graph — nodes, edges, density, average degree, connected components, and diameter — giving the user immediate quantitative context about the graph they are querying. A natural-language text input is presented below these components, ready to receive a query.

Phase 2 — Question answering and grounding verification. After the user submits a query, the system executes the full retrieval and generation pipeline (Section 3.3 and 3.4) and displays four consecutive panels.

Answer panel. The natural-language response R generated by the selected LLM, grounded in the retrieved subgraph G_q.

Knowledge Graph Triplets panel. A tabular view of all SPO triples in G_q, displayed in Subject / Predicate / Object columns. This panel makes the retrieval step fully inspectable: the user can see exactly which triples were passed as context to the LLM, providing direct evidence of what information was available during generation.

Neuro-Symbolic Grounding Verification panel. An aggregated metrics summary displaying total claims, supported, partial, and unsupported counts, together with both GF_{strict} and GF_{weighted} scores. Below the summary, a claim-level table lists each atomic claim extracted from R, its verification status (SUPPORTED, PARTIAL, or UNSUPPORTED), and the specific triples cited as evidence.

Grounding Summary panel. A brief natural-language explanation generated by the verifier, summarizing the overall grounding quality of the response and the reasoning behind the claim-level classifications. This panel makes the verification outcome interpretable to non-technical users who may not wish to inspect the claim-level table directly.

Together, these panels implement a layered transparency architecture: the Triplets panel exposes *what* evidence was retrieved, the Verification panel quantifies *how well* the response is grounded in that evidence, and the Grounding Summary explains *why* each claim received its classification — making every step from query to grounded response visible and auditable.

4. Evaluation Methodology

Our evaluation framework assesses the degree to which generated responses are grounded in the retrieved subgraph evidence across all combinations of knowledge graph and answer-generation LLM, using the Grounding Fidelity metrics introduced in Section 3.5. The following subsections describe the competency questions, parameter selection, evaluation matrix, and scoring scheme.

4.1. Competency Questions

Following established methodology for knowledge graph evaluation [10, 11], we assess each graph–LLM configuration using ten competency questions (CQs) that reflect the knowledge needs of a domain expert consulting the Carrara corpus. Ten competency questions were selected to balance domain coverage with evaluation tractability. The questions were formulated by marine science domain experts at CESIMAR-CONICET who are familiar with the Carrara corpus, ensuring that they reflect genuine knowledge needs rather than arbitrary retrieval tasks. The set spans the main knowledge types present in the corpus — taxonomic, morphological, geographic, procedural, temporal, and institutional — while keeping the total number of evaluations ($N \times M \times J = 240$) within a range that allows systematic analysis without being computationally prohibitive. Each question type is represented at least once, ensuring that grounding fidelity is assessed across structurally diverse retrieval scenarios rather than concentrated on a single knowledge category.

The CQs are presented in English for readability; all experiments were conducted using Spanish-language queries to ensure semantic alignment with the node labels in the knowledge graphs, which are encoded in Spanish. This decision is motivated by a cross-lingual retrieval gap identified during parameter selection (Section 4.2): when queries are issued in English against Spanish-language graphs, cosine similarity scores remain consistently below any tested threshold, causing $k_{\min}$ to dominate node selection and rendering the threshold parameter ineffective. The full list of competency questions is provided in Appendix B.

4.2. Parameter Selection

Two retrieval parameters required empirical justification prior to the main evaluation: the cosine similarity threshold τ and the minimum node count $k_{\min}$.

We conducted a sensitivity analysis over two representative graphs (KG_1 and KG_2) and four competency questions (CQ1, CQ3, CQ7, CQ9), varying $\tau \in \{0.45, 0.50, 0.55, 0.60, 0.65\}$ with $k_{\min} = 10$ fixed, and measuring subgraph size and GF_{weighted} for each configuration. The analysis was run twice: once with queries in English and once with queries in Spanish.

The English-query run revealed the cross-lingual retrieval gap described above: across all threshold values, virtually no node exceeded the similarity threshold ($n_{\text{above}} \approx 0$), causing $k_{\min}$ to govern node selection in every case and producing identical subgraphs and GF scores regardless of τ. This confirmed that English queries are semantically misaligned with the Spanish node labels and that the evaluation must be conducted in Spanish.

The Spanish-query run showed meaningful threshold sensitivity for queries whose relevant nodes are densely represented in the graph. The most informative case was CQ7 on KG_1: at $\tau = 0.45$, the retriever selected 13 seed nodes, producing a subgraph of 308 nodes and $GF_{\text{weighted}} = 0.28$, indicative of retrieval noise from over-inclusion; at $\tau = 0.55$, seed nodes collapsed to 10 (governed by $k_{\min}$) and GF_{weighted} rose to 0.67. However, results were not uniformly monotone across graphs and queries: for KG_2/CQ7, GF_{weighted} was 0.875 at $\tau = 0.45$ but collapsed to 0.0 from $\tau = 0.50$ onward, suggesting that retrieval quality under threshold variation is graph- and query-dependent.

Based on these observations, $\tau = 0.55$ was selected as a conservative default. It avoids the retrieval noise observed at lower thresholds (over-inclusion of weakly similar nodes) while preventing $k_{\min}$ from dominating excessively at higher thresholds where relevant nodes may be filtered out. We do not claim $\tau = 0.55$ to be globally optimal; rather, it represents a reasonable operating point for the main evaluation. Full sensitivity results for both language conditions are available as supplementary material in the project repository (see Appendix A).

4.3. Evaluation Matrix

The main evaluation is structured as a full factorial design over three dimensions:

- $N = 10$ competency questions (CQ1–CQ10)
- $M = 6$ knowledge graphs (KG_1–KG_6, one per construction LLM)
- $J = 4$ answer-generation LLMs (GPT-4o-mini, Llama 3.3 70B, Llama 3.1 8B, Gemini 2.5 Flash)

This yields $N \times M \times J = 10 \times 6 \times 4 = 240$ individual evaluations. For each evaluation cell (q_n, KG_m, LLM_j), the system executes the full pipeline described in Section 3: query embedding, top-k node retrieval, structural expansion, answer generation, and post-generation grounding verification. All evaluations use Spanish-language queries and a fixed configuration of $\tau = 0.55$, $k_{\min} = 10$.

The factorial design enables analysis along three axes: (i) *graph effect* — does the choice of construction LLM affect downstream GF scores?; (ii) *model effect* — does the choice of answer-generation LLM affect grounding fidelity independently of the graph?; and (iii) *interaction effect* — do specific graph–LLM pairings produce GF scores that diverge significantly from what the individual graph and model effects would predict in isolation?

4.4. Scoring Scheme

Each evaluation cell is assessed on two Grounding Fidelity variants as defined in Section 3.5: GF_{strict}, which counts only fully supported claims, and GF_{weighted}, which assigns partial credit to partially supported claims. Both are computed automatically by the neuro-symbolic grounding verifier and require no manual annotation.

We report both variants throughout: GF_{strict} serves as a conservative lower bound, while GF_{weighted} is the primary metric. The latter is preferred because partial evidence is common in knowledge graphs derived from historical narratives, where explicit statements are often distributed across multiple triples rather than captured in a single SPO assertion.

Aggregate results are reported as mean GF_{strict} and mean GF_{weighted} per graph (KG_m), per answer-generation LLM (LLM_j), and per competency question (q_n), enabling fine-grained analysis of where grounding succeeds or fails across the evaluation matrix.

5. Results

We report results for 235 valid evaluations out of 240 total ($N \times M \times J = 10 \times 6 \times 4$). Five cells produced parse errors during grounding verification and are excluded from all aggregations: four involved `llama-3.1-8b-instant` as the answer-generation LLM (on KG_1/CQ7, KG_5/CQ8, KG_6/CQ2, and KG_6/CQ10), and one involved `llama-3.3-70b-versatile` (KG_5/CQ7). These failures are attributable to malformed JSON output from the verifier in specific graph–CQ combinations.

5.1. Overall Grounding Fidelity

Across all 235 valid evaluations, the mean GF_{weighted} is 0.50 and the mean GF_{strict} is 0.47, indicating that on average half of the atomic claims in generated responses are fully or partially traceable to triples in the retrieved subgraph. The gap between the two variants (0.03) is small, suggesting that partially supported claims are relatively rare and that most claims are either clearly supported or clearly unsupported.

Table 3 reports mean GF_{weighted} and GF_{strict} for every graph–LLM combination, together with marginal means per graph and per answer-generation LLM.

Table 3
Mean GF_{weighted} / GF_{strict} per knowledge graph and answer-generation LLM, averaged over 10 competency questions. Each cell reports GF_{weighted} / GF_{strict}. Bold indicates the highest GF_{weighted} per row.

Graph	GPT-4o-mini	Llama 3.3 70B	Llama 3.1 8B	Gemini 2.5 Flash	Mean
KG_1 (GPT-4o-mini)	0.29/0.29	**0.95/0.91**	0.69/0.60	0.74/0.73	0.67/0.63
KG_2 (Mixtral 8×7B)	0.10/0.10	0.56/0.52	**0.62/0.47**	0.57/0.56	0.46/0.41
KG_3 (Mistral 7B)	0.18/0.17	0.30/0.30	**0.57/0.51**	0.46/0.45	0.38/0.36
KG_4 (Ministral 14B)	0.10/0.10	**0.73/0.72**	0.71/0.62	0.54/0.53	0.52/0.49
KG_5 (Mistral Large 3)	0.08/0.08	0.75/0.72	0.66/0.51	**0.78/0.77**	0.57/0.52
KG_6 (LLaMA 3 8B)	0.25/0.25	0.46/0.46	**0.59/0.52**	0.45/0.44	0.44/0.42
Mean	0.17/0.17	0.62/0.60	0.64/0.54	0.59/0.58	0.50/0.47

5.2. Effect of the Answer-Generation LLM

The choice of answer-generation LLM has a strong and consistent effect on grounding fidelity across all graphs. Llama 3.1 8B and Llama 3.3 70B achieve the highest mean GF_{weighted} (0.64 and 0.62 respectively), followed by Gemini 2.5 Flash (0.59). GPT-4o-mini performs substantially below the other three, with a mean GF_{weighted} of only 0.17 — a gap of 0.45 points relative to the next-lowest model.

This result is counterintuitive given that GPT-4o-mini is generally considered a highly capable model. An inspection of the claim-level data reveals a likely explanation: GPT-4o-mini produces an average of only 1.6 atomic claims per evaluation, compared to 3.7 for Llama 3.3 70B, 6.6 for Llama 3.1 8B, and 6.4 for Gemini 2.5 Flash. This suggests that GPT-4o-mini tends to generate shorter, more cautious responses that frequently acknowledge the limits of the retrieved subgraph rather than committing to positive factual assertions. Such negative meta-claims interact with the verifier's handling of negative assertions — where a claim stating that the subgraph lacks information about a topic is classified as SUPPORTED if no relevant triples exist — and the resulting low claim count amplifies any individual unsupported claim's impact on the final GF score. This behavior is discussed further in Section 6.

5.3. Effect of the Knowledge Graph

Graph construction quality also affects downstream grounding fidelity, though the effect is smaller than the LLM effect and interacts with the choice of answer model. KG_1 (constructed with GPT-4o-mini) achieves the highest mean GF_{weighted} across answer-generation LLMs (0.67), followed by KG_5 (Mistral Large 3, 0.57) and KG_4 (Ministral 14B, 0.52). KG_3 (Mistral 7B) performs worst (0.38), consistent with the lower structural quality reported for that graph in prior structural analyses.

These results suggest that graphs constructed by more capable models provide denser, more semantically coherent triple sets that are easier for the verifier to match against generated claims. KG_1 in particular benefits from the same model that produces the highest-quality triples in our corpus, consistent with findings in Section 3.1.

5.4. Interaction Effects

The best single graph–LLM combination is KG_1 × Llama 3.3 70B, achieving $GF_{\text{weighted}} = 0.95$ and $GF_{\text{strict}} = 0.91$ — near-perfect grounding across 10 competency questions. The second and third best combinations are KG_5 × Gemini 2.5 Flash (0.78) and KG_5 × Llama 3.3 70B (0.75), both involving KG_5.

The worst combinations all involve GPT-4o-mini as the answer-generation LLM: KG_5 × GPT-4o-mini (0.08), KG_2 × GPT-4o-mini and KG_4 × GPT-4o-mini (both 0.10).

A notable interaction pattern is that KG_3 (Mistral 7B) shows the lowest mean GF across all answer models, but Llama 3.1 8B achieves a relatively competitive 0.57 on this graph — suggesting that certain answer models may be more robust to lower-quality graph structures.

5.5. Results by Competency Question

Table 4 reports mean GF_{weighted} and GF_{strict} per competency question, averaged over all graph–LLM combinations.

Table 4
Mean Grounding Fidelity per competency question, averaged over all 24 graph–LLM combinations, sorted by GF_{weighted}.

CQ	GF_{weighted}	GF_{strict}
CQ1	0.737	0.681
CQ6	0.640	0.628
CQ8	0.534	0.524
CQ4	0.512	0.484
CQ7	0.505	0.468
CQ5	0.484	0.454
CQ3	0.468	0.434
CQ9	0.390	0.345
CQ10	0.389	0.353
CQ2	0.356	0.328

CQ1 (pinniped species studied) achieves the highest mean grounding fidelity ($GF_{\text{weighted}} = 0.74$), likely because species names and their locations are densely represented as explicit triples across all graphs. CQ6 (sexual dimorphism in *Otaria flavescens*) also scores well (0.64), reflecting the detailed morphological descriptions present in the Carrara corpus.

The lowest-scoring questions are CQ2 (years of official investigations, 0.36), CQ9 (locomotion of *Mirounga leonina*, 0.39), and CQ10 (role of aerial photography, 0.39). These questions target procedural, temporal, or contextual information that may be distributed across the corpus in ways that GraphRAG's entity-relation extraction does not consistently capture as explicit triples, resulting in subgraphs with limited direct evidence for the verifier to match against.

6. Discussion

6.1. Graph Construction Quality and Downstream Grounding

The results in Section 5 reveal a clear relationship between graph construction quality and downstream grounding fidelity. KG_1 (GPT-4o-mini) consistently achieves the highest mean GF_{weighted} across answer-generation LLMs, which aligns with its structural properties: 840 edges, an average degree of 2.66, and only 90 connected components — the best balance of connectivity and cohesion

among the six graphs. In contrast, KG_3 (Mistral 7B), with only 68 edges and 167 components, produces the lowest GF scores, consistent with its sparse and fragmented structure limiting the availability of direct triple evidence for the verifier.

This pattern extends findings from prior structural analysis [2]: models that generate more connected, semantically coherent graphs also produce better downstream retrieval and grounding outcomes, suggesting that graph construction quality propagates through the full pipeline to the final response. Notably, KG_5 (Mistral Large 3) — the second-best graph by mean GF — combines a high edge count (1660) with low fragmentation (39 components), reinforcing the importance of both density and cohesion in graph construction.

6.2. The GPT-4o-mini Anomaly

The most striking finding is the consistently low GF_{weighted} of GPT-4o-mini as an answer-generation LLM (0.17 mean, compared to 0.59–0.64 for the other three models). As discussed in Section 5.2, the claim-level data points to a behavioral explanation: GPT-4o-mini produces an average of only 1.6 atomic claims per evaluation — less than half the next-lowest model — suggesting it generates shorter, more cautious responses that frequently acknowledge the limits of the retrieved subgraph rather than committing to positive factual assertions.

This behavior interacts with the verifier's handling of negative assertions: when the Q&A LLM states that the subgraph does not contain sufficient information on a topic, the verifier evaluates this claim solely against the retrieved triples — its operational scope — and classifies it as SUPPORTED if no relevant triples exist. How to handle such cases optimally remains an open question that we identify as a direction for future work. However, models that generate such meta-claims produce fewer positive assertions overall, and any individual unsupported claim has a disproportionate impact on the final GF score when the total claim count is low.

This reveals a non-trivial design tension in LLM-based grounding evaluation: models that accurately acknowledge the limits of retrieved evidence may score lower than models that generate more (potentially unsupported) positive claims. Resolving this tension requires a more expressive claim taxonomy — for instance, distinguishing between *epistemic* claims (about what the graph contains) and *factual* claims (about domain content) — which we identify as a promising direction for future work.

6.3. Limitations

Self-consistency bias. The verifier uses the same LLM that generated the response, introducing a potential self-consistency bias: a model may rate its own claims more favorably than an independent judge [25]. The strict system prompt and the requirement to cite specific triples mitigate but do not eliminate this bias. GF scores have not yet been validated against independent human judgments. A natural extension is to use a fixed, independent LLM as verifier across all evaluation cells, enabling a controlled comparison of verifier behavior.

Single corpus. All six knowledge graphs are derived from the same historical source (Carrara, 1952). While this enables a controlled comparison of construction LLMs, it limits the generalizability of GF scores to other corpora, domains, or languages. Extending the evaluation to additional marine science texts and to graphs derived from structured observational data (e.g., OBIS occurrence records) is a priority for future work. A natural next step is to perform a staged evaluation by first incorporating additional historical documents from the marine science domain, followed by contemporary biodiversity datasets and multilingual corpora. This would allow assessing the robustness of the proposed metrics across different writing styles, knowledge densities, and graph topologies, while determining whether the observed behavior of GF generalizes beyond the specific corpus considered in this study.

Cross-lingual retrieval gap. Node labels in all six graphs are in Spanish, while the system is designed for multilingual use. The sensitivity analysis (Section 4.2) demonstrated that English queries

produce near-zero cosine similarities against Spanish node labels, effectively disabling threshold-based filtering. Addressing this limitation through multilingual sentence embedding models capable of aligning semantically equivalent concepts across languages (e.g., LaBSE [26]) or through cross-lingual knowledge graph alignment techniques (e.g., BootEA [27]) represents an important direction for future work. Such approaches could improve cross-lingual retrieval while preserving the original graph structure and avoiding manual translation of node labels.

Parse errors. Five out of 240 evaluations produced malformed JSON output from the verifier, all involving Llama 3.1 8B and Llama 3.3 70B on specific graph–CQ combinations. These failures suggest that JSON output compliance under strict formatting constraints varies across models regardless of parameter count, and that robust output parsing and fallback verification strategies should be incorporated in future versions of the pipeline. For the five affected cells, grounding fidelity could not be computed automatically; manual annotation by a domain expert would allow these cases to be scored and would additionally serve as a validation mechanism for the automatic GF scores across the full evaluation matrix — an assessment we identify as a priority for future work.

7. Conclusions and Future Work

This paper presented HERMES, a web-based platform for natural-language querying and comparative evaluation of marine knowledge graphs built with Microsoft GraphRAG. The system introduces a post-generation neuro-symbolic grounding verifier that audits generated responses against retrieved subgraph triples, and two novel metrics — GF_{strict} and GF_{weighted} — that quantify the degree to which response claims are traceable to graph evidence.

A full factorial evaluation across N=10 competency questions, M=6 knowledge graphs, and J=4 answer-generation LLMs (240 individual evaluations) produced three main findings. First, the choice of answer-generation LLM has a stronger and more consistent effect on grounding fidelity than the choice of construction LLM: Llama 3.1 8B and Llama 3.3 70B achieved mean GF_{weighted} of 0.64 and 0.62 respectively, while GPT-4o-mini scored only 0.17, a result attributed to its tendency to generate cautious, low-claim responses rather than positive factual assertions. Second, mean GF varied across graph configurations: KG_1 and KG_5 achieved the highest graph-level means, whereas the substantially sparser and more fragmented KG_3 obtained the lowest mean GF. Third, the best single combination — $KG_1\times$ Llama 3.3 70B — achieved near-perfect grounding ($GF_{\text{weighted}} = 0.95$), demonstrating that high-fidelity graph-grounded generation is attainable in domain-specific historical corpora.

Future work will pursue five directions. First, applying a fixed, independent LLM as verifier across all evaluation cells to eliminate self-consistency bias and enable controlled comparisons of verifier behavior. Second, extending the evaluation to additional marine science corpora — including structured observational data from OBIS and Argo — to assess the generalizability of GF beyond historical texts. Third, investigating multilingual embedding strategies to address the cross-lingual retrieval gap identified between English queries and Spanish graph node labels. Fourth, developing a more expressive claim taxonomy that distinguishes epistemic claims (about graph content) from factual claims (about domain content), enabling fairer evaluation of models that accurately acknowledge retrieval limits. Fifth, conducting expert human annotation of a representative subset of evaluation cells to validate the automatic GF scores and establish inter-annotator agreement — providing an independent quality assessment of the neuro-symbolic grounding verifier and a benchmark for future iterations of the metric.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (GPT-4) and Claude in order to: Paraphrase and reword and Grammar and Spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, et al., From local to global: A graphrag approach to query-focused summarization, https://arxiv.org/pdf/2404.16130, 2024.

[2] G. Nuñez, M. Zárate, D. Ceballos, P. Fillottrani, Digitizing and structuring early marine biodiversity records: A GraphRAG-based methodology, in: Proceedings of the 1st GOBLIN Workshop on Knowledge Graph Technologies (GOBLIN25), Zenodo, Leipzig, Germany, 2025. URL: https://doi.org/10.5281/zenodo.16912811. doi:10.5281/zenodo.16912811.

[3] S. Es, J. James, L. Espinosa Anke, S. Schockaert, RAGAs: Automated evaluation of retrieval augmented generation, in: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 150–158. URL: https://aclanthology.org/2024.eacl-demo.16.

[4] J. Saad-Falcon, O. Khattab, C. Potts, M. Zaharia, Ares: An automated evaluation framework for retrieval-augmented generation systems, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 338–354.

[5] I. S. Carrara, Lobos marinos, pingüinos y guaneras de las costas del litoral marítimo e islas adyacentes de la república argentina, in: Lobos marinos, pingüinos y guaneras de las costas del litoral marítimo e Islas adyacentes de la República Argentina, 1952, pp. 198–198.

[6] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering 36 (2024) 3580–3599.

[7] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, S. Tang, Graph retrieval-augmented generation: A survey, ACM Transactions on Information Systems 44 (2025) 1–52.

[8] N. Mihindukulasooriya, S. Tiwari, C. F. Enguix, K. Lata, Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text, in: International semantic web conference, Springer, 2023, pp. 247–265.

[9] S. Seo, H. Cheon, H. Kim, D. Hyun, Structural quality metrics to evaluate knowledge graph quality, Semantic Web (2022).

[10] M. Gruninger, Methodology for the design and evaluation of ontologies, in: Proc. IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing, 1995.

[11] M. Uschold, M. Gruninger, Ontologies: Principles, methods and applications, The knowledge engineering review 11 (1996) 93–136.

[12] J. Kim, Y. Kwon, Y. Jo, E. Choi, Kg-gpt: A general framework for reasoning on knowledge graphs using large language models, arXiv preprint arXiv:2310.11220 (2023).

[13] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, H. Wang, et al., Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 2 (2023) 32.

[14] A. d. Garcez, L. C. Lamb, Neurosymbolic AI: The 3rd wave, Artificial Intelligence Review 56 (2023) 12387–12406.

[15] G. Marra, S. Dumančić, R. Manhaeve, L. De Raedt, From statistical relational to neurosymbolic artificial intelligence: A survey, Artificial Intelligence 328 (2024) 104062.

[16] OpenAI, GPT-4o mini, Large language model accessed via OpenAI API, 2024. URL: https://platform.openai.com/docs/models/gpt-4o-mini, accessed: 2026-06-08.

[17] M. AI, Mixtral 8x7b instruct, https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1, 2023. Accessed: 2025-07-15.

[18] M. AI, Mistral 7b instruct, https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1, 2023. Accessed: 2025-07-30.

[19] Mistral AI, Ministral 3-14b, Large language model accessed via Mistral AI API, 2025. URL: https://docs.mistral.ai/models/model-cards/ministral-3-14b-25-12, model: ministral-14b-2412. Accessed: 2026-05-30.

[20] Mistral AI, Mistral large 3, Large language model accessed via Mistral AI API, 2025. URL: https://docs.mistral.ai/models/model-cards/mistral-large-3-25-12, model: mistral-large-2512. Accessed:

2026-02-20.
[21] Meta AI, Llama 3 8b instruct, https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct, 2024. Accessed: 2025-07-15.
[22] Meta AI, Llama 3.3 70b instruct, 2024. URL: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct, accessed: 2026-06-28.
[23] Meta AI, Llama 3.1 8b instruct, 2024. URL: https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, accessed: 2026-06-28. Deprecated by Groq as of August 2026; experiments conducted prior to deprecation.
[24] Google DeepMind, Gemini 2.5 flash, 2025. URL: https://deepmind.google/technologies/gemini/, accessed: 2026-06-08.
[25] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-bench and Chatbot Arena, in: Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 46595–46623. URL: https://arxiv.org/abs/2306.05685.
[26] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022, pp. 878–891.
[27] Z. Sun, W. Hu, Q. Zhang, Y. Qu, Bootstrapping entity alignment with knowledge graph embedding., in: Ijcai, volume 18, 2018.

A. Online Resources

The following resources support reproducibility of the experiments reported in this paper:

- **HERMES platform** (live system): https://hermes-project.streamlit.app
- **Source code, evaluation scripts, and supplementary material** (including sensitivity analysis results and evaluation CSV files): https://github.com/gustavomarcelonunez/hermes

B. Competency Questions

CQ1. What species of pinnipeds were studied along the Argentine maritime coast and adjacent islands?

CQ2. During which years were the official investigations on pinnipeds, penguins, and seabird guano conducted?

CQ3. Which governmental body endorsed and facilitated the investigations described in the text?

CQ4. What were the main methods and means of transportation used during the different research commissions?

CQ5. How are southern sea lions (*Otaria flavescens*) morphologically and behaviorally characterized?

CQ6. What are the main physical differences between male and female *Otaria flavescens*?

CQ7. Where were the colonies of two-haired sea lions (*Arctocephalus australis*) located according to the study?

CQ8. What reasons does the text give for the prohibition of hunting *Arctocephalus australis*?

CQ9. How is the locomotion of southern elephant seals (*Mirounga leonina*) described in the document?

CQ10. What role did aerial photography and mapping play in the data collection of the investigations?