RDF-Based Structured Quality Assessment Representation of Multilingual LLM Evaluations

Jonas Gwozdz and Andreas Both

Leipzig University of Applied Sciences, Leipzig, Germany
{jonas.gwozdz,andreas.both}@htwk-leipzig.de

Abstract. Large Language Models (LLMs) increasingly serve as knowledge interfaces, yet systematically assessing their reliability with conflicting information remains difficult. We propose an RDF-based framework to assess multilingual LLM quality, focusing on knowledge conflicts. Our approach captures model responses across four distinct context conditions (complete, incomplete, conflicting, and no-context information) in German and English. This structured representation enables the comprehensive analysis of knowledge leakage-where models favor training data over provided context-error detection, and multilingual consistency. We demonstrate the framework through a fire safety domain experiment, revealing critical patterns in context prioritization and language-specific performance, and demonstrating that our vocabulary was sufficient to express every assessment facet encountered in the 28-question study.

1 Introduction

As interfaces to vast amounts of knowledge, Large Language Models (LLMs) are increasingly prevalent in information processing. However, their tendency to blend knowledge from training data with provided context poses challenges for reliability assessment, particularly in critical domains where factual accuracy is essential [2]. With incomplete or conflicting information, LLMs may prioritize either training data or provided context, impacting their reliability. Current evaluation approaches often lack standardized, structured representations of assessment results, hindering systematic analysis and comparison. This gap is especially pronounced in multilingual evaluations, where performance can differ across languages, and in understanding the interplay between context-based and training-based knowledge, which remains underexplored. Moreover, without adhering to FAIR principles (Findable, Accessible, Interoperable, Reusable), evaluation results are poorly materialized, limiting their usability and broader adoption in research and practice. We address these challenges by introducing an RDF-based framework for the structured representation of LLM quality assessments across models and languages. Our approach offers several key contributions: (1) a comprehensive RDF vocabulary for representing LLM evaluation results that aligns with FAIR principles to ensure findability, accessibility, interoperability, and reusability, (2) a systematic methodology for testing LLM
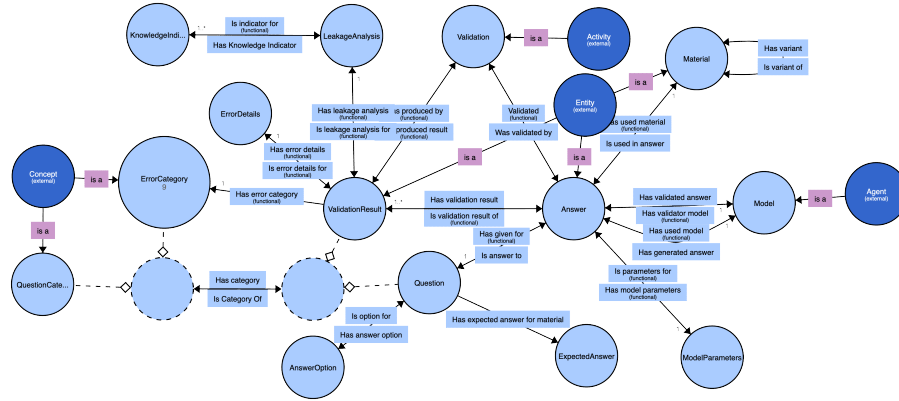
Fig. 1. Simplified visual representation of the RDF vocabulary for LLM evaluations.

responses across four distinct context conditions—complete, incomplete, conflicting, and no-context information—supporting multilingual evaluations with language-specific extensions to capture model- and language-specific behaviors, and (3) a demonstration validating our data model through a fire safety domain experiment with models like GPT-4o-mini[1] and Gemini-2.0-Flash[2] in German and English, exposing critical patterns in context prioritization and language-specific performance within the resulting dataset. Figure 1 illustrates the core structure of our RDF vocabulary, which underpins the evaluation framework.

2   Related Work

Recent studies explore LLMs with knowledge graphs and knowledge conflicts. Lavrinovics et al. [2] survey how knowledge graphs mitigate LLM hallucinations, while Kwan et al. [1] use them for factual accuracy. Xie et al. [4] and Tan et al. [3] highlight LLMs' bias toward context, even when incorrect. However, these efforts lack standardized vocabularies for systematic, multilingual evaluation. While valuable, current approaches often focus on narrow contexts or fail to provide structured, interoperable frameworks aligned with FAIR principles. Our work addresses this limitation by introducing an RDF-based representation for multilingual LLM assessments, enabling consistent, queryable analysis of knowledge conflicts and language-specific behaviors across varied context scenarios.

3   RDF Vocabulary for LLM Evaluation

Our RDF vocabulary links `:Question`, `:Answer`, `:ValidationResult`, and `:Material` to evaluate LLMs across languages and context conditions (complete,

---

[1] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[2] https://developers.googleblog.com/en/gemini-2-family-expands/

incomplete, conflicting, no-context). Multilinguality is supported via language-tagged literals for `:hasText`. Relationships such as `:hasGivenFor` (`:Question` to `:Answer`), `:hasUsedMaterial` (`:Answer` to `:Material`), and `:hasValidation-Result` (`:Answer` to `:ValidationResult`) enable SPARQL queries for analyzing knowledge leakage and cross-lingual consistency. Figure 1 illustrates this core structure with key relationships.

The T-Box defines 14 classes (e.g., `:Question`, `:Answer`) and 57 properties (e.g., `:isValid`, `:matchesFactual`), with OWL/SHACL constraints ensuring integrity. It aligns with FAIR principles using the PROV Ontology[3] and Dublin Core[4]. The full schema is available at `http://purl.org/sqare#`. The choice of RDF enables reasoning, linkage to external KGs, and federated SPARQL queries—advantages unattainable with flat CSV tables.

4   Use Case and Evaluation

We applied our approach to a fire safety domain experiment, showcasing its ability to structure LLM quality assessments. Fire safety was chosen due to its well-defined knowledge base and critical need for accuracy, making it ideal for testing the schema's capacity to capture context variations as well as multilingual responses. Beyond this domain, the framework generalizes to applications like educational content creation. By testing LLM responses against course materials, creators can pinpoint knowledge gaps and ensure content sufficiency, with the RDF structure tracking material-question alignments. Full experimental details are available in the online appendix (Git repository).

Experimental Setup The experiment tested LLM responses to 28 fire safety questions under four context conditions: (1) *Complete*: full, accurate information; (2) *Incomplete*: missing information; (3) *Conflicting*: factually contradicting information; (4) *No Context*: no supporting context. These conditions were chosen to assess how LLMs prioritize context versus training knowledge, revealing behaviors like context adherence or knowledge leakage. All prompts follow a *zero-shot, system-first* template detailed in the online appendix.

Data Collection and Analysis For each question and context condition, we collected responses from GPT-4o-mini and Gemini-2.0-Flash in both German and English, storing them in our RDF structure. Validation assessed correctness per fire safety standards and context expectations. The resulting RDF dataset captures anomalies and notable LLM behaviors, such as deviations from expected results, comprehensively reflecting the assessed LLMs' capabilities. SPARQL queries enabled the analysis of knowledge patterns, multilingual differences, and context reliance across models and languages.

---

[3] https://www.w3.org/TR/prov-o/
[4] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Table 1. Paired Statistical Comparison (German): Gemini-2.0-Flash vs. GPT-4o-mini

Context	Contingency (a,b;c,d)	McNemar $p^{\dagger}$	Δ-Acc (95 % CI) [pp]	Cohen's κ
Complete	(28, 0; 0, 0)	-	0.0 [0.0, 0.0]	- (κ undefined)
Incomplete	(10, 4; 8, 6)	0.3877	-14.3 [$-37.9, +9.4$]	0.143
Conflicting	(2, 0; 1,25)	-	-3.6 [$-10.4, +3.3$]	0.781
No Context	(24, 2; 2, 0)	-	0.0 [$-14.0, +14.0$]	-0.077

† Exact two-sided McNemar p-value; "–" indicates b + c < 5

Key Findings Tables 1 and 2 present the performance of GPT-4o-mini and Gemini-2.0-Flash in German and English under four context conditions: *Context Dominance*: All models strongly adhere to the given instructions and prioritize provided context, replicating incorrect information at rates of 89-93% rather than using their training knowledge; *Multilingual Differences*: English models handle incomplete information better, while German models demonstrate stronger baseline knowledge without context, revealing language-specific behaviors; *Educational Applications*: For course creators, these findings suggest focusing on information completeness in English materials while potentially leveraging German LLMs' stronger baseline knowledge for gap identification.

All such findings are reflected in the vocabulary. Hence, such findings can be generated using SPARQL queries[5], which validates our research task of representing the data from LLM assessments in a comprehensive and semantically rich form.

4.1   Paired Statistical Comparison of Models

To complement our RDF-based evaluation, we performed a paired analysis on the binary correctness labels (`is_valid`) for each question under each context (*complete*, *incomplete*, *conflicting*, *no_context*) and language (*de*, *en*). For each model pair (Gemini-2.0-Flash vs. GPT-4o-mini) we built the 2×2 contingency table

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} \text{both correct} & \text{Gemini correct, GPT wrong} \\ \text{Gemini wrong, GPT correct} & \text{both wrong} \end{bmatrix}$$

and computed:

1. **McNemar's exact test** (two-sided, no continuity correction) on the discordant cells b vs. c.
2. Δ-**accuracy** $= \text{Acc}_{\text{Gemini}} - \text{Acc}_{\text{GPT}}$, with 95 % Newcombe CI for paired proportions.
3. **Cohen's** κ as a measure of overall agreement beyond chance.

McNemar's test is non-significant ($p > 0.05$) in all German contexts, including *incomplete* ($p = 0.3877$). In other cases such as *no_context*, the number of discordant pairs was too low ($b + c < 5$) to permit meaningful testing, despite a noticeable accuracy gap of -14.3 percentage points in *incomplete*. In English,

---

[5] See our online appendix at `http://purl.org/sqare/repo#`.

Table 2. Paired Statistical Comparison (English): Gemini-2.0-Flash vs. GPT-4o-mini

Context	Contingency (a,b;c,d)	McNemar $p^{\dagger}$	Δ-Acc (95 % CI) [pp]	Cohen's κ
Complete	(28, 0; 0, 0)	-	0.0 [0.0, 0.0]	- (κ undefined)
Incomplete	(27, 1; 0, 0)	-	+3.6 [−3.3, +10.4]	0
Conflicting	(2, 1; 1,24)	-	0.0 [−9.9, +9.9]	0.627
No Context	(14, 0; 9, 5)	0.0039	−32.1 [−49.4, -14.8]	0.357

† Exact two-sided McNemar p-value; "–" indicates b + c < 5

only the *no_context* condition shows a significant discordance ($p = 0.0039$), with GPT-4o-mini outperforming Gemini by 32.1 pp. The Newcombe CIs reveal that many of these gaps are too wide to draw firm conclusions (e.g., German *incomplete* CI spans ± 25 pp), while Cohen's κ highlights very high agreement in error replication ($\kappa = 0.781$ de, 0.627 en) versus low agreement under manipulated prompts (e.g., $\kappa = 0.143$ de, undefined in en).

5   Conclusion and Future Work

Our RDF vocabulary (contribution 1) and fire safety experiment (contribution 2) assess LLM reliability across languages and contexts, focusing on knowledge conflicts critical for real-world use. Experiments with GPT-4o-mini and Gemini-2.0-Flash in German and English reveal how models handle contradictory information and language-specific differences, emphasizing the need for structured knowledge. Findings show that LLMs favor training data over context in 7-11% of conflicting cases, but mostly adhere to the given context, even when incorrect.

This semantically rich representation ensures reliability in critical systems, supporting standardized, queryable analysis for transparency and reproducibility. It advances structured LLM evaluation methodologies, tackling reliability challenges with conflicting information across languages. Future work will scale the question set, add evaluations for low-resource languages, and refine leakage metrics, as well as extend this framework to new domains, improve knowledge leakage detection, and define reliability metrics.

References

1. Kwan, L., Omran, P.G., Taylor, K.L.: Using knowledge graphs and agentic llms for factuality text assessment and improvement. In: International Workshop on the Semantic Web (2024), `https://api.semanticscholar.org/CorpusID:274281581`
2. Lavrinovics, E., Biswas, R., Bjerva, J., Hose, K.: Knowledge graphs, large language models, and hallucinations: An nlp perspective (2024). `https://doi.org/10.48550/arXiv.2411.14258`, `https://arxiv.org/abs/2411.14258`
3. Tan, H., Sun, F., Yang, W., Wang, Y., Cao, Q., Cheng, X.: Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? (2024), `https://arxiv.org/abs/2401.11911`
4. Xie, J., Zhang, K., Chen, J., Lou, R., Su, Y.: Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts (2024), `https://arxiv.org/abs/2305.13300`