Leveraging Large Language Models for Semantic Evaluation of RDF Triples

André Gomes Regino[1][a], Fernando Rezende Zagatti[1,4][b], Rodrigo Bonacin[1][c], Victor Jesus Sotelo Chico[2][d], Victor Hochgreb[2][e] and Julio Cesar dos Reis[3][f]

[1]*Center for Information Technology Renato Archer, Campinas, São Paulo, Brazil*

[2]*GoBots, Campinas, São Paulo, Brazil*

[3]*Institute of Computing, University of Campinas, Campinas, São Paulo, Brazil*

[4]*Department of Computing, UFScar, São Carlos, Brazil*

Keywords: LLM as a Judge, RDF Triple Generation, RDF Triple Validation.

Abstract: Knowledge Graphs (KGs) depend on accurate RDF triples, making the quality assurance of these triples a significant challenge. Large Language Models (LLMs) can serve as graders for RDF data, providing scalable alternatives to human validation. This study evaluates the feasibility of utilizing LLMs to assess the quality of RDF triples derived from natural language sentences in the e-commerce sector. We analyze 12 LLM configurations by comparing their Likert-scale ratings of triple quality with human evaluations, focusing on both complete triples and their individual components (subject, predicate, object). We employ statistical correlation measures (Spearman and Kendall Tau) to quantify the alignment between LLM and expert assessments. Our study examines whether justifications generated by LLMs can indicate higher-quality grading. Our findings reveal that some LLMs demonstrate moderate agreement with human annotators and none achieve full alignment. This study presents a replicable evaluation framework and emphasizes the current limitations and potential of LLMs as semantic validators. These results support efforts to incorporate LLM-based validation into KG construction processes and suggest avenues for prompt engineering and hybrid human-AI validation systems.

1 INTRODUCTION

Knowledge Graphs (KGs) are organized representations of information that support semantic reasoning and knowledge discovery in various fields. KGs consist of RDF (Resource Description Framework) triples formatted as (subject, predicate, object) (Bonatti et al., 2019). These triples represent factual statements and relationships between entities in a format that machines can read, which is important for applications like search engines, recommendation systems, and question answering. Building and maintaining a KG effectively relies on the precise extraction and verification of these triples from unstructured or semi-structured text sources.

In recent years, Large Language Models (LLMs) like GPT-4 (Achiam et al., 2023), Llama (Touvron et al., 2023) and DeepSeek (Liu et al., 2024) have shown significant success in various Natural Language Processing (NLP) tasks. These models capture complex linguistic and contextual patterns, leading to strong effectiveness in summarization (Zhang et al., 2025), translation (Elshin et al., 2024), and question answering (Zhuang et al., 2023). Due to their advanced generative and reasoning abilities, LLMs are being investigated as tools for automating the creation of RDF triples from natural language (Regino et al., 2023), offering a scalable method for constructing and enhancing KGs.

The outputs of the LLM generation need thorough evaluation, particularly regarding RDF triples. The reliability of these triples is not guaranteed, as LLMs may hallucinate, overlook contextual constraints, or generate outputs that are incompatible with the un-

[a] https://orcid.org/0000-0001-9814-1482
[b] https://orcid.org/0000-0002-7083-5789
[c] https://orcid.org/0000-0003-3441-0887
[d] https://orcid.org/0000-0001-9245-8753
[e] https://orcid.org/0000-0002-0529-7312
[f] https://orcid.org/0000-0002-9545-2098

Regino, A. G., Zagatti, F. R., Bonacin, R., Chico, V. J. S., Hochgreb, V. and Reis, J. C.
Leveraging Large Language Models for Semantic Evaluation of RDF Triples.
DOI: 10.5220/0013837600004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 2: KEOD and KMIS*, pages 74-85
ISBN: 978-989-758-769-6; ISSN: 2184-3228

derlying ontology. Without careful inspection, these triples can propagate factual errors, biases, or violate domain constraints, compromising the reliability and trustworthiness of the entire KG. This assessment is used to establish trust in the system, as users and downstream applications depend on the accuracy and relevance of the knowledge produced.

LLM-generated triples must conform to the ontological structure and semantics of the target KG. The evaluation process is typically manual, necessitating expert human reviewers, which renders it time-consuming, expensive, and challenging to implement at scale in practical applications. The systematic evaluation of these triples, particularly those generated by LLMs, remains an under-researched area (Regino and dos Reis, 2025), (Khorashadizadeh et al., 2024). There is a lack of established metrics, protocols, and frameworks for assessing triple quality in alignment with human expectations and ontological consistency. This intersection presents a valuable opportunity to investigate whether LLMs can serve as evaluators or 'judges" of structured data, potentially reducing the need for human reviewers and the heavy task that relies on them.

Based on our exploratory literature review, we observe that there is currently no widely accepted framework implemented as a software tool for automatically assessing the quality of RDF triples prior to their integration into a KG. The absence of automated and reliable evaluation mechanisms, particularly those that can closely mimic human judgment with quality, poses an obstacle in the construction and validation of KGs for real-world applications.

This article examines the effectiveness of LLMs as automated evaluators, or "judges", of RDF triples generated from natural language text. Although LLMs frequently create structured data from unstructured sources (Racharak et al., 2024), their ability to assess the semantic adequacy, clarity, and coherence of this structured output is not well-studied. We aim to assess whether current LLMs can achieve human-level evaluations of RDF triples in terms of their correctness and fidelity to the original sentences. Recognizing this capability can reduce the human efforts needed for triple validation, which is an obstacle in the construction and maintenance of adequate KGs.

We design and develop an evaluation pipeline with 12 configurations of LLMs, which include three model families (*Gemma*, *Qwen*, *Sabia*), two model sizes (*small*, *large*), and two prompting strategies (*zero-shot*, *few-shots*). Our study assesses each configuration using 300 Portuguese sentence-triple pairs, scoring them on four dimensions: subject, predicate, object, and the complete triple. Ten human annotators independently rated these triples to establish a gold standard. We analyze the ordinal ratings from the LLMs against the human judgments using two rank-based correlation metrics to determine the extent to which LLMs align with human preferences. Our investigation conducts a meta-evaluation to assess the clarity and coherence of the justifications provided by LLMs, thereby enhancing the quality assurance process.

The result analysis demonstrates that LLMs provide lower average scores than human annotators overall, except for the "subject" dimension, where LLMs rate more generously. Few-shot prompting, small models, and Sabia-family LLMs align most closely with human evaluations. This suggests that LLMs can capture some aspects of semantic plausibility; nevertheless, the overall correlation levels remain below those typically expected for high-stakes or expert-level annotation tasks. This highlights that, although promising, LLMs may not yet match human evaluators in accurately judging RDF triples across the dimensions of subject, predicate, object, and overall fidelity. Among the four evaluation dimensions, LLMs showed the strongest agreement with humans when assessing full triples, suggesting better effectiveness in holistic judgments compared to granular components, such as predicates or objects. The meta-evaluation confirms that few-shot prompts resulted in more coherent and interpretable outputs.

This study offers three key and original contributions:

1. **Benchmark Design:** We introduce a new benchmark consisting of 300 sentence-triple pairs in Portuguese language, annotated by experts across some semantic dimensions, to facilitate systematic evaluation of LLMs;
2. **Comprehensive Evaluation:** We conduct original empirical comparisons of 12 LLM configurations against human judgments using ordinal scales and correlation metrics, revealing patterns of alignment between models and human evaluations;
3. **Meta-Evaluation Strategy:** We establish a novel meta-evaluation step that employs a larger LLM to evaluate the quality of justifications from smaller models, emphasizing clarity and coherence;

The remaining of this article is organized as follows: Section 2 describes related work on LLM judges in RDF triple domain. Section 3 describes the experimental setup. Section 4 presents the results; Section 5 discusses the results; and Section 6 draws concluding remarks.

2 RELATED WORK

(Zheng et al., 2023) explored LLM as a judge for other LLMs' output. They introduced two benchmarks: MT-Bench, assessing multi-turn conversational and reasoning abilities, and Chatbot Arena, a crowdsourced platform for chatbot comparison. Their work investigated whether LLMs like GPT-4 can reliably approximate human preferences in dialogue evaluation. Through over 6,000 human judgments, they demonstrated that LLMs can achieve over 80% agreement with humans, comparable to human-human alignment, positioning the "LLM-as-a-judge" approach as a scalable alternative for conversational AI evaluation. While both studies explored LLMs for evaluation, our present study differs in domain, method, and focus. We target the semantic assessment of RDF triples derived from natural language, a task rooted in the Semantic Web rather than dialogue. Instead of agreement percentages, we measure alignment via correlation metrics (Spearman and Kendall Tau) on fine-grained components (subject, predicate, object, and whole triple).

Another relevant study, (Guerdan et al., 2025), addressed the challenges of validating LLM-as-a-judge systems, particularly in cases where human ratings are ambiguous or lack clear gold labels. Their work proposed a theoretical and empirical framework for understanding how rating task design, elicitation, aggregation schemes, and agreement metrics affect judge system validation. The authors demonstrated that current validation methods can select suboptimal judge systems—sometimes performing up to 34% worse than alternatives because they overlook the impact of task indeterminacy, where multiple valid ratings may exist. They demonstrated this through an empirical study using five commercial LLMs to evaluate "toxicity" on the Civil Comments dataset, highlighting how existing validation pipelines can produce misleading conclusions about both judge and target system performance. Their concern lies in the validation of LLM judges in the absence of clear ground truth. In contrast, our study focuses on the semantic evaluation of RDF triples, where the alignment between LLM and human judgments is assessed through statistical correlation.

Another relevant study explored the role of LLMs as judges in the context of KG construction (Huang et al., 2024). The proposed *GraphJudge* framework leverages open-source and closed-source LLMs to improve the quality of triples extracted from natural language. To address typical challenges such as information noise, domain knowledge gaps, and hallucinations, *GraphJudge* introduces three key components: a module for cleaning irrelevant information, another to adapt LLMs for triple validation tasks, and one for filtering incorrect triples. Their experiments on both general and domain-specific datasets demonstrated state-of-the-art effectiveness over baselines, achieving over 90% accuracy in triple validation. *GraphJudge* uses LLMs to directly filter and improve extracted triples as part of KG construction. Our present study analyzes the ability of LLMs to judge the semantic quality of given RDF triples against human evaluations. Their evaluation is benchmarked through classification accuracy, while ours relies on correlation metrics between human and LLM assessments.

Another recent study examined the use of LLMs as curators for validating RDF triple insertions into existing KGs (Regino and dos Reis, 2025). The authors proposed a systematic validation method covering four key aspects of RDF validation: class and property alignment, URI standardization, semantic consistency, and syntactic correctness. Using prompts to guide LLMs through each stage, they evaluated four models across these tasks. Results indicated that larger models, such as Llama-3 (70*B* Instruct), consistently outperform smaller ones, achieving high precision and recall, particularly in syntactic validation (accuracy of 0.99). They further highlighted practical challenges, including domain generalization, semantic drift, and the trade-off between cost and accuracy.

To the best of our knowledge, this study is the first to systematically investigate LLMs as semantic judges for RDF triples. It provides a detailed evaluation at the subject, predicate, object, and triple levels, and directly compares LLM judgments with human assessments through correlation analysis. Our current research investigates the impact of model size, language specificity, and prompting strategies, while also conducting a secondary evaluation of clarity and coherence. Our study provides new insights into the reliability, limitations, and practical applications of LLMs in KG curation and quality control, addressing a gap in previous research that has focused solely on extraction or structural validation.

3 EXPERIMENTAL EVALUATION METHOD

This section presents our evaluation method developed to assess the ability of LLMs to judge the semantic correctness of RDF triples extracted from natural language texts[1]. The method compares machine-

[1]The source code is available at:
https://github.com/andreregino/llm-as-a-judge

generated judgments to an original gold standard of human annotations, allowing an investigation of the reliability, consistency, and alignment of LLM-based evaluators in the context of semantic information extraction. Figure 1 presents the evaluation methodology.

The goal of this evaluation is to determine how accurately LLMs can act as evaluators of RDF triple quality. Given a text and its corresponding RDF triple, LLMs are prompted to judge the semantic plausibility of each triple and its components.

Subsection 3.1 describes the dataset used in the evaluation; Subsection 3.2 describes the LLMs; Subsection 3.3 reports on the prompt design; Subsection 3.4 shows the gold standard annotated by humans, Subsection 3.5 refers to the evaluation procedure; Subsection 3.6 presents the used evaluation metrics and Subsection 3.7 illustrates an example of the evaluation.

3.1 Dataset

The dataset (component 1 of Figure 1) used in this evaluation was constructed from real-world Brazilian e-commerce and marketplace platforms. It contains summarized text inputs and corresponding RDF triples generated by the [blind review], a question-answer-to-RDF-triple generation system developed for the e-commerce domain [blind review].

Each data instance represents a question-and-answer pair related to a product, automatically summarized into a single sentence, along with a semantic triple extracted from that summary. The source content includes customer questions, seller answers, and product names. These were processed to produce concise, representative sentences and corresponding structured triples.

The dataset comprises 300 Portuguese sentence–triple pairs collected from multiple stores and randomly sampled to ensure a diverse representation. It includes three distinct intent types: compatibility (e.g., “Is this tyre compatible with my vehicle?”), specification (e.g., “What is the memory capacity of this phone?”), and availability (e.g., “Is this item in stock in size M?”). Each intent type is represented equally, with 100 sentence–triple pairs per category. The examples span 20 different product domains, including automotive, electronics, apparel, and household goods, and reflect real usage scenarios from 2023.

The RDF triples follow the standard format subject-predicate-object and aim to capture the core semantics of the summarized statement. Figure 2 presents two examples of texts and their corresponding RDF triples from the dataset. We use the triple parts and the triple as a whole in the judgment process (component 2 of Figure 1).

3.2 Evaluated Large Language Models

A diverse set of LLMs was selected to act as graders (component 3 of Figure 1), automated judges responsible for evaluating the semantic correctness of RDF triples based on input texts. The motivation for using LLMs as judges stems from the nature of the content being evaluated. RDF triples are machine-readable structures intended for system interoperability, ontology alignment, and semantic reasoning — tasks primarily executed by machines rather than humans. Therefore, assessing whether a triple correctly captures the meaning of a sentence is, in this context, more naturally aligned with machine interpretation. If the evaluated output were intended for human consumption (e.g., a natural language summary), then a human-centric evaluation would be more appropriate.

The chosen models vary in architecture, size, and source, enabling a comparative analysis across different model families and capacities. The following models were used in the main evaluation:

- **Gemma 2 9B and 27B** – Google’s open-weight transformer model with 9 and 27 billion parameters;
- **Qwen 2.5 7B and 32B** – 7B and 32B parameter models from Alibaba’s Qwen series, optimized for instruction-following;
- **Sabiazinho 3 and Sabiá 3.1** – Both compact and more powerful Brazilian LLM fine-tuned for Portuguese understanding and instruction following;

The rationale behind selecting these models was: first, they cover a range of parameter sizes (from 3B to 32B), which enables an exploration of how model capacity influences evaluation quality; second, they represent different linguistic and regional orientations: while some models are general-purpose and multilingual (e.g., Qwen), others are specifically tuned for Portuguese and Brazilian contexts (e.g., Sabiazinho and Sabiá), which is the language of the dataset used in the evaluation process.

In addition to the graders listed above, DeepSeek R1 with 685B parameters, a much larger model, was used in a complementary role to assess the internal consistency of the judgments. Specifically, this model was tasked with evaluating whether the grading scores assigned by each LLM were clear and coherent with the textual justifications provided by those same LLMs. We call this a meta-evaluation model.

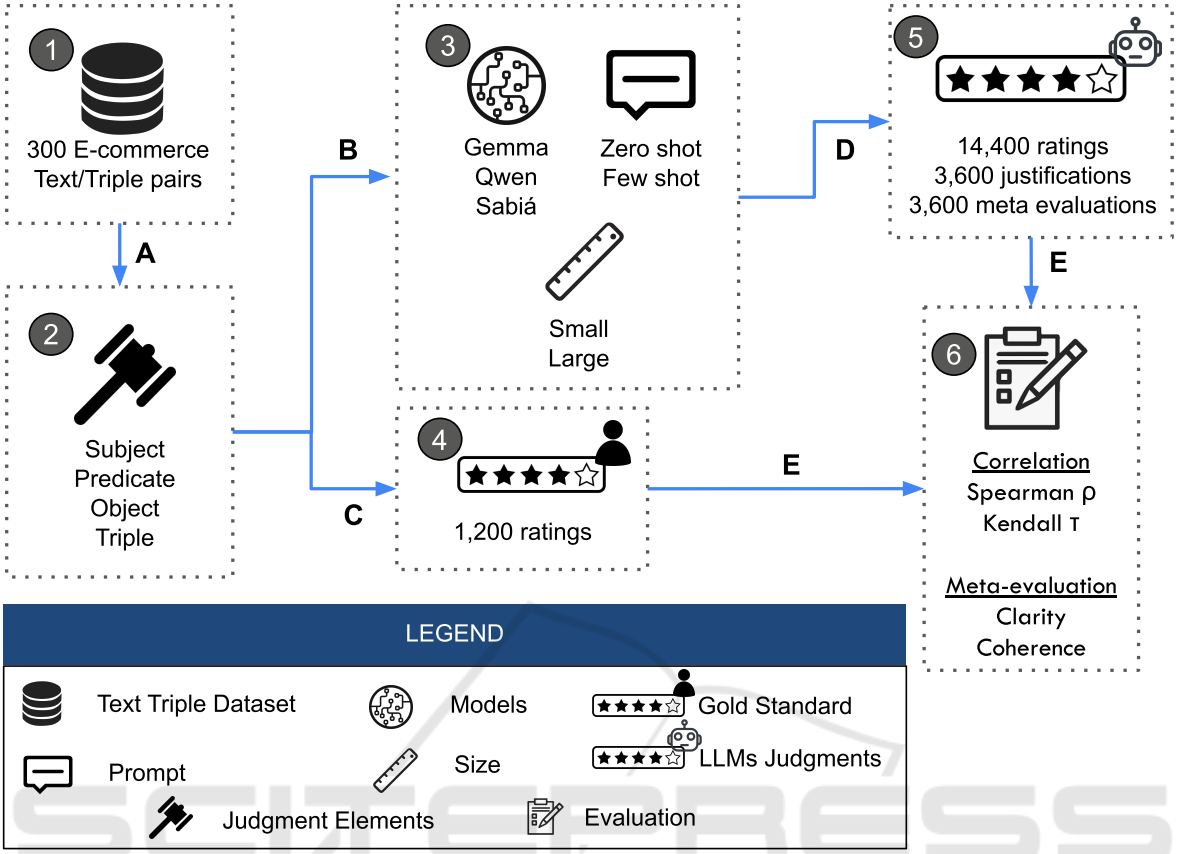

Figure 1: Methodology to evaluate LLMs' capacity in judging RDF generation from text. The numbers represent components and the letters the actions among them. A) Evaluation Components Definition; B) LLMs Setup Definition; C) Gold Standard Creation; D) LLM Judgment; E) Metrics Evaluation.

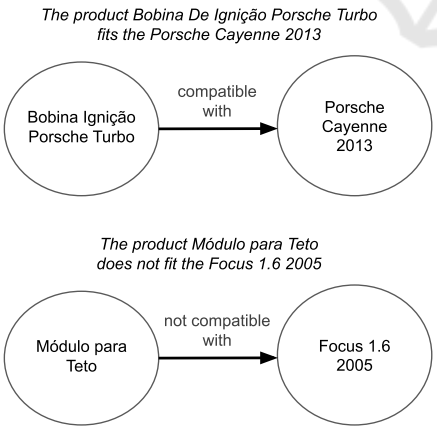

Figure 2: Two examples of summarized texts and their corresponding RDF triples.

This meta-evaluation model was chosen due to its significantly higher capacity, which allows for deeper reasoning and a stronger ability to detect logical inconsistencies between a justification and its corresponding evaluation (So et al., 2025). This consistency check was conducted separately from the main evaluation to preserve fairness and avoid biasing the comparative results (cf. Section 3.5).

3.3 Prompt Design

To evaluate the semantic plausibility of RDF triples, we designed two types of prompts (component 3 of Figure 1): one for the main evaluation LLMs (acting as graders), and another for a larger LLM used as a meta-evaluator (consistency checker).

The primary prompt was written in Portuguese and designed to simulate a structured evaluation scenario. It instructs the model to behave as an ontology and knowledge representation expert tasked with judging whether a given RDF triple accurately reflects the information in a corresponding natural language sentence. The input to this prompt consists of a summarized sentence derived from an e-commerce QA interaction and an RDF triple automatically extracted from the sentence.

The model is asked to provide four Likert-scale

ratings (from 1 to 5), assessing the correctness of the subject, predicate, and object, as well as the overall fidelity of the triple to the source sentence. A textual justification, with a maximum of 30 words, is also required to explain the rationale behind the scores.

Prompt for RDF Triple Generation from Text Judgment

You are an expert in natural language understanding and RDF. Your task is to judge **the quality of RDF triples generated from texts**. In other words, given a descriptive sentence and an RDF triple, you should act as a judger, grading the quality of the triples.

Here is the explanation of each grade:

{**likert_grades_and_explanations**}

Here is each of the judgments:

{**list_of_judgments**}

Below are some examples of the sentences, triples and their corresponding judgments:

Examples: {examples}

With the examples above as reference, generate the grades for the following sentence and triple:

Text: {text}
Triple: {triple}

Judgments:

This prompt format standardizes the evaluation across all grader LLMs and supports later comparison with human annotations.

Two versions of the grader prompt were used in the experiments:

- **Zero-Shot Prompt:** Contains only the evaluation instruction and the specific text–triple pair;
- **Few-Shot Prompt:** Includes two randomly selected examples of evaluated sentences and triples, followed by the new instance to be judged.

Few-shot prompting, a form of in-context learning (Brown et al., 2020), was chosen over tuning-based approaches due to its flexibility and lower data requirements. Fine-tuning would demand thousands of annotated examples, which is impractical in this case. Additionally, by testing both zero-shot and few-shot configurations, we aim to assess whether the presence of examples influences grading behavior or score inflation.

To assess the internal consistency of the LLM-generated judgments, we employed a separate prompt, that receives as input: the original sentence; the extracted RDF triple and the judgment provided by another LLM, consisting of four scores and a justification.

The task of this consistency evaluator is to assess two aspects:

1. **Clarity of the Justification:** Is the justification provided by the LLM understandable and well-written?
2. **Coherence of the Ratings:** Do the numeric scores logically match the justification?

This additional layer of evaluation is intended to validate whether the LLMs' justifications are not only syntactically well-formed but also logically consistent with their numerical evaluations. A larger model was chosen for this step to provide deeper reasoning capabilities and stronger metalinguistic analysis.

Both aspects are rated using a 1–5 Likert scale, and the output is structured as the following example:

```
{
  "EVALUATION 1": 4,
  "EVALUATION 2": 5
}
```

3.4 Gold Standard

To serve as a point of comparison for the LLM-generated judgments, we constructed a gold standard dataset composed of human evaluations (component 4 of Figure 1). These annotations were performed by a group of ten human evaluators, each one with prior exposure to Semantic Web technologies, Knowledge Graphs, and RDF modeling. The group includes professionals from the public sector, graduate students from academic institutions, and employees from private companies with technical backgrounds.

Each human evaluator was randomly assigned 30 summarized sentences and their corresponding RDF triples, totaling 300 unique sentence–triple pairs across the full dataset. The evaluators were instructed to assess the quality of the transformation from sentence to triple according to the same four dimensions used in the LLM evaluation prompts:

1. **Subject Correctness:** Was the subject correctly extracted?
2. **Predicate Correctness:** Does the predicate accurately represent the relationship expressed in the sentence?

3. **Object Correctness:** Does the object reflect the target of the relation?
4. **Triple Fidelity:** Does the complete triple faithfully represent the sentence's meaning?

For each question, the evaluators assigned a score on a 5-point Likert scale, which was clearly defined in the annotation protocol to minimize ambiguity:

1 – *Incorrect*

2 – *Many errors*

3 – *Approximately half correct*

4 – *Few errors*

5 – *Correct*

Given the cognitive efforts required to make accurate semantic judgments, the number of evaluations per annotator (30) was chosen to balance annotation reliability with fatigue reduction. The total number of annotated instances (300) reflects the practical constraints of recruiting and coordinating qualified human annotators familiar with the nuances of RDF triple structures and the Semantic Web, expertise that is not widespread even among computer science researchers.

3.5 Evaluation Procedure

The evaluation procedure was designed to systematically compare the judgments of human evaluators and multiple LLMs across 300 text-to-triple transformations. Both types of evaluators assessed the same semantic dimensions, allowing for direct comparison.

We evaluated three different LLM families—**Gemma**, **Qwen**, and **Sabiá**—each in two sizes (a smaller and a larger version), and under two prompt engineering strategies: *zero-shot* and *few-shot*. This resulted in 12 distinct evaluation setups:

$$3 \text{ models} \times 2 \text{ sizes} \times 2 \text{ prompt styles} = 12 \text{ setups}$$

Each of these 12 setups was used to independently evaluate all 300 sentence–triple pairs across the four semantic dimensions, generating a total of $14,400$ individual scores (12 setups × 300 triples × 4 ratings), along with $3,600$ textual justifications.

To quantify agreement levels, we calculated the average scores for each of the four dimensions in the human-annotated gold standard. Similarly, we computed the average scores produced by each of the 12 LLM setups. These aggregate values are compared in Section 4, allowing us to analyze how closely LLM judgments align with human expectations.

For each of the $3,600$ judgments generated by the grader models, DeepSeek model was used as meta-evaluator and prompted to assess the criteria presented in Section 3.3.

3.6 Metrics

For each of the four evaluation dimensions (subject, predicate, object, and full triple), we computed the correlation between the human scores and the scores produced by each of the 12 LLM setups across all 300 sentence–triple pairs (component 6 of Figure 1). The use of rank-based correlation metrics allows us to quantify the strength and direction of association between human and machine-generated judgments, even when the absolute values differ.

- **Spearman's** ρ (Spearman, 1987) evaluates the monotonic relationship between two variables. It is useful when we are interested in whether the ordering of scores (rather than their exact values) is preserved between annotators. This is appropriate in our case, where some variation in scoring behavior is expected, but consistency in ranking is desirable. The coefficient ranges from –1 to 1, where 1 indicates perfect agreement in ranking (i.e., identical order), 0 indicates no correlation, and –1 indicates a perfectly reversed ranking.
- **Kendall's** τ (Kendall, 1938) measures the number of concordant and discordant pairs between two rankings. It is more conservative than Spearman's ρ and more sensitive to small rank differences, making it a complementary metric that provides robustness to the analysis. Like Spearman's ρ, Kendall's τ ranges from –1 to 1.

Both metrics are well-suited for ordinal data such as Likert-scale judgments, and together they offer a comprehensive view of how similarly LLMs and human evaluators perceive the semantic quality of RDF triples.

3.7 Running Example

To illustrate the evaluation method in practice, we present a representative example from the dataset, including the original sentence, the corresponding RDF triple, and the evaluation provided by one of the LLM graders under the few-shot prompting condition.

Summarized Sentence. "O produto Kit Quadros Decorativos Sala Quarto Moldura está disponível em tamanho maior, desde que seja solicitado o tamanho

desejado." [original sentence in Portuguese][2]
RDF Triple. Kit Quadros Decorativos Sala Quarto Moldura — disponível — tamanho maior[original triple in Portuguese][3]
LLM Judgments. The LLM evaluated the semantic plausibility of the triple based on four distinct dimensions, assigning the following Likert-scale scores:

- **Subject Correctness:** 5;
- **Predicate Correctness:** 5;
- **Object Correctness:** 5;
- **Triple Fidelity:** 5.

Justification. *"The triple correctly represents the sentence, with the correct subject, predicate, and object well extracted from the context."*

This example illustrates a high-agreement case, where the LLM accurately interprets and assesses the semantic alignment between the natural language input and the RDF triple. All dimensions received the maximum score, and the justification shows logical clarity and alignment with the numeric judgments.

4 RESULTS

This section presents the obtained results of the evaluation experiments, organized into three parts. We first report average scores given by human and LLM evaluators (Subsection 4.1). Next, we analyze the correlation between LLM and human scores using Spearman's ρ and Kendall's τ (Subsection 4.2). Finally, we present the results of the meta-evaluation using a larger language model to assess the clarity and coherence of the LLMs' judgments (Subsection 4.3).

4.1 Average Ratings from Humans and LLMs

Human annotators (Gold Standard, GS) and each of the 12 LLM setups evaluated the same 300 sentence–triple pairs across four grading dimensions: subject, predicate, object, and full triple. The first step in our analysis was to compute mean scores for each grading dimension, both for human evaluators and for the combined LLMs.

The overall average score assigned by all LLMs across all dimensions was 4.300, whereas the average score from the GS was 4.498. This suggests that human annotators rated the triple transformations slightly more favorably than the LLMs.

When disaggregating the averages by evaluation dimension, we observed a pattern (Figure 3): LLMs assigned higher scores than humans only in the *subject* dimension. For *predicate*, *object*, and *triple*, human evaluators assigned higher grades.

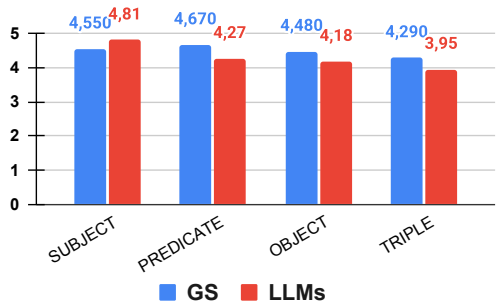

Figure 3: Grading average by dimensions (triple part) using Likert scale.

We computed average scores grouped by three comparative axes: (i) model family (Gemma, Qwen, Sabia), (ii) model size (small or large), and (iii) prompt type (zero-shot vs. few-shot). For each model, averages were calculated over all four executions (e.g., all versions of Gemma). Figure 4 presents these comparisons. The closest match to the human gold standard came from:

- **Few-shot prompts** ($\mu = 4.418$),
- **Sabia models** ($\mu = 4.451$), and
- **Small models** (higher than large counterparts).

Despite these proximities, no LLM setup surpassed the human average in any of the categories.

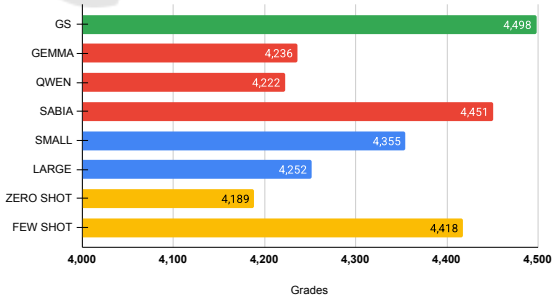

Figure 4: Comparison among average gradings of Gold Standard dataset (green), LLM family (red), LLM size (blue), and prompt type (yellow).

[2]The Decorative Picture Kit Living Room Bedroom Frame product is available in a larger size, as long as the desired size is requested. – translated by the authors.

[3]Decorative Picture Kit Living Room Bedroom Frame — available — larger size

4.2 Correlation Between LLMs and Human Ratings

To quantify the alignment between human and LLM ratings, we computed Spearman's ρ and Kendall's τ correlation coefficients for each of the 12 LLM setups, across each of the four evaluation dimensions. Table 1 (Spearman) and Table 2 (Kendall) present the attained results.

Table 1: Spearman scores for each component of the triple. The highest value in each column is highlighted in bold. In the first column, the LLM setup identifier is composed by three parts: the LLM (e.g GE for Gemma), the size (e.g SM for small) and prompt type (e.g ZS for zero shot).

LLM	Subject	Predicate	Object	Triple
GE_SM_ZS	0.112	0.216	0.350	0.372
GE_SM_FS	0.119	0.225	0.435	0.435
GE_LA_ZS	0.165	**0.244**	0.428	0.480
GE_LA_FS	0.134	0.198	0.414	0.443
QW_SM_ZS	0.075	0.110	-0.055	0.057
QW_SM_FS	0.104	0.147	0.406	0.426
QW_LA_ZS	**0.560**	0.172	0.350	0.358
QW_LA_FS	0.252	0.183	0.303	0.396
SA_SM_ZS	0.137	0.157	0.392	0.467
SA_SM_FS	0.078	0.148	**0.460**	0.421
SA_LA_ZS	0.265	0.241	0.310	0.375
SA_LA_FS	0.333	0.193	0.330	**0.491**

Table 2: Kendall Tau scores for each component of the triple. The highest value in each column is highlighted in bold. In the first column, the LLM setup identifier is composed by three parts: the LLM (e.g GE for Gemma), the size (e.g SM for small) and prompt type (e.g ZS for zero shot).

LLM	Subject	Predicate	Object	Triple
GE_SM_ZS	0.107	0.199	0.315	0.334
GE_SM_FS	0.115	0.214	0.405	0.395
GE_LA_ZS	0.157	0.217	0.388	0.424
GE_LA_FS	0.129	0.184	0.382	0.395
QW_SM_ZS	0.070	0.101	-0.050	0.051
QW_SM_FS	0.100	0.140	0.366	0.382
QW_LA_ZS	**0.530**	0.155	0.316	0.314
QW_LA_FS	0.241	0.167	0.272	0.354
SA_SM_ZS	0.128	0.143	0.353	0.423
SA_SM_FS	0.075	0.143	**0.436**	0.386
SA_LA_ZS	0.250	**0.226**	0.293	0.339
SA_LA_FS	0.322	0.183	0.309	**0.451**

Highest Correlations per Dimension

- **Subject:** Highest $\rho = 0.560$ (Qwen Large Zero-Shot), $\tau = 0.530$ (same set).
- **Predicate:** Highest $\rho = 0.244$ (Gemma Large Zero-Shot), $\tau = 0.217$ (same set).
- **Object:** Highest $\rho = 0.460$ (Sabia Small Few-Shot), $\tau = 0.436$ (same set).
- **Triple:** Highest $\rho = 0.491$ (Sabia Large Few-Shot), $\tau = 0.451$ (same set).

No correlation score exceeded 0.56 in any case, indicating a moderate agreement at best. To facilitate interpretation, we computed:

- **Mean correlation per LLM setup** (averaging across the four dimensions), yielding 12 Spearman and 12 Kendall values (Figure 5).
- **Mean correlation per evaluation dimension** (averaging across the 12 setups), yielding four mean Spearman and four mean Kendall scores (Figure 6).

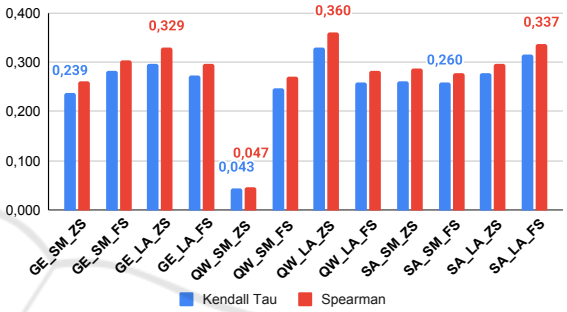

Figure 5: Correlation results by each setup. X axis shows the 12 setups (e.g GE_SM_ZS stands for Gemma small zero shot setup) and the y axis shows the values of Kendall and Spearman.

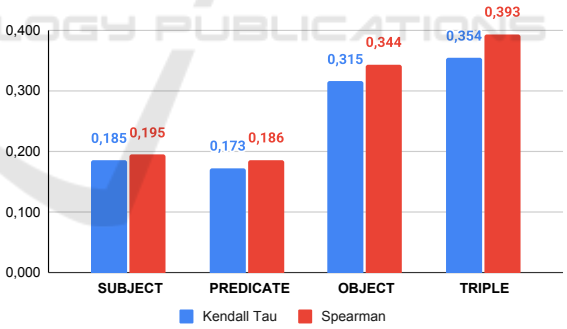

Figure 6: Correlation results by each dimension. X axis shows the 4 dimensions and the y axis shows the values of Kendall and Spearman.

The majority of LLM setups (11 out of 12) yielded mean correlations in the range of 0.2–0.35. When aggregated by dimension, we found that correlations for **subject** and **predicate** were notably lower than for **object** and **triple**. The highest average correlation appeared in the *triple* dimension:

- $\rho = 0.393$, $\tau = 0.354$

These results indicate that LLMs tend to align more closely with human evaluators when judging the semantic quality of the complete triple rather than its

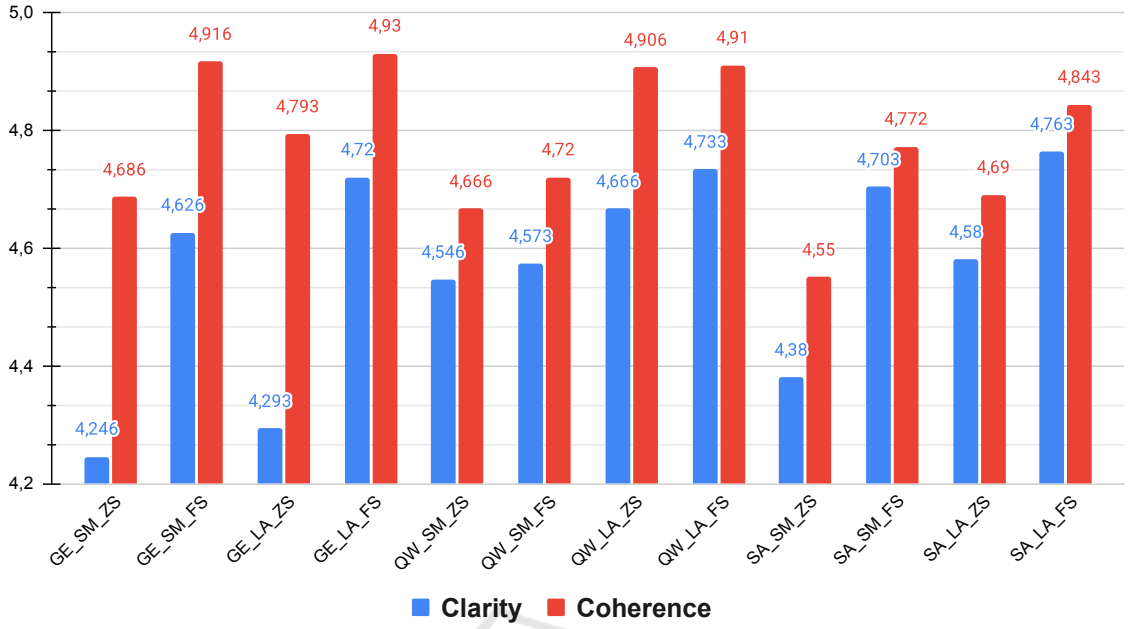

Figure 7: Clarity and coherence values by each setup.

individual components. The fact that both correlation metrics point to the same pattern strengthens the reliability of this conclusion.

4.3 Meta-Evaluation Using a Large Language Model

To assess the quality of the judgments produced by the 12 LLM setups, we employed DeepSeek R1 as a meta-evaluator. This model assessed each LLM judgment based on two criteria:

1. **Clarity of the justification**, and
2. **Coherence of the ratings with the justification**.

Figure 7 presents the results. Average scores per LLM setup ranged from 4.245 to 4.930 for both dimensions.

- **Coherence** scores were consistently higher than **clarity** scores across all models.
- **Few-shot setups** (denoted by "FS" in the Figure 7) achieved higher average scores than their zero-shot counterparts.

These results suggest that the judgments provided by the smaller LLMs were, on average, both interpretable and internally consistent—particularly when few-shot prompting was used.

5 DISCUSSION

This study originally explored an evaluation framework that combines expert human judgments with automated assessments from LLMs across twelve configurations. Each configuration differed in model family (Gemma, Qwen, and Sabia), model size (small vs. large), and prompting strategy (zero-shot vs. few-shots). This design enabled a thorough examination of the study's core question: Can LLMs reliably evaluate structured outputs, such as RDF triples? Our results indicate that although LLMs show promising capabilities, they do not consistently achieve human-level discernment across all evaluation criteria.

The credibility of the gold standard annotations arises from the expertise of all annotators in Semantic Web technologies and KG construction. A carefully designed Likert scale facilitated nuanced evaluations across four dimensions: subject, predicate, object, and full triple. Each annotator assessed 30 distinct instances to maintain a balanced distribution of labor, and the overall design emphasized consistency despite the sample size (N=300). Recruiting experts for this specialized task presented logistical challenges, highlighting the need to explore automated grading alternatives while preserving the rigor of the gold standard.

The LLMs' efficacy varied significantly across the 12 evaluation setups. Larger models tended to produce scores more aligned with human judgments, particularly under the few-shot prompting paradigm.

Few-shot setups provided the models with exemplars that likely improved semantic grounding and scoring consistency. However, some smaller models unexpectedly demonstrated sharper precision in isolated cases, particularly in predicate evaluation. These inconsistencies highlight that model size alone is not a sufficient predictor of alignment; prompting context and model architecture also play important roles. The results underscore that while LLMs can replicate some human-like judgment behavior, their reliability fluctuates depending on setup.

We used Spearman and Kendall Tau metrics to measure the alignment between human and LLM scores across all four evaluation axes. Some configurations, particularly those with large models and few-shot prompting, displayed moderate positive correlations, especially in subject and full triple evaluations. Many setups produced low or inconsistent correlations, indicating that LLMs, in their current state and configuration (open-source, mid-sized models in the e-commerce domain), are not yet suitable for fully replacing human evaluators. Their best application may be as copilots in the RDF triple generation pipeline, aiding in draft evaluations that humans subsequently review.

We utilized DeepSeek to analyze the quality of 30-word justifications provided by LLM graders, validating their reliability. This secondary meta evaluation focused on two aspects: the clarity of justifications and the coherence of assigned scores. Notably, most setups received high scores from DeepSeek, indicating that LLMs often articulated their reasoning coherently, even when their ratings differed from human assessments. This finding highlights that the quality of justification may improve trust and transparency, despite potential misalignment in scores. However, DeepSeek's evaluations did not strongly correlate with human ratings, revealing a gap between the clarity of explanations and the accuracy of judgments.

Our findings indicate significant implications for using LLMs in Semantic Web evaluation tasks. Although LLMs cannot yet replace expert annotators, they can serve as valuable auxiliary evaluators, especially for large-scale RDF generation from text. Their capability to produce structured justifications is particularly beneficial in hybrid pipelines, where automated scoring complements human review. Future research should investigate multi-agent architectures, enabling several LLMs to collaboratively assess and vote on outputs, which may enhance evaluation reliability. Improvements through optimized prompt design, focused fine-tuning, or human-in-the-loop frameworks can lead to more trustworthy and effective systems in semantic evaluation contexts.

Both the dataset and the Sabia models employed in this study were optimized for Portuguese. This introduces a language-specific constraint that may limit the generalizability of our results. While the findings provide strong evidence within the Portuguese domain, further work is needed to validate whether similar outcomes hold across other languages. As a natural next step, future research should explore multilingual evaluation settings, testing the robustness of the proposed approaches in diverse linguistic contexts and broadening their applicability to global Semantic Web scenarios.

6 CONCLUSION

Knowledge Graphs are central to organizing structured information, making the automation of their validation important for scaling their real-world applications. Large Language Models can interpret both natural and formal languages, offering an alternative to the labor-intensive process of human evaluation. This study examined 12 configurations of LLMs as graders of RDF triples generated from e-commerce texts, comparing their scores to those from expert human annotators. We measured the alignment between LLM and human evaluations using Spearman and Kendall Tau correlations, analyzing both complete triples and their components (subject, predicate, object). The results indicated that while some configurations achieved moderate correlation, no model consistently matched human reliability across all components. We also assessed justification-based ratings using the DeepSeek R1 reasoning model, which highlighted potential indicators of grader quality. We found that these indicators did not strongly correlate with human judgments. Our findings indicate that while LLMs cannot fully replace expert reviewers, they serve as promising supportive tools for semantic validation tasks, providing scalability and facilitating preliminary assessments. Importantly, the moderate correlation scores highlight that the proposed approach is not yet ready for deployment in high-stakes KG validation scenarios where accuracy and reliability are critical. At present, LLM-based validation should be regarded as complementary, augmenting but not substituting expert judgment. Our evaluation analysis study lays the groundwork for future research that integrates model reasoning, prompt refinement, multilingual testing, and human-in-the-loop systems to improve the accuracy, trustworthiness, and efficiency of RDF validation workflows.

ACKNOWLEDGMENTS

We thank the National Council of Technological and Scientific Development (CNPq), Brazil, grant #301337/2025-0.

REFERENCES

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bonatti, P. A., Decker, S., Polleres, A., and Presutti, V. (2019). Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371). *Dagstuhl reports*, 8(9):29–111.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Elshin, D., Karpachev, N., Gruzdev, B., Golovanov, I., Ivanov, G., Antonov, A., Skachkov, N., Latypova, E., Layner, V., Enikeeva, E., et al. (2024). From general llm to translation: How we dramatically improve translation quality using human evaluation data for llm finetuning. In *Proceedings of the Ninth Conference on Machine Translation*, pages 247–252.

Guerdan, L., Barocas, S., Holstein, K., Wallach, H. M., Wu, Z. S., and Chouldechova, A. (2025). Validating llm-as-a-judge systems in the absence of gold labels. *CoRR*.

Huang, H., Chen, C., He, C., Li, Y., Jiang, J., and Zhang, W. (2024). Can llms be good graph judger for knowledge graph construction? *arXiv preprint arXiv:2411.17388*.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Khorashadizadeh, H., Amara, F. Z., Ezzabady, M., Ieng, F., Tiwari, S., Mihindukulasooriya, N., Groppe, J., Sahri, S., Benamara, F., and Groppe, S. (2024). Research trends for the interplay between large language models and knowledge graphs. *arXiv preprint arXiv:2406.08223*.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Racharak, T., Wang, T., and Jearanaiwongkul, W. (2024). An automated medical rdf knowledge graph construction from text using in-context learning. In *2024 16th International Conference on Knowledge and System Engineering (KSE)*, pages 465–471. IEEE.

Regino, A. and dos Reis, J. C. (2025). Can llms be knowledge graph curators for validating triple insertions? In Genet Asefa Gesese, Harald Sack, H. P. A. M.-P. and Chen, L., editors, *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK) co-located with the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 87–99, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Regino, A. G., Caus, R. O., Hochgreb, V., and dos Reis, J. C. (2023). From natural language texts to rdf triples: A novel approach to generating e-commerce knowledge graphs. In Coenen, F., Fred, A., Aveiro, D., Dietz, J., Bernardino, J., Masciari, E., and Filipe, J., editors, *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 149–174. Springer Nature Switzerland.

So, C. C., Sun, Y., Wang, J.-M., Yung, S. P., Loh, A. W. K., and Chau, C. P. (2025). Are large language models capable of deep relational reasoning? insights from deepseek-r1 and benchmark comparisons. *arXiv preprint arXiv:2506.23128*.

Spearman, C. (1987). The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Zhang, H., Yu, P. S., and Zhang, J. (2025). A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Zhuang, Y., Yu, Y., Wang, K., Sun, H., and Zhang, C. (2023). Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36:50117–50143.