September 2026

Ontology-Aware Prompting for Knowledge Graph Construction from Text

Balram TIWARI [a,b,1], Italo Lopes OLIVEIRA [a,2], Asep Fajar FIRMANSYAH [a], Hamada M. ZAHERA [a], Frank HOPFGARTNER [b], and Axel-Cyrille NGONGA NGOMO [a]

[a] *Data Science Group, Heinz Nixdorf Institute, Paderborn University, Germany*
[b] *Institute for Web Science and Technologies, University of Koblenz, Germany*
ORCiD ID: Balram Tiwari https://orcid.org/0009-0005-2192-7512, Italo Lopes Oliveira https://orcid.org/0000-0002-2357-5814, Asep Fajar Firmansyah https://orcid.org/0000-0001-5775-6966, Hamada M. Zahera https://orcid.org/0000-0003-0215-1278, Frank Hopfgartner https://orcid.org/0000-0003-0380-6088, Axel-Cyrille Ngonga Ngomo https://orcid.org/0000-0001-7112-3516

Abstract. Constructing Knowledge Graphs (KGs) from unstructured text has traditionally relied on multi-stage pipelines combining Named Entity Recognition, Relation Extraction, and Entity Linking components. These architectures are costly to adapt to new domains and suffer from error propagation across stages. While Large Language Models (LLMs) enable end-to-end triple generation from text, their direct application introduces two key challenges: (1) hallucination, where generated facts are not supported by the input text, and (2) ontology misalignment, where extracted entities and relations violate the constraints of a target schema. To address these limitations, we propose an ontology-aware multi-prompt framework that integrates the target ontology into the prompting process and applies hierarchical verification after triple generation. Specifically, we combine Tree-of-Thoughts-style reasoning for structured exploration, schema-constrained OpenIE, and relaxed triple generation to produce a diverse set of candidate triples while maintaining ontological awareness. A rule-based three-filter filtering module then refines these candidates using cross-prompt agreement, evidence grounding, and surface-form alignment, without relying on external toolchains. Experiments on Text2KGBench, including DBpedia-WebNLG and Wikidata-TekGen, show improvements of up to 2.3 times over LLM-based baselines, together with hallucination rates below 0.20 in most domains and ontology conformance above 0.90 on most DBpedia-WebNLG domains. The framework is released as open-source and available at https://github.com/dice-group/ontology-aware-kg-construction.

Keywords. Knowledge Graph Construction, Large Language Models, Ontology-Aware Prompting

[1] Corresponding Author: Balram Tiwari, balram@uni-paderborn.de
[2] Corresponding Author: Italo Lopes Oliveira, italo.lopes.oliveira@uni-paderborn.de

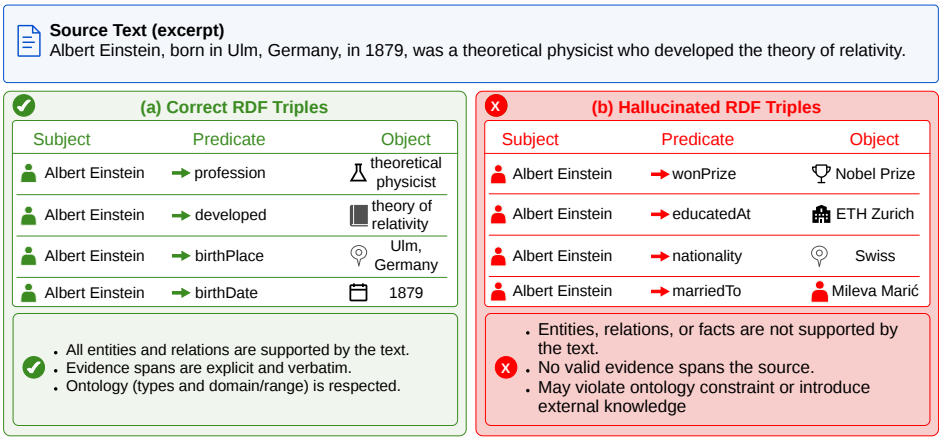

Figure 1. Examples of correct (a) vs. hallucinated (b) RDF triples extracted from text.

1. Introduction

Knowledge graphs (KGs) represent entities, concepts, and their relationships as Resource Description Framework (RDF) triples. They support a wide range of downstream tasks, including semantic search [1], question answering [2], and recommendation systems [3]. Traditionally, KG construction is decomposed into a sequence of specialized tasks: Named Entity Recognition (NER), Entity Linking (EL), and Relation Extraction (RE) [4,5]. These tasks transform unstructured text into structured RDF triples in one pipeline that populate a target ontology. However, these pipelines are costly to engineer and tightly coupled to fixed schemas, which has motivated a shift toward generative approaches [6].

Recent advances in Large Language Models (LLMs) introduce a new paradigm for automated KG construction [7]. Instead of chaining NER, EL, and RE, an LLM can generate RDF triples directly from text in a single step. Figure 1(a) illustrates this on the sentence *"Albert Einstein, born in Ulm, Germany, in 1879, was a theoretical physicist who developed the theory of relativity."*. From this input, the LLM produces the triple *<Albert_Einstein, profession, theoretical_physicist>*, where *Albert_Einstein* is the subject, *theoretical_physicist* is the object, and *profession* is the relation that links them. Despite this promise, LLM-based extraction suffers from two persistent limitations. The first is *hallucination*, where the model generates a triple that is plausible but not supported by the source text [8]. The second is *ontology misalignment*, where the generated entities or relations violate the type, predicate, or domain/range constraints of the target ontology. Figure 1(b) illustrates the first failure with the triple *<Albert_Einstein, nationality, Swiss>*. This statement is historically true, but it is not present in the source sentence. Plausible but unsupported claims of this kind are particularly damaging in high-precision domains such as healthcare, law, and scientific knowledge management. Both limitations have received attention in the literature. However, to the best of our knowledge, no prior work tackles them jointly. For instance, COMET [9] and AutoKG [10] extract triples directly from text, yet provide no mechanism to enforce factual grounding or ontology compliance. More recent work begins to address the ontology side. Norouzi

et al. [11] and ODKE+ [12] guide the LLM through ontology-aware prompts, while Aljabari et al. [13] apply the same idea to ontology alignment. However, all three rely on a single prompting strategy, limiting the available signals to detect possible hallucinations. Among them, only ODKE+ introduces explicit safeguards against hallucination, and even there, the coverage at the entity and relation level remains partial. Consequently, to the best of our knowledge, no existing approach combines multi-prompt extraction with explicit hallucination control under an ontology.

In this work, we present an ontology-aware, multi-prompt framework for end-to-end KG construction. Our framework includes two stages: i) a multi-prompt extractor that produces candidate triples through three complementary strategies, each occupying a different point on the precision–schema–recall trade-off. ii) a three-filter validator that decides which candidates to keep. Each filter applies a different acceptance criterion. *Filter 1* accepts triples on which multiple prompts agree and whose arguments are textually anchored in the source. *Filter 2* accepts triples that arrive with an explicit evidence span and pass a weighted evidence score. *Filter 3* accepts the remaining candidates through a surface-grounding fallback. We evaluate the framework on the Text2KGBench benchmark [14], covering the DBpedia-WebNLG and Wikidata-TekGen datasets across 29 ontologies. The results show consistent improvements up to 2.3 times in precision, recall, and F1-score over fine-tuned baselines such as ALPACA-LORA-13B and VICUNA-13B. In addition, the framework reduces the three Text2KGBench hallucination metrics — Subject Hallucination (SH), Object Hallucination (OH), and Relation Hallucination (RH) — on both datasets, with values below 0.20. An ablation study further isolates the contribution of each component to the overall performance. We summarize the main contributions of this work as follows:

1. A multi-prompt extraction stage that generates candidate triples through three complementary strategies, *Tree-of-Thoughts Search*, *Schema-Constrained Open IE*, and *Relaxed Extraction*, each occupying a different position on the precision–schema–recall trade-off.
2. A three-filter, evidence-driven validator that accepts a triple only when one of three criteria is satisfied: cross-prompt agreement with lexical anchoring (*Filter 1*), an explicit evidence span passing a weighted score (*Filter 2*), or a surface-grounding fallback (*Filter 3*).
3. A comprehensive evaluation on the Text2KGBench benchmark across 29 ontologies and two domains (DBpedia-WebNLG, Wikidata-TekGen), showing consistent F1 improvements over fine-tuned baselines (ALPACA-LORA-13B, VICUNA-13B) and reduced Subject, Object, and Relation Hallucination rates.

2. Related Work

2.1. Traditional KG Construction

Traditional KG construction relies on pipeline-based methods that combine NER [4], RE [5], Event Extraction (EE) [15], and EL [16] from unstructured text. These pipelines are implemented using either rule-based or machine learning approaches. Rule-based systems employ pattern matching and dependency path heuristics, offering interpretability and high precision in domain-specific settings [17,18]. Statistical and neural methods,

including supervised learning and distant supervision, improve scalability and generalization to unseen data [19,20]. Open Information Extraction (OpenIE) methods [21,22] enable open-domain extraction without predefined ontologies but produce noisy and redundant outputs due to limited semantic constraints. Despite high-quality results when carefully tuned, these pipelines require substantial human effort, domain expertise, and labeled data, and they suffer from error propagation across components [23,24].

2.2. Generative KG Construction Using LLMs

Recent advances in LLMs have enabled direct triple generation from unstructured text without intermediate NLP components [24]. This end-to-end approach offers scalability but introduces two key challenges: hallucination, the generation of plausible but unsupported facts [8], and ontology misalignment when triples do not adhere to target schemas [24]. Several studies focus on open-domain extraction without explicit ontology constraints. For instance, ChatIE [25] formulates information extraction as a multi-turn question-answering problem to extract entities and relations in a zero-shot setting. AutoRE [26] introduces a Relation-Head-Facts paradigm for document-level relation extraction without requiring predefined relation types. Similarly, Papaluca et al. [27], and KGGen [28] evaluate zero-shot triplet extraction. Recent work embeds chain-of-thought reasoning into LLM prompts for KG construction [29], though without ontology context to guide reasoning steps or dedicated mechanisms to filter hallucinated outputs. Early generative efforts include COMET [9] for commonsense inference. Text2KGBench [14] established baselines (Alpaca-LoRA [30,31], Vicuna [32]) on ontology-aligned datasets. AutoKG [10] combines LLMs with external tools via agent coordination. iText2KG [33] enables zero-shot incremental construction. KARMA [34] employs multiple agents for extraction and validation in biomedical domains.

2.3. Ontology-Aware KG Construction

Our work belongs to the category of ontology-aware KG construction, where schema information is explicitly incorporated to guide LLM-based extraction and improve ontology compliance. Kommineni et al. [35] proposed a semi-automatic approach that first extracts an ontology from text using user-defined competency questions and then uses the extracted ontology to retrieve triples. Feng, Wu, and Meng [36] extended this direction by using an LLM to generate competency questions automatically for ontology extraction, followed by guided triple extraction on Wiki-NRE [37], SciERC [38], and WebNLG [39]. In the medical domain, Bhattarai et al. [40] proposed an ontology-guided prompting approach that extracts concepts from the UMLS ontology to personalize prompts for triple extraction. Their evaluation on the n2c2 [41] and ADE [42] datasets demonstrated improvements over standard retrieval-augmented generation techniques. Recent work systematically investigates the role of ontology in LLM-based KG construction. Norouzi et al. [11] studied LLM-based ontology population, highlighting the importance of prompt design and ontology structure for improving extraction accuracy and reducing hallucinations. Their modular ontology method dynamically builds context by summarizing knowledge base texts and retrieving relevant information through embedding-based search. Khorshidi et al. [12] proposed ODKE+, a production-grade system that dynamically generates ontology snippets tailored to each entity type to align

extractions with schema constraints. The system employs post-extraction validation using additional LLM calls to verify and correct extracted triples against the ontology, demonstrating effectiveness in reducing schema violations.

2.4. Positioning of Our Work

Our work differs from prior ontology-aware approaches in two key respects. *First*, while existing methods such as ODKE+ [12] and Norouzi et al. [11] inject ontology terms as unstructured context into prompts, our framework structures the generation process using Tree-of-Thoughts reasoning to explore candidate entity types and relation constraints before committing to triple outputs. Moreover, we employ additional prompting strategies, including OpenIE and one-shot prompting, to extract more RDF triples in different domains. *Second*, unlike approaches that rely on external toolchains (e.g., AutoKG [10]) or multiple LLM calls for validation (e.g., ODKE+ [12]), our method employs a lightweight rule-based three-filter filtering module that operates at the entity, relation, and triple levels without requiring additional API calls.

3. Approach

We propose an ontology-aware, multi-prompt framework for end-to-end RDF triple extraction from unstructured text, as shown in Figure 2. Given input text x and a target ontology Ω, our framework works in two stages: (1) a *multi-prompt extractor* that generates a diverse pool of candidate triples P using three distinct LLM prompting strategies, and (2) a *hierarchical, rule-based validator* that filters these candidates through a three-tier process to ensure factual grounding and ontology compliance. Ontology awareness in our framework means that the target ontology is used to guide prompting, candidate generation, and lightweight validation. The ontology provides canonical predicates, type descriptions, and domain/range information that influence which triples are proposed and accepted. The validator, however, does not perform full reasoning over OWL axioms or exhaustive checks of class membership and logical consistency.

3.1. Stage 1: Generating Candidate Triples

Instead of relying on a single extraction method, we use three prompting strategies that vary in reasoning depth, schema strictness, and recall orientation. This diversity provides the downstream evaluator with three complementary candidate pools ($P_1, P_2, P_3 \subseteq P$), respectively: high-precision triples from structured reasoning, ontology-aligned triples with explicit evidence, and high-recall triples from a more relaxed extraction process.

Strategy 1: Tree-of-Thoughts Semantic Extraction The first strategy focuses on precision by framing RDF extraction as a search problem, following the Tree-of-Thoughts paradigm [43]. The search begins from an empty state S_0 and proceeds over partial triple sets $S \subseteq \mathrm{T}$, where each state S represents a set of candidate triples for the current input x under ontology Ω. At each step, a generator $\texttt{Gen}(p_{\theta_g}, S)$ proposes up to k new triples, where p_{θ_g} is the generator prompt. Each resulting state $S' = S \cup \{(s,r,o)\}$ is then scored in $[0,1]$ by an evaluator $\texttt{Eval}(p_{\theta_e}, S')$, where p_{θ_e} is the prompt used for evaluation. This score reflects two properties: how well the triple is grounded in the source text x, and

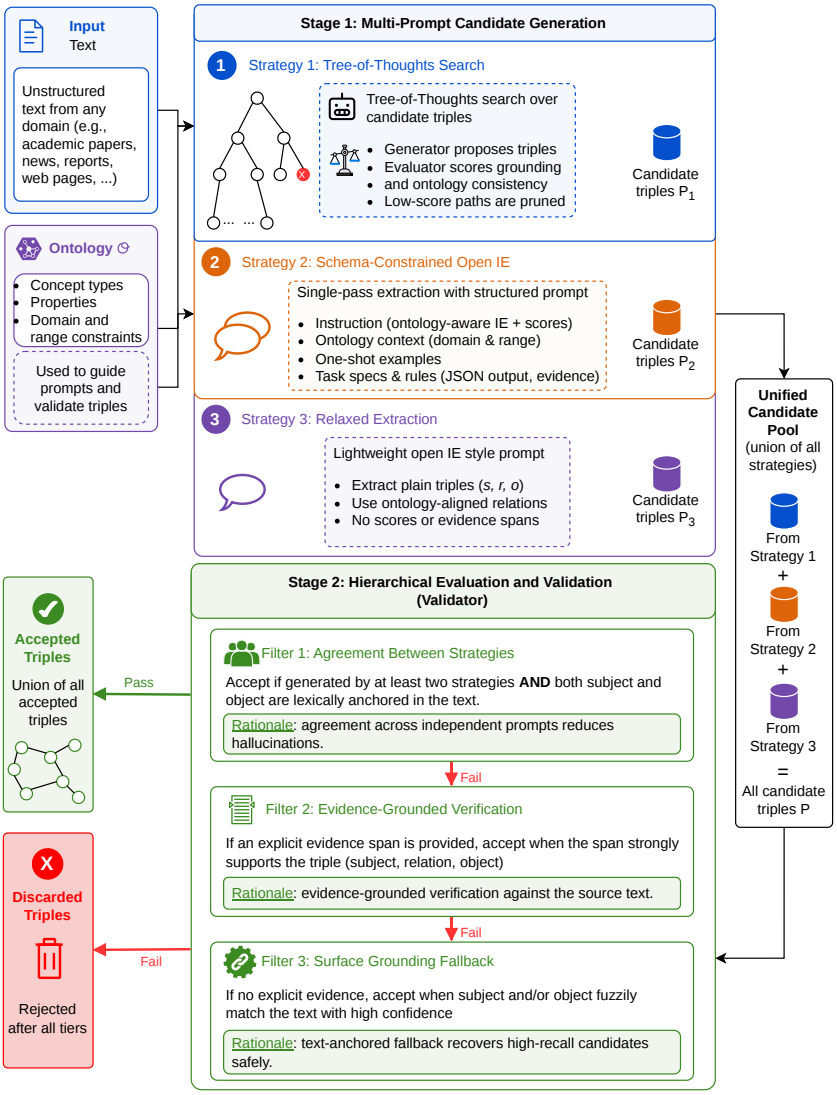

Figure 2. Ontology-aware multi-prompt framework for constructing knowledge graphs from text. Stage 1 combines three LLM prompting strategies to generate candidate triples, and Stage 2 applies a three-tier validator to filter hallucinations and enforce ontology compliance.

how well it aligns with the ontology Ω. If the score falls below a threshold τ, the branch is pruned, removing weak candidates early on. Expansion continues until a maximum depth $D_{\max}$. The highest-scoring deep paths form the output S^*, and each accepted triple is stored alongside its supporting textual span $\texttt{support}(s, r, o) \subseteq x$.

Strategy 2: Schema-Constrained Open IE The second strategy aims for schema-aligned extraction in a single pass by explicitly binding the model to the ontology. The prompt includes four parts. (i) *instructions* that frame the task as Open IE with confidence scores in $[0, 1]$. (ii) an *ontology block* that lists the `domain` and `range` for each predicate

as short textual descriptions rather than raw RDF/OWL syntax. (iii) a *one-shot example* showing a valid triple alongside its evidence span, and (iv) a *task specification* that presents the input text x and requests strict JSON output containing exact quotes from the input text. This produces ontology-aligned triples paired with direct textual support.

Strategy 3: Relaxed Extraction The third strategy maximizes recall by removing structural constraints. The model receives only the input text (x) and triples set P, and generates plain triples (s,r,o). It omits confidence scores, reasoning traces, and evidence spans. Because rigid formats can cause the model to miss valid triples, relaxing these constraints recovers additional candidates. Although this increases noise, the trade-off is acceptable because Stage 2 filters it.

3.2. Stage 2: Three-Tier Validation

Stage 1 produces a mixed pool of candidates P. To separate reliable triples from noise, the validator asks the question for each candidate triple t: What kind of evidence supports this triple? Since different triples carry different types of evidence (cross-prompt agreement, explicit spans, or none), we use a three-filter cascade approach. A triple is accepted as soon as it satisfies one of the filters, following from stronger evidence to weaker ones. Filter 1 requires both multi-prompt agreement and lexical anchoring for the subject and object, and therefore tends to admit fewer but highly supported triples. Filter 2 focuses on explicit evidence spans and admits triples whose supporting span achieves a high evidence score E_B. Filter 3 serves as a permissive fallback: it recovers additional candidates based on fuzzy grounding and overall text–triple coherence when no explicit span is available. This cascade prioritizes high-confidence triples first while still recovering recall through softer conditions at the final stage. Algorithm 1 summarizes this process. Here, $\mathrm{sim}(\cdot,\cdot) \in [0,1]$ denotes a string-matching score between an entity surface form and the source text x, implemented as a normalized edit similarity.[44]. The functions $E_B, E_C \in [0,1]$ are evidence scores and used consistently throughout the paper; Section 4.2 specifies the exact thresholds used in our experiments.

Filter 1: Cross-Prompt Consensus with Lexical Anchoring. Filter 1 of Algorithm 1 (lines 4–5) exploits agreement across prompting strategies. The intuition is: if multiple independent prompts generate the same triple, that triple is unlikely to be a shared hallucination. We achieve this with the two complementary conditions in line 4, both of which must hold for a candidate to be accepted. The first condition, *multi-prompt agreement*, requires the candidate t to appear in at least two of the three candidate pools, that is, $|\{\, j \mid t \in P_j \,\}| \geq 2$. This extends the self-consistency principle [45] from a single model's sampled reasoning paths to a multi-prompt ensemble. However, agreement alone is not enough to filter all hallucinated triples, since independent prompts may still converge on a plausible but text-absent triple. To rule out this case, the second condition, *lexical anchoring*, requires the subject and object of t to each match the source text x above the grounding thresholds τ_s^{anc} and τ_o^{anc}; concretely, $\mathrm{sim}(s,x) \geq \tau_s^{\mathrm{anc}}$ and $\mathrm{sim}(o,x) \geq \tau_o^{\mathrm{anc}}$. Whenever both conditions hold, line 5 adds t to the validated set T_1.

Filter 2: Evidence-Grounded Verification. A triple that fails Filter 1 may still be reliable if its originating prompt provided a verbatim span $\texttt{support} \subseteq x$. The cascade therefore falls through to Filter 2 of Algorithm 1 (lines 6–7). We evaluate the triple against its own evidence by computing a normalized score E_B, defined as the weighted sum of four

Algorithm 1 Hierarchical Ontology-Aware Validation

Require: Candidate pools P_1, P_2, P_3; source text x; anchoring thresholds $\tau_s^{\text{anc}}, \tau_o^{\text{anc}}$; fuzzy-match thresholds $\tau_s^{\text{fuz}}, \tau_o^{\text{fuz}}$; acceptance thresholds τ_B, τ_C.
Ensure: Validated triple set $\hat{G}$

1: $\hat{G} \leftarrow \emptyset$
2: $T_1 \leftarrow T_2 \leftarrow T_3 \leftarrow \emptyset$
3: **for all** $t = (s, r, o) \in P_1 \cup P_2 \cup P_3$ **do**
4: **if** $|\{j \mid t \in P_j\}| \geq 2$ **and** $\text{sim}(s, x) \geq \tau_s^{\text{anc}}$ **and** $\text{sim}(o, x) \geq \tau_o^{\text{anc}}$ **then**
5: $T_1 \leftarrow T_1 \cup \{t\}$ ▷ Filter 1: agreement + lexical anchoring
6: **else if** t carries supporting span $\textit{support}(s, r, o) \subseteq x$ **and** $E_B(t, x) \geq \tau_B$ **then**
7: $T_2 \leftarrow T_2 \cup \{t\}$ ▷ Filter 2: evidence score E_B
8: **else if** $E_C(t, x) \geq \tau_C$ **then**
9: $T_3 \leftarrow T_3 \cup \{t\}$ ▷ Filter 3: surface fallback score E_C
10: **end if**
11: **end for**
12: $\hat{G} \leftarrow T_1 \cup T_2 \cup T_3$
13: **return** $\hat{G}$

heuristic features: complete-match, subject-match, object-match, and span-conciseness. Respectively, they are weighted by $\alpha_B, \beta_B, \gamma_B$, and δ_B. The complete-match returns 1 if the subject, predicate, and object all appear within the evidence span; this carries the highest weight, since joint surface presence is the strongest single indicator of textual support. The subject-match and object-match check whether each argument appears in the evidence span. Finally, the span-conciseness is computed as the Jaccard similarity [46] between the evidence span and the source text, penalizing overly broad spans that would otherwise inflate the score. When $E_B \geq \tau_B$, line 7 admits t into the Filter 2 set T_2. This design follows the claim–evidence pairing principles established in automated fact-verification research [47].

Filter 3: Surface Grounding Fallback. Candidates that fail both earlier branches reach the Filter 3 of Algorithm 1 (lines 8–9). They have neither multi-prompt agreement nor an explicit evidence span, a situation that is common for Strategy 3 outputs. To verify whether they are nonetheless supported by x, we apply a surface-level fallback score E_C, computed as a weighted sum of three heuristic features: Subject Grounding, Object Grounding, and Triple Coherence. Respectively, they are weighted by α_C, β_C, and γ_C. The dominant term, *Subject Grounding*, returns 1 if the subject matches a span in the text above the strict threshold τ_s^{fuz}, measured by the normalized Levenshtein ratio [44]; the strict threshold and higher weight reflect that subjects are typically stable named entities, and that a hallucinated subject would corrupt an entire entity node in the resulting graph. The *Object Grounding* feature applies the same test to the object, but against the more relaxed threshold τ_o^{fuz}, since objects often contain free-text descriptions. The final feature, *Triple Coherence* is computed as the Jaccard similarity between the concatenated triple elements and the source text, supplying a weak overall coherence signal. When $E_C \geq \tau_C$, line 9 admits t into the Filter 3 set T_3, closing the cascade.

3.3. Final Knowledge Graph

Let T_i denote the set of triples accepted by Filter i in Stage 2. The final knowledge graph is the union $\hat{G} = T_1 \cup T_2 \cup T_3$. The two stages work complementarily: Stage 1 maximizes recall through diverse prompts, while Stage 2 safeguards precision by demanding grounded evidence at one of three levels. This design separates the task of *finding* candidate triples from the task of *validating* them.

4. Experiments

We design our experiments to answer the following research questions.

- **RQ**$_1$: *How effectively can our framework construct KGs from unstructured text?*
- **RQ**$_2$: *To what extent can our framework mitigate hallucinations during KG construction?*

4.1. Evaluation Setup

Datasets. We evaluate on *Text2KGBench* [14], an ontology-guided benchmark for text-to-KG construction. It contains two datasets: *DBpedia-WebNLG*, comprising 4,860 sentences across 19 ontologies with multiple gold triples per sentence, and *Wikidata-TekGen*, containing 13,474 sentences across 10 Wikidata-derived ontologies with a single annotated triple per sentence. The datasets differ substantially in quality. *Wikidata-TekGen* is constructed through distant supervision and therefore contains incomplete and noisy sentence–triple alignments. For example, a sentence may express multiple relations while only one is annotated, or may omit explicit mentions of entities contained in the associated triple. To reduce this limitation, we enrich the training data before fine-tuning.

Wikidata-TekGen Enrichment. Following prior data-enrichment approaches [31,48, 49,50], we use *Qwen2.5-14B-Instruct* [51] to generate additional ontology-consistent triples. The original gold triple is used as an anchor during generation to preserve factual grounding. Generated triples are retained only if their relations belong to the target ontology.

Baselines. We compare against four open-source baselines that report results on *Text2KGBench*. *Alpaca-LoRA-13B* [14] is an instruction-tuned LLaMA-13B fine-tuned with LoRA [30] on GPT-4-generated data. *Vicuna-13B* [14] is a LLaMA-13B fine-tuned on multi-turn ShareGPT conversations. *T5-large* [52] is a 738M-parameter encoder–decoder model that frames triple extraction as sequence-to-sequence generation. *ChatGLM-6B* [29] is an open bilingual LLM; we adopt its CoT-with-ontology variant, which reports the best results. We exclude Norouzi et al. [11] and ODKE+ [12] from this comparison: although both are conceptually closer to our framework, neither releases full prompts or code, which makes faithful reimplementation infeasible.

Evaluation Metrics. For **RQ**$_1$, we report Precision, Recall, and F1-score using exact matching between normalized predicted and gold triples. Normalization includes lowercasing, removal of non-alphanumeric characters, and whitespace stripping for all labels. Moreover, we also perform underscore replacement and WordNet lemmatization for subject and object labels. For **RQ**$_2$, we use four hallucination-related metrics from

Text2KGBench: Subject Hallucination (SH), Object Hallucination (OH), Ontology Conformance (OC), and Relation Hallucination (RH). SH and OH measure the proportion of extracted subjects and objects unsupported by the source text after stemming and tokenization, defined as $\mathrm{SH} = |\mathrm{HS}|/|\hat{G}|$ and $\mathrm{OH} = |\mathrm{HO}|/|\hat{G}|$, where HS and HO denote hallucinated subject and object sets in the generated graph $\hat{G}$. The OC metric measures the proportion of extracted triples whose relations belong to the target ontology, $\mathrm{OC} = |R_{\mathrm{ont}}|/|\hat{G}|$, where $R_{\mathrm{ont}} \subseteq \hat{G}$. Meanwhile, the RH metric is defined as its complement, $\mathrm{RH} = 1 - \mathrm{OC}$. We highlight that neither OC nor RH metrics validate complete class constraints, domain/range compatibility, or logical consistency.

4.2. Implementation Details

Base Model and Training. We use *LLaMA-3-8B-Instruct* as the base model across all framework components and set decoding temperature to 0 for reproducibility. The model is fine-tuned using 4-bit QLoRA on the official Text2KGBench training split, combining DBpedia-WebNLG training data and enriched Wikidata-TekGen training data. The 95/5 split is used only within this training portion for internal training and validation during model selection. All results reported in Tables 1 and 2 are computed on the separate held-out Text2KGBench test split, which is not used for fine-tuning or validation. All experiments run on a single NVIDIA RTX 3090 GPU (24 GB). All details about the prompts, models, and hyperparameters can be found on the GitHub repository for reproducing our experiments and supporting further research.

Strategy and Validator Configuration. The three prompting strategies use different model configurations. *Tree-of-Thoughts Search* employs the fine-tuned model for candidate generation, while retaining the base model for state evaluation and evidence extraction to avoid self-validation effects. We employ $k = 2$, $D_{max} = 2$, and $\tau = 0.7$ for ToT Search hyperparameters. *Schema-Constrained Open IE* uses the base model only, following prior ontology-guided extraction approaches. *Relaxed Extraction* employs the fine-tuned model because structured generation benefits from task adaptation. Validator thresholds were tuned through grid search on a held-out 5% subset of DBpedia-WebNLG and kept fixed across all experiments. The hyperparameters are set as follows. For *Filter 1*, we set $\tau_s^{anc} = \tau_o^{anc} = 0.9$. For *Filter 2*, we set $\tau_B = 0.9$, $\alpha_B = 0.40$, $\beta_B = \gamma_B = 0.25$, and $\delta_B = 0.10$. For *Filter 3*, we set $\tau_s^{fuz} = 0.8$, $\tau_o^{fuz} = 0.55$, $\alpha_C = 0.50$, $\beta_C = 0.40$, and $\gamma_C = 0.10$.

5. Results and Discussion

We evaluate our framework against four baselines across 29 domains on two benchmarks, addressing RQ_1 (extraction quality) and RQ_2 (hallucination mitigation). Our results consistently demonstrate that ontology-aware, multi-strategy extraction with evidence-driven validation outperforms all baselines in extraction quality and substantially reduces hallucinations across most domains.

5.1. Extraction Quality (RQ_1)

Table 1 reports per-domain Precision, Recall, and F1-score. Our framework achieves the highest F1-score in all 19 DBpedia-WebNLG domains and in 7 of 10 Wikidata-TekGen

Table 1. Comparative performance analysis (Precision, Recall, and F1) across **DBpedia** and **Wikidata** domains ($\mathbf{RQ_1}$). Higher is better. Best values are highlighted in **bold**.

Domain	Alpaca-LoRA			Vicuna			T5-large			ChatGLM			Our Approach		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DBpedia-WebNLG															
University	0.29	0.16	0.20	0.31	0.19	0.23	0.38	0.38	0.38	0.32	0.14	0.19	**0.79**	**0.63**	**0.68**
MusicalWork	0.18	0.15	0.15	0.20	0.18	0.18	0.39	0.40	0.40	0.22	0.15	0.17	**0.59**	**0.56**	**0.56**
Airport	0.28	0.21	0.20	0.32	0.23	0.23	0.40	0.40	0.40	0.23	0.16	0.18	**0.77**	**0.66**	**0.69**
Building	0.35	0.27	0.29	0.48	0.33	0.38	0.40	0.40	0.40	0.40	0.23	0.28	**0.72**	**0.71**	**0.71**
Athlete	0.38	0.30	0.32	0.39	0.26	0.29	0.38	0.39	0.39	0.28	0.22	0.24	**0.79**	**0.76**	**0.76**
Politician	0.39	0.27	0.30	0.39	0.28	0.32	0.39	0.39	0.39	0.29	0.12	0.16	**0.66**	**0.63**	**0.64**
Company	0.35	0.21	0.25	0.48	0.30	0.34	0.35	0.36	0.36	0.40	0.25	0.30	**0.86**	**0.67**	**0.73**
CelestialBody	0.52	0.41	0.47	0.48	0.46	0.46	0.38	0.38	0.38	0.59	0.49	0.52	**0.91**	**0.94**	**0.92**
Astronaut	0.34	0.24	0.27	0.48	0.36	0.40	0.39	0.39	0.39	0.45	0.21	0.27	**0.89**	**0.81**	**0.84**
ComicsCharacter	0.52	0.41	0.45	0.45	0.41	0.42	0.32	0.32	0.32	0.17	0.14	0.15	**0.70**	**0.68**	**0.68**
Transport*	0.13	0.07	0.08	0.14	0.09	0.10	0.40	0.40	0.40	0.33	0.23	0.25	**0.69**	**0.61**	**0.63**
Monument	0.11	0.07	0.08	0.05	0.05	0.05	0.37	0.37	0.37	0.13	0.12	0.12	**0.58**	**0.57**	**0.55**
Food	0.38	0.28	0.31	0.40	0.32	0.34	0.34	0.34	0.34	0.51	0.37	0.41	**0.76**	**0.81**	**0.77**
WrittenWork	0.39	0.35	0.36	0.34	0.34	0.34	0.38	0.38	0.38	0.52	0.31	0.38	**0.90**	**0.90**	**0.90**
SportsTeam	0.41	0.29	0.31	0.40	0.32	0.34	0.34	0.34	0.34	0.16	0.14	0.14	**0.74**	**0.71**	**0.71**
City	0.10	0.09	0.09	0.12	0.12	0.12	0.50	0.50	0.50	0.12	0.09	0.10	**0.61**	**0.62**	**0.60**
Artist	0.35	0.33	0.34	0.52	0.43	0.46	0.39	0.40	0.39	0.30	0.21	0.24	**0.54**	**0.56**	**0.54**
Scientist	0.42	0.33	0.35	0.53	0.43	0.46	0.39	0.40	0.39	0.33	0.23	0.25	**0.81**	**0.80**	**0.80**
Film	0.18	0.15	0.13	0.23	0.14	0.14	0.39	0.39	0.39	0.14	0.09	0.11	**0.45**	**0.44**	**0.44**
Count as Best	0	0	0	0	0	0	0	0	0	0	0	0	**19**	**19**	**19**
Wikidata-TekGen															
Movie	0.28	0.14	0.17	0.33	0.23	0.25	0.27	0.29	0.28	0.34	**0.57**	**0.43**	**0.46**	0.41	0.41
Music	0.33	0.19	0.22	0.42	0.28	0.32	0.29	0.30	0.29	0.29	0.47	0.36	**0.54**	**0.51**	**0.50**
Sport	0.51	0.44	0.45	**0.57**	0.52	0.54	0.50	0.56	0.53	0.21	0.37	0.27	**0.57**	**0.64**	**0.58**
Book	0.25	0.21	0.21	0.35	0.28	0.30	0.35	0.29	0.37	0.32	**0.55**	0.41	**0.42**	0.46	**0.42**
Military	0.21	0.21	0.21	0.24	0.23	0.23	0.39	0.41	0.40	**0.50**	**0.79**	**0.62**	0.49	0.60	0.52
Computer	0.29	0.27	0.27	0.28	0.25	0.26	0.33	0.34	0.33	0.33	**0.57**	0.42	**0.48**	0.51	**0.48**
Space	0.58	0.55	0.52	0.67	0.63	0.63	0.52	0.54	0.53	0.53	0.74	0.62	**0.76**	**0.78**	**0.76**
Politics	0.22	0.21	0.21	0.32	0.32	0.32	0.42	0.42	0.42	0.27	0.39	0.32	**0.59**	**0.65**	**0.60**
Nature	0.15	0.15	0.15	0.27	0.24	0.25	0.42	0.53	0.47	0.22	0.29	0.25	**0.54**	**0.59**	**0.55**
Culture	0.15	0.16	0.15	0.31	0.31	0.31	**0.45**	0.45	**0.45**	0.32	0.32	0.32	0.42	**0.48**	0.44
Count as Best	0	0	0	1	0	0	1	0	1	1	4	2	**8**	6	7

domains. The largest margins arise where the ontology declares a rich predicate set and fine-grained class types: *CelestialBody* (F1 = 0.92 vs. 0.52 for ChatGLM, the strongest baseline), *WrittenWork* (0.90 vs. 0.38 for T5-Large), *Astronaut* (0.84 vs. 0.40 for Vicuna), *Food* (0.77 vs. 0.41 for ChatGLM), and *Airport* (0.69 vs. 0.40 for T5-Large). In each of these domains, our F1 exceeds twice that of the best baseline—a pattern directly attributable to ontology grounding. When the schema provides discriminative domain and range information, our framework uses it to guide candidate generation and vali-

dation, differently from the baselines, which do not include an explicit evidence-driven validation stage. On Wikidata-TekGen, the framework still wins in most domains, but with smaller margins. The strongest gains here are in *Space* (F1 = 0.76), *Politics* (0.60), *Nature* (0.55), and *Sport* (0.58). We attribute this to the distant supervision artifacts described in Section 4.1, which limit the achievable F1 on this benchmark. All results are obtained using *LLaMA-3-8B-Instruct* (8B parameters), while *Alpaca-LoRA* and *Vicuna* both use 13B parameters. Ontology-aware prompting, therefore, appears to provide an inductive bias that partially offsets the reduced model scale.

5.2. Hallucination Mitigation (RQ_2)

Table 2 reports the hallucination-related metrics OC, SH, RH, and OH. We compare only against *Alpaca-LoRA-13B* and *Vicuna-13B* baselines, since published *Text2KGBench* evaluations for *T5-large* and *ChatGLM-6B* do not report these metrics. On DBpedia-WebNLG, OC is consistently high, reaching 0.97–1.00 in domains such as *University*, *WrittenWork*, *SportsTeam*, and *CelestialBody*, with RH close to zero. Two design choices contribute to this outcome: *Schema-Constrained Open IE* restricts the model's relation vocabulary to ontology predicates, while *Filter 2* of the validator accepts a triple when the evidence span explicitly contains the predicted relation. OH decreases substantially even in surface-ambiguous domains, particularly for *City* (OH decreases from 0.68 under *Alpaca-LoRA* to 0.26) and *Artist* (from 0.26 to 0.09). This effect is largely driven by the joint-presence term ($\alpha_B = 0.40$), where an object is added to $\hat{G}$ only if it co-occurs with its subject and relation within the cited evidence span, preventing generation plausibility alone from validating a triple. Wikidata-TekGen exhibits less consistent behavior across domains. OC remains high and OH remains low in well-defined domains (*Politics*, *Culture*, *Music*). However, RH increases in domains with sparse or ambiguous ontology coverage, particularly *Space*. Examination of failure cases shows that source sentences contain relations not represented in the underlying Wikidata-derived ontology; consequently, the evaluator labels these otherwise valid extractions as RH. This effect is largely attributable to benchmark limitations rather than systematic failures of the framework, and explains the remaining RH observed on this dataset.

The main exception on DBpedia-WebNLG is *MusicalWork*, where SH increases to 0.54 (*Alpaca-LoRA*: 0.32, *Vicuna*: 0.32). This apparent degradation is due to the *MusicalWork* structural characteristics, which contain several fine-grained subject types (e.g., `dbo:MusicalWork`, `dbo:Album`, `dbo:Single`). Text2KGBench computes SH using exact type matching against these fine-grained classes. Our framework may predict a semantically correct sibling class for the domain, which remains ontologically plausible but is counted as a hallucination. In contrast, fine-tuned baselines such as *Alpaca-LoRA* and *Vicuna* generate fewer distinct entity-type predictions and therefore incur fewer type mismatches. Importantly, the high OC (0.97) and low RH (0.03) for *MusicalWork* indicate that relation extraction and overall OC remain strong, while only fine-grained subject-type resolution is affected.

5.3. Ablation Study (RQ_1 and RQ_2)

Table 3 reports an ablation of the three prompting strategies and the hierarchical validator. Configurations are denoted by the strategies included: `T` = *Tree-of-Thoughts Search*,

Table 2. Performance evaluation of all models in mitigating hallucinations, averaged per domain ($\mathbf{RQ}_2$). Best values per metric are highlighted in **bold**. OC ↑: higher is better. SH, RH, OH ↓: lower is better.

Domain	ALPACA-LORA				VICUNA				**Our Approach**			
	OC	SH	RH	OH	OC	SH	RH	OH	OC	SH	RH	OH
DBpedia-WebNLG												
University	0.89	0.13	0.11	0.26	0.92	0.11	0.08	0.21	**0.97**	**0.00**	**0.03**	**0.02**
MusicalWork	0.85	**0.32**	0.15	0.33	0.89	**0.32**	0.11	0.27	**0.97**	0.54	**0.03**	**0.23**
Airport	0.94	0.12	0.06	0.45	0.92	**0.03**	0.08	0.27	**0.97**	0.04	**0.03**	**0.16**
Building	0.93	0.17	0.07	0.32	**0.98**	**0.02**	**0.02**	0.22	**0.98**	**0.02**	**0.02**	**0.07**
Athlete	0.94	0.11	0.06	0.27	0.92	**0.01**	0.08	**0.13**	**0.97**	0.11	**0.03**	**0.13**
Politician	0.92	0.15	0.08	0.38	0.89	**0.11**	0.11	0.29	**0.97**	0.16	**0.03**	**0.19**
Company	0.95	0.17	0.05	0.44	**1.00**	0.09	**0.00**	0.36	0.97	**0.01**	0.03	**0.15**
CelestialBody	0.96	0.12	0.04	0.49	0.97	0.05	0.03	0.46	**1.00**	**0.01**	**0.00**	**0.34**
Astronaut	0.87	0.08	0.13	0.58	0.87	0.06	0.13	**0.28**	**0.98**	**0.00**	**0.02**	**0.28**
ComicsChar.	0.93	0.62	0.07	0.42	0.97	0.64	0.03	0.28	**0.98**	**0.41**	**0.02**	**0.08**
Transport*	0.97	0.18	0.03	0.36	0.94	**0.14**	0.06	0.41	**0.98**	0.17	**0.02**	**0.16**
Monument	0.97	0.13	0.03	**0.31**	0.94	0.18	0.06	**0.31**	**1.00**	**0.00**	**0.00**	**0.31**
Food	0.92	0.12	0.08	0.26	0.94	**0.05**	0.06	0.20	**0.99**	0.06	**0.01**	**0.11**
WrittenWork	0.90	0.20	0.10	0.50	0.92	0.12	0.08	0.33	**0.99**	**0.05**	**0.01**	**0.15**
SportsTeam	0.90	0.11	0.10	0.24	0.91	0.04	0.09	0.11	**0.99**	**0.03**	**0.01**	**0.08**
City	0.96	0.07	0.04	0.68	**0.98**	**0.04**	**0.02**	0.67	0.95	0.28	0.05	**0.26**
Artist	0.83	0.07	0.17	0.26	0.89	**0.03**	0.11	0.13	**0.99**	0.37	**0.01**	**0.09**
Scientist	0.92	0.14	0.08	0.43	0.95	0.05	0.05	0.30	**0.97**	**0.02**	**0.03**	**0.23**
Film	0.82	0.08	0.18	0.33	0.94	0.30	0.06	0.19	**0.98**	**0.01**	**0.02**	**0.14**
Count as Best	0	1	0	1	3	8	3	3	**16**	**10**	**16**	**17**
Wikidata-TekGen												
Movie	**0.92**	0.25	**0.08**	0.24	0.89	0.26	0.11	0.26	0.90	**0.23**	0.10	**0.06**
Music	0.91	0.18	0.09	0.24	**0.94**	0.16	**0.06**	0.22	**0.94**	**0.13**	**0.06**	**0.04**
Sport	**0.96**	0.18	**0.04**	0.11	0.85	0.22	0.15	0.13	0.90	**0.14**	0.10	**0.06**
Book	**0.94**	0.17	**0.06**	0.19	0.92	0.16	0.08	0.23	0.88	**0.12**	0.12	**0.05**
Military	**0.93**	0.21	**0.07**	0.26	0.80	0.19	0.20	0.26	0.87	**0.16**	0.13	**0.08**
Computer	0.83	0.17	0.17	0.13	**0.85**	0.15	**0.15**	0.11	0.81	**0.13**	0.19	**0.02**
Space	0.90	0.14	0.10	0.10	**0.93**	0.15	**0.07**	0.08	0.79	**0.09**	0.21	**0.04**
Politics	0.90	0.19	0.10	0.15	0.92	0.17	0.08	0.15	**0.95**	**0.15**	**0.05**	**0.04**
Nature	**0.93**	0.22	0.07	0.20	0.68	0.10	**0.04**	0.14	0.91	**0.08**	0.09	**0.06**
Culture	0.54	0.16	0.46	0.14	0.59	0.15	0.39	0.12	**0.83**	**0.11**	**0.17**	**0.06**
Count as Best	**4**	0	**4**	0	3	1	**4**	0	**4**	**10**	3	**10**

S = *Schema-Constrained Open IE*, and R = *Relaxed Extraction*. The suffix *w/o Evaluator* indicates that the candidate pool is added directly to $\hat{G}$ without the three-filter validator, while *Full* denotes the complete pipeline.

Effect of the Validator. The validator's contribution is isolated by comparing *Full* with *TSR w/o Evaluator*, since both use all three prompting strategies. On DBpedia-WebNLG, the validator improves F1 from 0.64 to 0.69 and OC from 0.92 to 0.98, and reduces RH

Table 3. Average F1, OC, SH, RH, and OH scores across ablation variants of our approach. Full denotes the complete pipeline, including the evaluator. The remaining variants use subsets of prompting strategies: *T* (Tree-of-Thoughts), *S* (Schema-constrained Open IE), and *R* (Relaxed Extraction). Best results are highlighted in **bold**.

Variation	DBpedia-WebNLG					Wikidata-TekGen				
	F1 ↑	OC ↑	SH ↓	RH ↓	OH ↓	F1 ↑	OC ↑	SH ↓	RH ↓	OH ↓
Full	**0.69**	0.98	0.12	0.02	0.16	**0.54**	0.91	0.17	0.09	**0.05**
T w/o Evaluator	0.65	**0.99**	0.13	**0.01**	0.20	**0.54**	**0.94**	0.17	**0.06**	0.12
S w/o Evaluator	0.42	0.86	**0.05**	0.14	**0.11**	0.29	0.77	0.02	0.23	0.13
R w/o Evaluator	0.66	**0.99**	0.11	**0.01**	0.20	0.50	**0.94**	0.17	**0.06**	0.10
TS w/o Evaluator	0.62	0.91	0.09	0.09	0.17	0.51	0.83	**0.01**	0.17	0.14
TR w/o Evaluator	0.68	**0.99**	0.12	**0.01**	0.20	0.50	0.93	0.17	0.07	0.13
SR w/o Evaluator	0.62	0.91	0.08	0.09	0.17	0.48	0.81	0.08	0.19	0.14
TSR w/o Evaluator	0.64	0.92	0.09	0.08	0.18	0.52	0.83	0.10	0.14	0.15

from 0.08 to 0.02. Similar gains on Wikidata-TekGen, where F1 increases from 0.52 to 0.54, OC from 0.83 to 0.91, and OH decreases substantially from 0.15 to 0.05. These results show that the validator converts a diverse but noisy candidate pool into a graph with stronger ontology alignment and fewer hallucinations. In contrast, SH increases slightly after introducing the validator (DBpedia-WebNLG: $0.09 \rightarrow 0.12$; Wikidata-TekGen: $0.10 \rightarrow 0.17$). We interpret this as a recall–precision trade-off induced by *Filter 3*. The validator preserves additional triples that would otherwise be discarded, some of which rely on fuzzy subject matching. Although this introduces more subject mismatches, the overall effect remains beneficial because F1 and OC improve simultaneously.

Strategies in Isolation. Among individual strategies, `T` performs best, achieving F1 scores of 0.65 on DBpedia-WebNLG and 0.54 on Wikidata-TekGen, with corresponding OC values of 0.99 and 0.94. This suggests that the iterative ontology-guided search effectively filters low-quality candidates during generation. `R` also performs competitively (0.66 and 0.50), indicating that relaxed output constraints improve recall by recovering otherwise discarded triples. In contrast, `S` performs worst in isolation (0.42 and 0.29) and exhibits the highest RH values, suggesting that schema constraints alone are insufficient without downstream validation.

Strategy Combinations. The interaction between strategies further highlights their complementary roles. On DBpedia-WebNLG, *TR w/o Evaluator* achieves F1 $= 0.68$, approaching the full system (0.69), indicating that `T` contributes precision while `R` improves coverage. Adding `S` without the validator does not improve performance, as additional candidates primarily introduce noise. On Wikidata-TekGen, *TSR w/o Evaluator* performs best among non-validator configurations (F1 $= 0.52$), but still falls below the complete pipeline in ontology conformance (0.83 vs. 0.91). These results suggest that prompt diversity alone can improve recall, whereas the validator is required to convert this diversity into reliable graph construction.

Summary. Overall, `T` is the primary contributor to extraction quality, `R` provides additional recall, and `S` contributes diversity that becomes beneficial only when combined

with structured validation. Together, these findings answer RQ_1 and RQ_2: prompt diversity increases candidate coverage, while evidence-driven validation transforms this diversity into a reliable knowledge graph.

5.4. Limitations

We acknowledge three limitations of this work. First, performance depends on the quality and completeness of the supplied ontology, and robustness to incomplete or noisy ontologies remains untested. Performance may also vary with the model, prompt, decoding settings, ontology structure, and dataset. A broader robustness study across model families, prompt variants, and ontology types is complementary to quantify the robustness. Second, both *Text2KGBench* datasets contain independent short sentences, limiting conclusions regarding longer contexts such as paragraphs or documents. Third, the generated knowledge graph is not linked to external resources such as DBpedia or Wikidata. Integrating entity linking could enable bidirectional enrichment with existing KGs.

6. Conclusions and Future Work

We presented an ontology-aware, multi-prompt framework for RDF triple extraction that decouples candidate generation from validation. Three prompts balance precision, schema adherence, and recall, while a three-filter, evidence-driven validator admits only triples supported by cross-prompt agreement, evidence spans, or surface grounding. On the *Text2KGBench* benchmark across 29 ontologies, the framework outperforms four baselines in F1-score on all DBpedia-WebNLG domains and most Wikidata-TekGen domains, despite using an 8B-parameter backbone. It substantially reduces object and relation hallucinations while improving ontology conformance, and ablation results identify the validator as the key component that transforms prompt diversity into a high-precision graph. Future work will evaluate the framework's robustness to ontology quality, assess its performance under longer contexts, such as paragraphs and full documents, and integrate entity linking with public knowledge graphs, including DBpedia and Wikidata.

Acknowledgments

This work has been supported by the German Federal Ministry of Research, Technology and Space (BMFTR) within the project KI-Akademie OWL under the grant no. 16IS24057B, by the Federal Ministry for Economic Affairs and Energy (BMWE) on the basis of a decision by the German Bundestag within the project ikDS under the grant no. KK5175206LO4, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the project TRR 318/3 2026 under the grant no. 438445824.

Declaration on Generative AI

During the preparation of this work, the authors used *ChatGPT*, *Gemini*, and *Grammarly* for grammar and spelling correction, paraphrasing, and linguistic polishing of human-written text. The authors carefully reviewed and edited the content and take full responsibility for the integrity and content of the publication.

References

[1] Ayvaz S, Aydar M. Using RDF Summary Graph For Keyword-based Semantic Searches. CoRR. 2017;abs/1707.03602. Available from: http://arxiv.org/abs/1707.03602.

[2] Allemang D, Sequeda J. Increasing the LLM Accuracy for Question Answering: Ontologies to the Rescue! CoRR. 2024;abs/2405.11706. Available from: https://doi.org/10.48550/arXiv.2405.11706.

[3] Xu S, Hua W, Zhang Y. OpenP5: An Open-Source Platform for Developing, Training, and Evaluating LLM-based Recommender Systems. In: Yang GH, Wang H, Han S, Hauff C, Zuccon G, Zhang Y, editors. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024. ACM; 2024. p. 386-94. Available from: https://doi.org/10.1145/3626772.3657883.

[4] Speck R, Ngomo AN. Ensemble Learning of Named Entity Recognition Algorithms using Multilayer Perceptron for the Multilingual Web of Data. In: Corcho Ó, Janowicz K, Rizzo G, Tiddi I, Garijo D, editors. Proceedings of the Knowledge Capture Conference, K-CAP 2017, Austin, TX, USA, December 4-6, 2017. ACM; 2017. p. 26:1-26:4. Available from: https://doi.org/10.1145/3148011.3154471.

[5] Giorgi JM, Wang X, Sahar N, Shin WY, Bader GD, Wang B. End-to-end Named Entity Recognition and Relation Extraction using Pre-trained Language Models. CoRR. 2019;abs/1912.13415. Available from: http://arxiv.org/abs/1912.13415.

[6] Jiang Y, Huang K, Zhang X, Xu W, Liang Z, Yang Y, et al. A BERT-Assisted LLM Framework for Knowledge Graph Construction. In: 31th IEEE International Conference on Parallel and Distributed Systems, ICPADS 2025, Hefei, China, December 14-18, 2025. IEEE; 2025. p. 1-10. Available from: https://doi.org/10.1109/ICPADS67057.2025.11323011.

[7] Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, et al. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models (CRFM). 2023. Available from: https://crfm.stanford.edu/2023/03/13/alpaca.html.

[8] Manakul P, Liusie A, Gales MJF. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In: Bouamor H, Pino J, Bali K, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023. Association for Computational Linguistics; 2023. p. 9004-17. Available from: https://doi.org/10.18653/v1/2023.emnlp-main.557.

[9] Bosselut A, Rashkin H, Sap M, Malaviya C, Celikyilmaz A, Choi Y. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In: Korhonen A, Traum DR, Màrquez L, editors. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics; 2019. p. 4762-79. Available from: https://doi.org/10.18653/v1/p19-1470.

[10] Zhu Y, Wang X, Chen J, Qiao S, Ou Y, Yao Y, et al. LLMs for knowledge graph construction and reasoning: recent capabilities and future opportunities. World Wide Web (WWW). 2024;27(5):58. Available from: https://doi.org/10.1007/s11280-024-01297-w.

[11] Norouzi SS, Barua A, Christou A, Gautam N, Eells A, Hitzler P, et al. Ontology Population using LLMs. CoRR. 2024;abs/2411.01612. Available from: https://doi.org/10.48550/arXiv.2411.01612.

[12] Khorshidi S, Nikfarjam A, Shankar S, Sang Y, Govind Y, Jang H, et al. ODKE+: Ontology-Guided Open-Domain Knowledge Extraction with LLMs. CoRR. 2025;abs/2509.04696. Available from: https://doi.org/10.48550/arXiv.2509.04696.

[13] Aljabari A, Hamad N, Khalilia M, Jarrar M. WojoodOntology: Ontology-Driven LLM Prompting for Unified Information Extraction Tasks. In: Proceedings of The Third Arabic Natural Language Processing Conference; 2025. p. 179-93.

[14] Mihindukulasooriya N, Tiwari S, Enguix CF, Lata K. Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text. CoRR. 2023;abs/2308.02357. Available from: https://doi.org/10.48550/arXiv.2308.02357.

[15] Liu K, Chen Y, Liu J, Zuo X, Zhao J. Extracting Events and Their Relations from Texts: A Survey on Recent Research Progress and Challenges. AI Open. 2020;1:22-39. Available from: https://doi.org/10.1016/j.aiopen.2021.02.004.

[16] Shen W, Wang J, Han J. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. IEEE Trans Knowl Data Eng. 2015;27(2):443-60. Available from: https://doi.org/10.1109/TKDE.2014.2327028.

[17] Etzioni O, Cafarella MJ, Downey D, Popescu A, Shaked T, Soderland S, et al. Unsupervised named-entity extraction from the Web: An experimental study. Artif Intell. 2005;165(1):91-134. Available from: https://doi.org/10.1016/j.artint.2005.03.001.

[18] Jurafsky D, Martin JH. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International; 2009. Available from: https://www.worldcat.org/oclc/315913020.

[19] Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In: Su K, Su J, Wiebe J, editors. ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore. The Association for Computer Linguistics; 2009. p. 1003-11. Available from: https://aclanthology.org/P09-1113/.

[20] Zhang Y, Zhong V, Chen D, Angeli G, Manning CD. Position-aware Attention and Supervised Data Improve Slot Filling. In: Palmer M, Hwa R, Riedel S, editors. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017. Association for Computational Linguistics; 2017. p. 35-45. Available from: https://doi.org/10.18653/v1/d17-1004.

[21] Banko M, Cafarella MJ, Soderland S, Broadhead M, Etzioni O. Open Information Extraction from the Web. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI); 2007. p. 2670-6. Available from: http://ijcai.org/Proceedings/07/Papers/429.pdf.

[22] Stanovsky G, Michael J, Zettlemoyer L, Dagan I. Supervised Open Information Extraction. In: Walker MA, Ji H, Stent A, editors. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers). Association for Computational Linguistics; 2018. p. 885-95. Available from: https://doi.org/10.18653/v1/n18-1081.

[23] Zhigang Zhao MC Xiong Luo, Ma L. A Survey of Knowledge Graph Construction Using Machine Learning. Computer Modeling in Engineering & Sciences. 2024;139(1):225-57. Available from: http://www.techscience.com/CMES/v139n1/55119.

[24] Hofer M, Obraczka D, Saeedi A, Köpcke H, Rahm E. Construction of Knowledge Graphs: Current State and Challenges. Inf. 2024;15(8):509. Available from: https://doi.org/10.3390/info15080509.

[25] Wei X, Cui X, Cheng N, Wang X, Zhang X, Huang S, et al. ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT. arXiv preprint arXiv:230210205. 2024. Available from: https://arxiv.org/abs/2302.10205.

[26] Xue L, Zhang D, Dong Y, Tang J. AutoRE: Document-Level Relation Extraction with Large Language Models. In: Cao Y, Feng Y, Xiong D, editors. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), ACL 2024, Bangkok, Thailand, August 11-16, 2024. Association for Computational Linguistics; 2024. p. 211-20. Available from: https://doi.org/10.18653/v1/2024.acl-demos.20.

[27] Papaluca A, Krefl D, Rodriguez SM, Lenskiy A, Suominen H. Zero- and Few-Shots Knowledge Graph Triplet Extraction with Large Language Models. CoRR. 2023;abs/2312.01954. Available from: https://doi.org/10.48550/arXiv.2312.01954.

[28] Mo B, Yu K, Kazdan J, Mpala P, Yu L, Cundy C, et al. KGGen: Extracting Knowledge Graphs from Plain Text with Language Models. CoRR. 2025;abs/2502.09956. Available from: https://doi.org/10.48550/arXiv.2502.09956.

[29] Nie J, Hou X, Song W, Wang X, Jin X, Zhang X, et al. Knowledge Graph Efficient Construction: Embedding Chain-of-Thought into LLMs. In: Proceedings of Workshops at the 50th International Conference on Very Large Data Bases, VLDB 2024, Guangzhou, China, August 26-30, 2024. VLDB.org; 2024. p. 8097. Available from: https://vldb.org/workshops/2024/proceedings/LLM+KG/LLM+KG-4.pdf.

[30] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net; 2022. p. 3. Available from: https://openreview.net/forum?id=nZeVKeeFYf9.

[31] Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, et al.. Stanford alpaca: An instruction-following llama model. Stanford, CA, USA; 2023.

[32] Chiang WL, Li Z, Lin Z, Sheng Y, Wu Z, Zhang H, et al. Vicuna: An Open-Source Chatbot Impress-

ing GPT-4 with 90Quality. LMSYS Org Blog. 2023. Available from: https://lmsys.org/blog/2023-03-30-vicuna/.

[33] Lairgi Y, Moncla L, Cazabet R, Benabdeslem K, Cléau P. iText2KG: Incremental Knowledge Graphs Construction Using Large Language Models. In: Barhamgi M, Wang H, Wang X, editors. Web Information Systems Engineering - WISE 2024 - 25th International Conference, Doha, Qatar, December 2-5, 2024, Proceedings, Part IV. Lecture Notes in Computer Science. Springer; 2024. p. 214-29. Available from: https://doi.org/10.1007/978-981-96-0573-6_16.

[34] Lu Y, Wang J. KARMA: Leveraging Multi-Agent LLMs for Automated Knowledge Graph Enrichment. CoRR. 2025;abs/2502.06472. Available from: https://doi.org/10.48550/arXiv.2502.06472.

[35] Kommineni VK, König-Ries B, Samuel S. Towards the Automation of Knowledge Graph Construction using Large Language Models. In: Vakaj E, Iranmanesh S, Rizou S, Mihindukulasooriya N, Tiwari S, Ortiz-Rodríguez F, et al., editors. Proceedings of the 3rd International Workshop on Natural Language Processing for Knowledge Graph Creation co-located with 20th International Conference on Semantic Systems (SEMANTiCS 2024), Amsterdam, The Netherlands, September 17, 2024. CEUR Workshop Proceedings. CEUR-WS.org; 2024. p. 19-34. Available from: https://ceur-ws.org/Vol-3874/paper2.pdf.

[36] Feng X, Wu X, Meng H. Ontology-grounded Automatic Knowledge Graph Construction by LLM under Wikidata schema. In: Ciravegna G, Zarlenga ME, Barbiero P, Shams Z, Giannini F, Garreau D, et al., editors. Proceedings of the KDD Workshop on Human-Interpretable AI 2024 co-located with 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024), Centre de Convencions Internacional de Barcelona, Spain, August 26, 2024. CEUR Workshop Proceedings. CEUR-WS.org; 2024. p. 117-35. Available from: https://ceur-ws.org/Vol-3841/Paper19.pdf.

[37] Trisedya BD, Weikum G, Qi J, Zhang R. Neural Relation Extraction for Knowledge Base Enrichment. In: Korhonen A, Traum DR, Màrquez L, editors. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics; 2019. p. 229-40. Available from: https://doi.org/10.18653/v1/p19-1023.

[38] Luan Y, He L, Ostendorf M, Hajishirzi H. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, editors. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. Association for Computational Linguistics; 2018. p. 3219-32. Available from: https://doi.org/10.18653/v1/d18-1360.

[39] Castro Ferreira T, Gardent C, Ilinykh N, van der Lee C, Mille S, Moussallem D, et al. The 2020 Bilingual, Bi-Directional WebNLG+ Shared Task: Overview and Evaluation Results (WebNLG+ 2020). In: Castro Ferreira T, Gardent C, Ilinykh N, van der Lee C, Mille S, Moussallem D, et al., editors. Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+). Dublin, Ireland (Virtual): Association for Computational Linguistics; 2020. p. 55-76. Available from: https://aclanthology.org/2020.webnlg-1.7/.

[40] Bhattarai K, Oh IY, Abrams ZB, Lai AM. Document-level Clinical Entity and Relation extraction via Knowledge Base-Guided Generation. In: Demner-Fushman D, Ananiadou S, Miwa M, Roberts K, Tsujii J, editors. Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, BioNLP@ACL 2024, Bangkok, Thailand, August 16, 2024. Association for Computational Linguistics; 2024. p. 318-27. Available from: https://doi.org/10.18653/v1/2024.bionlp-1.24.

[41] Stubbs A, Filannino M, Soysal E, Henry S, Uzuner Ö. Cohort selection for clinical trials: n2c2 2018 shared task track 1. J Am Medical Informatics Assoc. 2019;26(11):1163-71. Available from: https://doi.org/10.1093/jamia/ocz163.

[42] Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Informatics. 2012;45(5):885-92. Available from: https://doi.org/10.1016/j.jbi.2012.04.008.

[43] Yao S, Yu D, Zhao J, Shafran I, Griffiths T, Cao Y, et al. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023; 2023. p. 11809-22. Available from: http://papers.nips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

[44] Yujian L, Bo L. A normalized Levenshtein distance metric. IEEE transactions on pattern analysis and machine intelligence. 2007;29(6):1091-5. Available from: https://doi.org/10.1109/TPAMI.2007.1078.

[45] Wang X, Wei J, Schuurmans D, Le QV, Chi EH, Narang S, et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net; 2023. Available from: https://openreview.net/forum?id=1PL1NIMMrw.

[46] Jaccard P. The distribution of the flora in the alpine zone. New phytologist. 1912;11(2):37-50.

[47] Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); 2018. p. 809-19. Available from: https://doi.org/10.18653/v1/n18-1074.

[48] Wang Y, Kordi Y, Mishra S, Liu A, Smith NA, Khashabi D, et al. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In: Rogers A, Boyd-Graber JL, Okazaki N, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023. Association for Computational Linguistics; 2023. p. 13484-508. Available from: https://doi.org/10.18653/v1/2023.acl-long.754.

[49] Chen L, Li S, Yan J, Wang H, Gunaratna K, Yadav V, et al. AlpaGasus: Training a Better Alpaca with Fewer Data. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net; 2024. p. 34767-97. Available from: https://openreview.net/forum?id=FdVXgSJhvz.

[50] Shirgaonkar A, Pandey N, Abay NC, Aktas T, Aski V. Knowledge Distillation Using Frontier Open-source LLMs: Generalizability and the Role of Synthetic Data. CoRR. 2024;abs/2410.18588. Available from: https://doi.org/10.48550/arXiv.2410.18588.

[51] Yang A, Yu B, Li C, Liu D, Huang F, Huang H, et al. Qwen2.5-1M Technical Report. CoRR. 2025;abs/2501.15383. Available from: https://doi.org/10.48550/arXiv.2501.15383.

[52] Meng Z, Zhan S, Xu R, Mayer W, Zhu Y, Zhang HY, et al. Domain Ontology-Driven Knowledge Graph Generation from Text. ACM Trans Probab Mach Learn. 2025 Dec;1(4). Available from: https://doi.org/10.1145/3708478.