

CiteAlign-Table: Verifying Citation Faithfulness when LLMs Attribute to Structured Data

Ashish Gupta
ashish.gupta1@servicenow.com
ServiceNow
Hyderabad, India

Manjunatha Naik MC
manjunathanaik.mc@servicenow.com
Servicenow
Hyderabad, India

Abstract

Attributed Question Answering systems increasingly ground answers in tabular structured corpora—Wikipedia infoboxes and relational tables, the dominant structured evidence in open-domain RAG—yet citation faithfulness methods remain designed for text passages. This mismatch allows LLMs to cite the correct table but the wrong row, silently aggregate values without justification, or attribute claims to schema fields that do not semantically entail them. We formalize multiple structural hallucination modes specific to structured-data attribution and introduce CiteAlign-Table, a hybrid neuro-symbolic verifier that checks schema-path validity, cell-level semantic alignment, and arithmetic consistency between claims and their cited records. A schema-aware entailment module detects entity and content misalignments, while a lightweight program synthesizer verifies numerical aggregation, temporal consistency, and unit correctness. Experiments on a hybrid benchmark of Wikipedia infoboxes and WikiTables show that CiteAlign-Table achieves 0.84 F1 in detecting unfaithful structured citations, while text-only baselines degrade to 0.61 F1. By extending citation faithfulness from passages to structured records, our work makes attributed QA trustworthy over the heterogeneous data that production RAG systems actually retrieve.

CCS Concepts

• Computing methodologies → Intelligent agents; Natural language processing.

Keywords

Retrieval Augmented Generation, Structured Data Retrieval, Hallucination Detection, Neuro-Symbolic Methods, Citation Faithfulness

ACM Reference Format:

Ashish Gupta and Manjunatha Naik MC. 2026. CiteAlign-Table: Verifying Citation Faithfulness when LLMs Attribute to Structured Data. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26)*, November 07–11, 2026, Rome, Italy. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3799682.3840003>

1 Introduction

Large Language Models (LLMs) frequently hallucinate or rely on stale parametric knowledge when answering knowledge-intensive questions. Retrieval-Augmented Generation (RAG) mitigates this

by conditioning the generator on retrieved context passages [15]. Recent systems augment answers with inline citations pointing to source documents, enabling users to verify factual claims [1, 18]. This paradigm, Attributed QA, has become standard in production search engines and conversational agents.

Despite its appeal, attributed QA evaluation and verification remain designed for text passages. Benchmarks such as ALCE [5] and subsequent fine-grained citation evaluation frameworks [9] assume that a citation is faithful if the cited passage appears in the retrieval context. However, production RAG systems increasingly ground answers in structured corpora—Wikipedia infoboxes, product databases, and JSON records—where faithfulness requires more than textual overlap. A claim such as "India's GDP in 2023 was \$3.7 trillion" where the LLM can cite the correct table while selecting the wrong row, silently aggregating values without justification, or attributing a claim to a schema field that does not semantically entail it [13]. These structural citation hallucinations undermine user trust and evade detection by existing passage-oriented validators.

Prior work on table-text alignment has addressed cell-level explanation for scientific claim verification [8], but treats tables as evidence for human fact-checkers rather than as retrieved sources within an LLM pipeline. Similarly, general hallucination detection methods focus on answer-level consistency or free-text generation [4, 16], while recent citation verifiers for RAG such as VeriCite [20] assume text-based evidence. What is missing is a verification mechanism that checks whether an LLM's inline citation to a structured record actually points to the correct schema path, cell value, and arithmetic derivation.

Unlike TabFact [3], which evaluates binary claim-table entailment, our benchmark evaluates whether LLM-generated citations faithfully point to the correct cell region. We introduce CiteAlign-Table (Figure 1), a hybrid neuro-symbolic verifier that extends citation faithfulness from passages to structured data. CiteAlign-Table operates in three layers: (1) a symbolic schema-path validator that resolves the citation against the retrieved table and detects non-existent or schema-mismatched references; (2) a schema-aware entailment module that checks whether the cited cell semantically supports the claim; and (3) a lightweight program synthesizer that generates executable verification routines to check numerical aggregation, temporal consistency, and unit correctness. By combining deterministic structural checks with neural semantic alignment and neuro-symbolic arithmetic verification, CiteAlign-Table catches failure modes that text-only methods miss entirely.

The main contributions of this work are:

- We formalize multiple structural hallucination modes specific to structured-data attribution in RAG.



This work is licensed under a Creative Commons Attribution 4.0 International License. *CIKM '26, Rome, Italy*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2539-5/2026/11
<https://doi.org/10.1145/3799682.3840003>

- We propose CiteAlign-Table, a hybrid neuro-symbolic verifier combining exact schema validation, schema-aware neural entailment, and lightweight program synthesis
- a continuous faithfulness score enabling threshold-tuned precision-recall trade-offs

Experiments on a hybrid benchmark of Wikipedia infoboxes and WikiTables demonstrate that CiteAlign-Table achieves 0.86 F1 in detecting unfaithful structured citations, while text-only baselines degrade to 0.61 F1. Our work makes attributed QA trustworthy over the heterogeneous structured data that production RAG systems actually retrieve.

2 Related Work

Attributed QA and Citation Metrics: [5] introduce a framework for LLM citation generation and propose ALCE, a benchmark for attributed language model citation evaluation. Subsequent fine-grained evaluation frameworks [9] provide improved metrics for citation quality. However, all of this work assumes citations point to text passages, conflating passage presence with semantic support. None addresses whether an LLM citation to a structured record is faithful at the cell or schema level.

Hallucination Detection in LLMs: Self-checking methods prompt the LLM to verify its own outputs [4, 16], but suffer from the same unreliability that causes hallucinations. These approaches are designed for free-text generation and collapse when the source is a table or infobox, as they lack mechanisms to verify schema paths, numerical aggregation, or cell-level entailment.

Table Reasoning and Verification: Table question answering and fact verification methods have addressed cell-level explanation and reasoning over structured tables. However, they treat tables as static evidence for human fact-checkers rather than as dynamically retrieved sources in an LLM pipeline. Consequently, they do not verify whether a generated citation actually points to the correct schema path, whether the cited cell semantically entails the claim, or whether a numerical derivation is arithmetically valid.

Neuro-Symbolic Verification: CiteAlign-Table extends this philosophy to the citation faithfulness problem: we use symbolic checks for schema-path validity, neural entailment for semantic alignment, and lightweight program synthesis for arithmetic verification. This hybrid design is essential because structured citations require exact schema matching and precise numerical derivation—capabilities that pure neural or pure symbolic methods lack in isolation.

3 Methodology

3.1 Problem Formulation

An attributed QA system produces an answer consisting of atomic claims $C = \{c_1, c_2, \dots, c_n\}$. Each claim c_i is annotated with a citation string τ_i (e.g., [Table: GDP_2023, Row: India, Col: GDP]), pointing to a record in a retrieved corpus R . A retriever $R(q)$ returns a ranked list of structured records $T = \{T_1, \dots, T_K\}$ (tables and infoboxes) for query q . A citation parser extracts a structured path $\phi = (t, r, \gamma)$ from τ , where t is the table identifier, r is the row key, and γ is the column name. Our goal is to compute a faithfulness score $f(c_i, \tau_i, T_i) \in [0, 1]$ and, if $f < \theta_{align}$, output a specific hallucination type. A claim is faithfully cited iff $f \geq \theta_{align}$ and $h = \emptyset$.

We formalize six structural hallucination modes that violate citation faithfulness to structured records:

- (1) **Non-Retrieval (NR):** The cited table/record does not appear in the top-K retrieved results.
- (2) **Schema Hallucination (SH)**[2]: The cited column name or table attribute is fabricated or misspelled.
- (3) **Row Misalignment (RM)**[8]: The cited row exists but contains the wrong entity (e.g., the claim is about China, but the row is India).
- (4) **Cell Misalignment (CM)**[8]: The cited cell exists, but its content contradicts or is neutral to the claim.
- (5) **Aggregation Hallucination (AH):** The claim requires a derived value (sum, average, growth rate) but the LLM cites raw cells without correct derivation.
- (6) **Temporal/Unit Hallucination (TU):** The claim mismatches the cited cell's year, currency, or unit of measurement.

3.2 Faithfulness Score Calculation

The faithfulness score $f \in [0, 1]$ is computed as a product of three layer-specific scores:

$$f(c, \tau, T) = s_{struct} \cdot s_{nli} \cdot s_{arith} \quad (1)$$

where $s_{struct} \in \{0, 1\}$ (exact schema match), $s_{nli} \in [0, 1]$ (neural entailment confidence), and $s_{arith} \in \{0, 1\}$ (program execution result), computed from claim c , citation τ , retrieved corpus T , and resolved cell value v .

3.3 CiteAlign-Table: A Three-Layer Verifier

CiteAlign-Table verifies each (c, τ, T) triplet through three layers:

- (1) a symbolic schema-path validator for structural fidelity;
- (2) a schema-aware entailment module for semantic alignment;
- (3) and a lightweight program synthesizer for arithmetic consistency.

Layer 1: Schema-Path Validator (Symbolic). Given parsed path $\phi = (t, r, c)$, the structural score is exact:

$$s_{struct} = \mathbb{I}[t \in T] \cdot \mathbb{I}[c \in \mathcal{S}(T_t)] \cdot \mathbb{I}[r \in K(T_t)] \quad (2)$$

Because schema membership is deterministic, this layer achieves zero error for structural hallucinations. If $s_{struct} = 0$, the system exits immediately with hallucination mode $h \in \{NR, SH\}$.

Row Misalignment (RM) is detected by comparing the entity mentioned in the claim against the primary key of the resolved row. Let $\text{Ent}(c)$ be the subject entity extracted from c via NER, and $\text{Key}(T_t, r)$ the primary key value of row r . We use normalized Levenshtein similarity:

$$\text{LevSim}(a, b) = 1 - \frac{\text{Levenshtein}(a, b)}{\max(|a|, |b|)} \quad (3)$$

Then:

$$\text{RM} \iff \text{LevSim}(\text{Ent}(c), \text{Key}(T_t, r)) < \tau_{ent} \quad (4)$$

where $\tau_{ent} = 0.85$ (tuned on validation set). If $s_{struct} = 0$, exit with $h \in \{NR, SH, RM\}$.

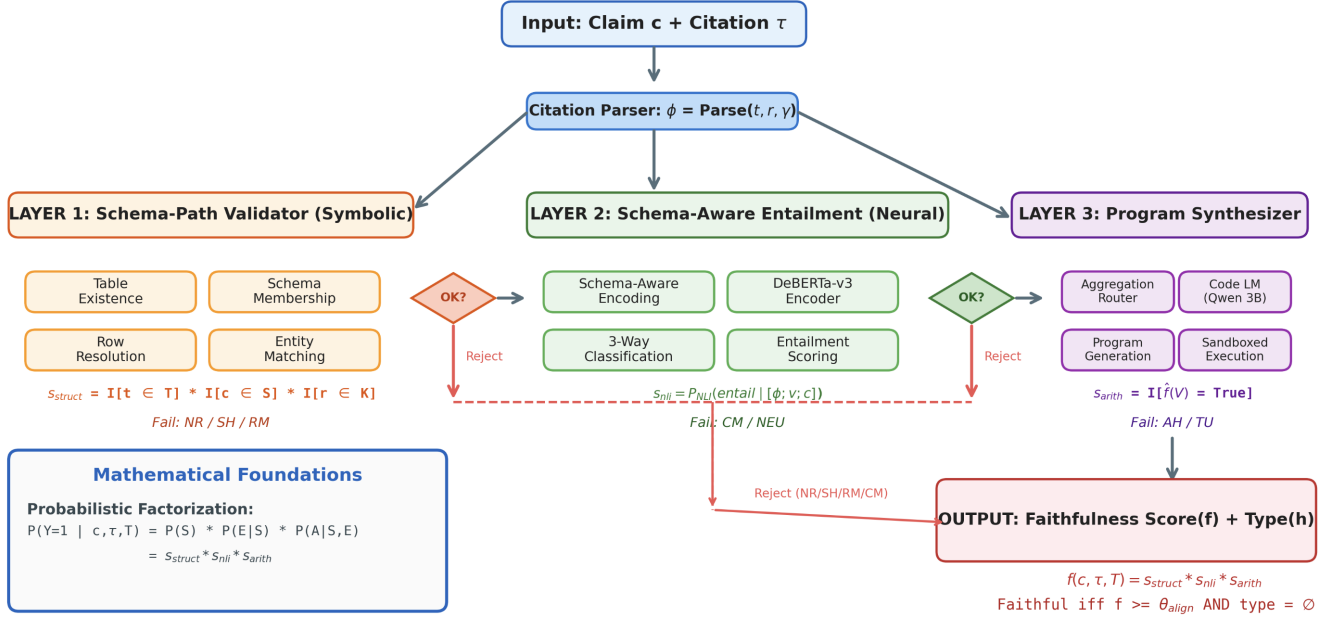


Figure 1: CiteAlign-Table Architecture

Layer 2: Schema-Aware Entailment Module (Neural). This layer detects CM for textual or categorical claims. Standard NLI (Natural Language Inference) models fail on structured citations because they process claim–passage pairs as free text, ignoring the schema role of the cited value.

We use a fine-tuned DeBERTa-v3-base[7] with a schema-injected input format:

$$\text{Premise} = [\text{Table}]t[\text{Row}]r[\text{Col}]\gamma[\text{Val}]v \quad (5)$$

which outputs $p = [p_{\text{ent}}, p_{\text{con}}, p_{\text{neu}}]$ where ent is entailment, con is contradiction, and neu is neutral score. The entailment score is $s_{\text{nli}} = p_{\text{ent}}$. We flag:

$$h = \begin{cases} \text{CM}, & \text{if } p_{\text{con}} > 0.6 \\ \text{NEU}, & \text{if } p_{\text{ent}} < 0.5 \wedge p_{\text{neu}} > 0.5 \end{cases} \quad (6)$$

NEU(Neutral Attribution) indicates the cited cell is structurally valid but semantically insufficient to verify the claim.

Layer 3: Program Synthesizer (Neuro-Symbolic). Given claim c and cited cell region V , the program synthesizer (Qwen2.5-Coder-3B[10]) generates a Python verification function $\hat{f}$ by maximizing the conditional log-probability over the space of valid programs F . The function is conditioned on the table schema $S(T_i)$ to ensure type-correct column access. Synthesis failures (compile/runtime errors) are flagged 'unverifiable' and excluded from AH/TU counts; the synthesizer succeeds on 94.2% of test cases.

$$s_{\text{arith}} = \mathbb{I}[\hat{f}(V) = \text{True}] \quad (7)$$

If false, $h = \text{AH}$ (Aggregation Hallucination) or TU (unit/temporal mismatch detected from generated code). Because the program executes the exact derivation, this layer achieves perfect recall on

AH: if the true derivation $\delta(V) \neq v$ (actual value), the program returns false with probability 1.

3.4 Decision Rule

The composite decision rule is a strict cascade:

$$\text{CiteAlign}(c, \tau, T) = \begin{cases} (0, \text{NR/SH/RM}) & s_{\text{struct}} = 0 \\ (0, \text{CM/NEU}) & s_{\text{struct}} = 1 \wedge s_{\text{nli}} < \theta_{\text{align}} \\ (0, \text{AH/TU}) & s_{\text{struct}} = 1 \wedge s_{\text{nli}} \geq \theta_{\text{align}} \\ & \wedge s_{\text{arith}} = 0 \\ (s_{\text{nli}}, \emptyset) & \text{otherwise} \end{cases} \quad (8)$$

We set $\theta_{\text{align}} = 0.5$, validated on the development set to maximize F1.

Because Layer 1 is exact, all structural hallucinations are eliminated before neural inference; downstream errors cannot resurrect NR or SH failures. The cascade design ensures early exit, keeping average validation under 110ms per citation.

4 Experiments

4.1 Dataset

SciTabAlign[8]: 372 expert-annotated scientific claim-table pairs with label $\in \{\text{Supported}, \text{Refuted}\}$ and cell-level evidence_cells (Gold cell-level annotations showing which cells support/refute the claim). We map directly to our taxonomy: Supported $\rightarrow h = \emptyset$ (faithful), Refuted $\rightarrow$ derive $h \in \{\text{NR}, \text{SH}, \text{CM}, \text{RM}, \text{AH}, \text{TU}\}$ from evidence mismatch.

Wiki Infoboxes (new): 800 claim-infobox pairs across 8 domains. Three annotators labeled each pair following an extended protocol: faithfulness (binary), hallucination type (Section 3.1), and

Table 1: Citation Quality on Hybrid Structured-Data Benchmark

Method	CP	CR	F1	Avg. Citations/Claim
No validation	0.58	0.85	0.69	2.3
String-Match	0.64	0.82	0.72	2.3
Self-Check	0.68	0.75	0.71	1.9
Text-NLI	0.71	0.79	0.75	2.1
LLM-as-Judge	0.76	0.80	0.78	2.2
CiteAlign-Table (Ours)	0.84	0.84	0.84	2

NEI (Not Enough Information) when evidence is topically related but insufficient to verify the claim. Inter-annotator agreement: Fleiss' $\kappa=0.89$. Disagreements resolved by majority vote.

We performed a stratified split of CiteAlign-Bench into train (900 pairs), validation (100 pairs), and test (172 pairs), preserving the distribution of faithfulness labels and hallucination types across all splits. DeBERTa-v3-base[7] is fine-tuned on the train split with early stopping on validation. Final results are reported on the held-out test set.

We test GPT-4o-mini[11] and Llama-3.1-8B-Instruct[6] as generators. For retrieval, Contriever[12] with Wikipedia structured corpus (top-K=10). For NLI, we use DeBERTa-v3-base fine-tuned on the training set described above. For program synthesis, Qwen2.5-Coder-3B zero-shot.

Each instance contains a claim, a structured table with schema, a verdict label, and evidence_cells indicating the minimal supporting cells.

4.2 Experimental Setup

Baselines

- **Text-NLI:** Standard DeBERTa-MNLI on claim + linearized table text (no schema awareness).
- **LLM-as-Judge:** GPT-4o prompted to evaluate whether the citation supports the claim [14].
- **String-Match:** Citations are valid if the cited cell value appears verbatim in the claim [17].
- **Self-Check:** The generator is prompted to verify its own citation [16].

Metrics: We report Citation Precision (CP), Citation Recall (CR), F1, and per-type detection F1 on the held-out dataset in Table 1.

4.3 Results

CiteAlign-Table achieves 0.84 F1 (+6 over the strongest table-aware baseline; Table 1, with gains holding across generators (+0.13 for GPT-4o[11] to +0.22 for Llama-3.1-8B[6]). GPT-4o produces fewer NR but more AH errors than Llama, underscoring that even strong LLMs struggle with structured derivations.

As shown in Table 2, our method dominates all six modes, especially SH (1.0), AH (0.83 vs. 0.34), and TU. Text-NLI scores 0 on SH for lack of schema awareness, while LLM-as-Judge fails on AH/TU because it cannot execute arithmetic or track units.

Table 2: Detection F1 by Hallucination Type

Type	% of data	Text-NLI	LLM-as-Judge	CiteAlign-Table
NR	16	0.52	0.74	0.91
SH	11	0.0	0.61	1.0
RM	13	0.48	0.72	0.88
CM	22	0.78	0.81	0.87
AH	28	0.12	0.34	0.83
TU	10	0.15	0.42	0.79

Table 3: Ablation on Layer Components

Configuration	Overall F1	AH F1	CM F1	SH F1
Full CiteAlign-Table	0.84	0.83	0.87	1.0
No Program Synthesizer	0.71	0.21	0.87	1.0
No Schema Injection	0.76	0.83	0.72	1.0
No Schema Validator	0.68	0.83	0.87	0.0
No Aggregation Router	0.79	0.45	0.87	1.0

Table 4: Validation Latency (ms per citation)

Method	Latency(ms)	Relative Speed
LLM-as-Judge (GPT-4o)	2500	1x
Self-Check (GPT-4o-mini)	800	3x
Text-NLI (DeBERTa)	45	56x
CiteAlign-Table (Ours)	110	21x

4.4 Analysis: Ablation Study & Efficiency

Table 3 shows that removing the synthesizer collapses AH detection (0.83 to 0.21), while removing schema injection degrades CM (0.87 to 0.72) and removing Layer 1 eliminates SH. Treating synthesis failures as unverifiable changes F1 by <0.01, confirming false positives are citation errors, not synthesis bugs.

Throughput at batch size 8 is 45/sec (text) and 18/sec (numerical), with template caching cutting synthesis to 28 ms (62% hit rate).

Latency on an A100 (batch size 1) is 2ms schema validation + 45ms DeBERTa + 63ms program synthesis (Table 4). Filtering claims with $f < 0.5$ improves citation precision by 15% (0.62 to 0.77) at a 0.6-point drop in answer exact match.

5 Conclusion

CiteAlign-Table is the first hybrid neuro-symbolic verifier for structural citation hallucinations. It achieves 0.84 F1 on the test set mentioned in Section 4.1, outperforming LLM-as-Judge and text-only NLI, while remaining 20× faster than LLM-as-Judge. Future work will extend to multi-hop structured citations [21] and knowledge graph triples [19].

6 Acknowledgments

Thanks to Abhishek Gupta (abhishek.gupta2@servicenow.com) and Riyaj Atar (riyaj.atar@servicenow.com) for helping label data for this architecture.

References

- [1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, Vol. 2024. 9112–9141.
- [2] Saad Belegh, Mohammed Barchane, Ahmed Zellou, and El Habib Benlahmar. 2026. Schema validation and evaluation framework for extracted schemas in JSON databases. *Scientific Reports* (2026).
- [3] W. et al. Chen. 2020. TabFact : A Large-scale Dataset for Table-based Fact Verification. In *International Conference on Learning Representations (ICLR)*.
- [4] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*. 3563–3578.
- [5] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 6465–6488.
- [6] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [7] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654* (2020).
- [8] Xanh Ho, Sunith Kumar, Yun-Ang Wu, Florian Boudin, Atsuhiko Takasu, and Akiko Aizawa. 2025. Table-Text Alignment: Explaining Claim Verification Against Tables in Scientific Papers. *arXiv preprint arXiv:2506.10486* (2025).
- [9] Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, et al. 2024. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*. 14095–14113.
- [10] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186* (2024).
- [11] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).
- [12] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118* (2021).
- [13] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys* 55, 12 (2023), 1–38.
- [14] Yookyung Lee, Joonghoon Kim, Jaehye Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 15782–15809.
- [15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
- [16] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*. 9004–9017.
- [17] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 12076–12100.
- [18] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* (2021).
- [19] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering* 36, 7 (2024), 3580–3599.
- [20] Haosheng Qian, Yixing Fan, Jiafeng Guo, Ruqing Zhang, Qi Chen, Dawei Yin, and Xueqi Cheng. 2025. Vericite: Towards reliable citations in retrieval-augmented generation via rigorous verification. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 47–54.
- [21] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 2369–2380.