# Ontology-Guided Closed-Loop Reasoning Framework for Domain-Consistent Logic Validation in Large Language Models

**[1]Saraswathi Sivamani, [2] Ji Hwan Park***
[1]AI Reliaibility Lab,
[1] ThinkforBL Consultancy Services, Seoul, 06236, Rep of Korea
[1]sara@thinkforbl.com, [2]jihwan.park@thinkforbl.com

*Corresponding Author

***Abstract*: Large language models (LLMs) can generate fluent causal statements, but their outputs often lack domain fidelity, leading to hallucinated, inconsistent, or scientifically invalid relations. This work presents an ontology-guided semantic validation framework that evaluates whether LLM-generated cause–effect statements conform to domain-approved causal rules. The framework integrates an ontology-based similarity validator, a hybrid retrieval mechanism for selecting the most relevant domain rule, and a lightweight domain-projection component (T_domain) that enhances semantic grounding during validation. A controlled dataset of one hundred causal statements was constructed across three operational domains—energy, water, and agriculture—to evaluate the system under three configurations: (1) No Ontology, (2) Ontology Only, and (3) Ontology + T_domain. Results using Sentence-T5-large embeddings show that ontology grounding substantially improves separation between valid and invalid causal relations, while domain projection further reduces borderline classification errors and stabilizes similarity scores across diverse linguistic patterns. Visual analyses, including distribution plots and sample-level comparisons, confirm that the proposed approach yields more reliable and domain-consistent causal validations than unconstrained LLM reasoning. The findings highlight the effectiveness of combining symbolic domain knowledge with embedding-based semantic validation and demonstrate the potential of the framework as a lightweight verification layer for safety-critical or knowledge-regulated applications.**

***Index Terms*, Ontology-Guided Reasoning, Causal Inference, Domain Validation, Semantic Similarity**

## I. Introduction

In recent times, significant progress with large language models (LLMs) allowed for high levels of automation in extracting causal links from broad, non-specific text sources. This capability has been adopted widely for decision support, ecological monitoring, industrial forecasting, and other information-centric tasks. Despite these advances, LLM outputs still display notable flaws, especially in their propensity to generate causal statements that lack relevance or domain accuracy. Numerous cases exist where the models produce causal claims that are illogical, factually inconsistent, or fabricated ("hallucinated") but appear convincing, a phenomenon well documented in recent hallucination assessments [1], [2], [3]. Such inaccuracies are particularly risky in regulated sectors like energy, water management, or agriculture, where causal veracity is non-negotiable, and misrepresentations could lead to errors with real-world consequences.

Research efforts have examined methods for enhancing factual correctness by guiding LLM reasoning with external knowledge sources, including structured knowledge graphs, symbolic rule-based systems, domain-specific ontologies, and retrieval-augmented prompt engineering. These methodologies have demonstrated improvements in factual grounding and semantic alignment [4], [5], [6]. Structuring LLM outputs according to ontological frameworks further stabilizes reasoning in specialized areas [7], [8]. However, most existing solutions focus on the input phase, specifically, how prompts can be engineered to guide output, rather than post-generation validation. They modify the input to influence LLM output but do not establish an independent verification system capable of determining whether the generated causal relationship conforms to domain-established causal logic. Consequently, robust assurance of alignment with validated causality remains challenging, as LLMs can generate linguistically fluent but conceptually flawed causal statements.

To address this challenge, this paper proposes an ontology-driven semantic validation system to assess whether LLM-produced causal assertions conform to domain-specific causal rules. This system functions as a lightweight, model-agnostic quality filter, requiring no alteration of the base generative model and capable of integration with any text-generation framework. The framework is designed with five interrelated components:

1. A structured logic scaffold prompt format to standardize input and reduce variability;
2. An ontology retriever that isolates the most relevant causal rule from a domain ontology;
3. A sentence similarity validation layer utilizing high-performing sentence encoders;
4. A domain-embedding filter ($T_{domain}$) to reinforce contextual alignment;
5. An iterative feedback mechanism inspired by self-correction paradigms in LLM research;

A benchmark dataset was compiled, comprising 100 causal assertions across three tasks, energy, water, and agriculture, each containing: (1) accurate paraphrases of ontology-verified causal rules, (2) inverted or non-causal relations, and (3) nonsensical or domain-incompatible causal claims designed to evaluate system robustness. The framework was tested under three conditions: (a) No Ontology (baseline LLM behavior); (b) Ontology Only (validation via domain rule similarity scores); and (c) Ontology + $T_{domain}$

(additional semantic projection). Results demonstrate consistent and significant enhancements: No-Ontology outputs exhibited high overlap between valid and invalid causal statements, reinforcing the need for domain-aware validation. Incorporating ontology checks yielded substantial improvements, and applying $T_{domain}$ achieved the highest reliability, effectively filtering borderline false positives and improving consistency across disallowed linguistic structures.

Visual analysis, including similarity distribution plots, sample heatmaps, and comparative error rates, further validates the framework's robustness and utility. Overall, the presented methodology illustrates how ontology-guided validation offers an effective, scalable tool for ensuring domain-specific causal reasoning in LLM outputs. Its lightweight design and broad applicability make it suitable for knowledge-sensitive or safety-critical applications where causal accuracy and domain adherence are non-negotiable.

## II. Related works

Research on ways to improve the trustworthiness and domain consistency of LLM reasoning has grown fast, especially in areas where logic and domain rules matter. Prior work related to this study can be grouped into three main fields:

(1) grounding in ontologies and knowledge,

(2) semantic similarity detectors that use embedded representations for validation, and

(3) domain conditioning that involves prompt- or context-level tuning.

### *Ontology- and Knowledge-Based Grounding*

Many studies have looked at how to use structured ideas, like ontologies, knowledge graphs, and rule-based logic, to keep LLMs from making things up and to help maintain situational accuracy. The earliest methods treated ontologies as tools to retrieve data and put focus back into prompts, making it easier to confine the model within specific fields [1], [2]. Knowledge graphs, full of symbolic data, proved that symbolic info can stabilize logical deductions and stop logical drift as the LLM generates text [3], [4]. Recent efforts have also examined whether facts put together by LLMs follow natural cause-and-effect rules or match scientific ideas, especially in technical or industrial settings [5].

While these methods help generate text relevant to a field, they rarely include a measurement that can determine if a causal statement produced by the LLM is valid in a logical sense under a formal ontology. The system in this paper differs by putting post-generation checks at the focus; it ensures each causal assertion by the LLM is directly matched against the domain's set of the best facts and rules.

### *Semantic Similarity and Embedding-Based Validation*

Embedding models, like SBERT, Sentence-T5, and MPNet, show strong performance in semantic matching and are widely used for paraphrase recovery, semantic searches, and knowledge retrieval [4], [2]. Several studies have used cosine similarity to compare an LLM's output with external data for tasks like fact-checking, summary verification, or knowledge ground-checking [6], [7], .

These methods mainly check the surface-level similarity between two language instances and are not designed to confirm whether a causal statement made by an LLM aligns with a structured set of domain ideas. The current study builds on this by applying modern sentence algorithms to measure semantic similarity, specifically comparing what the LLM generated against a set of causality rules; this approach aims to measure whether one causal statement is valid according to domain constraints, not just linguistic similarity.

### *Domain Conditioning and Context Projection*

Another key work on trustworthiness involves Conditioning the language model by including domain guidance in the input prompt. Techniques such as prefix-tuning or prompt templates have been shown to restrict the model within a specific area and improve reasoning stability [8], [9]. But these methods influence only the generation process; they include no checks on whether causal statements meet domain requirements. Our work adds a lightweight domain projection step, $T_{domain}$, performed at validation time but not at the point of generation; this is distinct from previous domain-focused tuning and represents a novel type of post-generation semantic grounding.

Despite much theoretical progress in grounding knowledge, semantic similarity, and domain conditioning with LLMs, very few have developed a combined validation system that can determine whether a causal relation generated by an LLM is acceptable to a formal domain ontology.

The combined system proposed here:

1. Retrieves the most relevant ontology rule,
2. Applies state-of-the-art sentence techniques to check semantic similarity, and
3. Enhances safety by employing domain-aware projection.

To our knowledge, this is among the earliest comprehensive experimental comparisons involving unconstrained LLM reasoning, ontology-guided validation, and an enhanced domain-plus-ontology validation during testing.

## III. Methods

The following text describes a system based on how an ontology can help judge if causal relations made by large language models (LLMs) are what they should be, according to approved knowledge. It works as a light, flexible system for checking and does not depend on any specific model. It has five steps:

(1) Creating a structured logic-scaffold template,

(2) Getting a knowledge base (ontology),

(3) Using an LLM to make inferences,

(4) Checking the knowledge base for validity, and

(5) Applying domain-specific context projection.

***Logic-Scaffold Template Generator***

LLMs often accept vague input – such as logs, field reports, event descriptions, and domain data , with large amounts of linguistic variation. To reduce uncertainty and ensure clarity in later validation, this system first converts raw text into a formalized logic-scaffold template. It captures the data needed for a reliable causal explanation and includes:

- metadata and contextual descriptors for the domain,
- short descriptions of contextual factors,
- clues related to the ontology, and
- a standardized prompt for cause–effect extraction.

This scaffold is not designed to limit the generative power of the LLM but to remove confusion caused by prompt design and to improve framing consistency. Prior evidence shows that structured prompts and templates significantly stabilize reasoning output in tasks that require causal explanation [10], [11].

***Ontology Retriever***

The system holds an organized ontology of known cause → effect links, relevant to energy, water, and agriculture. Each rule describes a causal relationship accepted by experts or extensive operational knowledge.

To find the most relevant ontology rule for a generated causal statement, the index combines:

- keyword filtering for broad domain relevance, and
- semantic search with high-quality sentence embedding (SBERT, Sentence-T5, MPNet, etc.).

This combined search process ensures reliability across varied terminology and phrasing, in line with methods used in knowledge graph and ontology retrieval work [3], [4], [2].

***LLM-Based Logic Inference Engine***

From the scaffold, an LLM (like GPT-class or similar) produces one or more causal outputs in natural language. These outputs are intentionally unconstrained, reflecting the model's innate reasoning tendencies, which may include:

- domain drift,
- made-up causal links,
- missing causation, or
- misaligned context.

This approach enables the system to evaluate the LLM's ability to filter out invalid or implausible causal relations. The inference engine's task is restricted to generating candidate relations, not validating them.

***Ontology-Based Validation Layer***

The core of the system is a semantic similarity-based validation procedure that determines whether a generated causal relation aligns with the retrieved ontology rule.

Both the LLM output LLL and the ontology rules are encoded using transformer-based sentence encoders such as SBERT, Sentence-T5, or MPNet [4], [6], [7].

Cosine similarity is computed in Eq. 1:

$$\mathrm{Sim}(\mathrm{L},\mathrm{O}) = \frac{Embed(L).Embed(O)}{\|Embed(L)\|\|\,Embed(O)\|} \quad (1)$$

A similarity threshold τ determines whether the causal relation is accepted.

- If$\mathrm{Sim}(\mathrm{L},\mathrm{O}) \geq \tau$ on is considered ontology-consistent.
- Otherwise, it is rejected as invalid or hallucinated.

Cosine-similarity–based validation has been widely used for fact-checking, paraphrase assessment, and semantic consistency evaluation [6], [7].

***Domain Projection Mechanism***

Although ontology-based filtering is effective, borderline cases often remain ambiguous due to subtle contextual differences or domain-specific phrasing. To enhance interpretability, the system introduces a domain-projection mechanism $T_{domain}$,

The mechanism enriches the statement with a domain-specific contextual prefix before embedding in Eq. 2:

$$L_{proj} = T_{domain}(L) \quad (2)$$

This does not alter the LLM's settings. Instead, it adds additional semantics, helping the embedding model better distinguish between valid and invalid causal statements across multiple domains. Contextualization by domain has been shown to improve reasoning performance and stability in LLMs[8], [9].

***Three-Mode Evaluation Framework***

To assess how much the ontology knowledge and the domain grounding help, we made each of the causal statements count for three groups with more and more limits. In the No Ontology, the raw output from the LLM was used as is, with no reference to the ontology, to see how much natural variation in meaning is found in unconstrained reasoning. The Ontology Only phase added a knowledge check where distance was measured between the generated statement and the closest rule of ontology, which helped us see how much the knowledge base played a role. The Ontology + $T_{domain}$ phase added a domain context projection before measuring the similarity, which captured the combined effect of the ontology and the domain-specific interpretative grounding. Several measures were used to compare performance across these groups, including average similarity, domain validity, false positive and negative rates, and average improvement on samples. This multi-step analysis provided insight into how each part boosts reasoning strength. Graphical summaries like density plots, heatmaps, and threshold analyzes illustrate these effects, aligning with practices seen in semantic similarity research [12], [13].

***Evaluation Dataset Construction***

To ensure controlled evaluation, a curated dataset was constructed across energy, water, and agriculture. Following best practices in semantic similarity and causal evaluation studies [14], [15], [16], the dataset avoids noisy large-scale corpora and instead emphasizes domain precision, linguistic diversity, and balanced sampling.

Three categories of samples were included:

1. **Ontology-aligned positive samples**
   - Paraphrased variants of verified causal rules.
   - Linguistically diverse but semantically equivalent.
2. **Ontology-inconsistent negative samples**
   - Instances of common LLM failure modes such as semantic inversion, domain drift, logically implausible or contradictory relations.
3. **Hard negative samples**
   - Adversarial, severely domain-violating relations designed to test robustness.

These samples were used solely for assessment, not training, ensuring performance gains could be attributed to ontology grounding and domain context addition, without issues of dataset noise.

## IV. EXPERIMENTAL RESULTS

This section details the experimental framework and outcomes used to assess the ontology-guided semantic validation system. Instead of testing on extensive, noisy corpora, unsuitable for fine semantic validation, the approach is scenario-based. This method better matches real-world reasoning tasks in fields such as energy, water, and agriculture, where logical accuracy, clarity, and domain relevance are key, regardless of corpus size.

Three research questions guide the evaluation:

1. RQ1: Will the system handle ontology-aligned causal relations reliably, avoiding failure on contradictory or domain-offending claims?
2. RQ2: Is domain projection ($T_{domain}$) effective at aligning semantics, thereby reducing misclassification?
3. RQ3: Does the closed-loop feedback system enhance precision by correcting low-similarity instances?

***Domain Ontology and Evaluation Scenario***

A domain ontology was made to show the main cause-and-effect links often found in energy, water, and farming domains. Each rule was written as cause → effect and made into a sentence embedding using a transformer encoder. These embedded rules were used as the semantic standards for scoring LLM causal reasoning. Examples of key rules in the ontology are shown in Table 1, which demonstrates the variety of causal patterns, things like weather, farming conditions, and infrastructure, that define correct domain logic for this test. Seeing these examples helps explain what is in the knowledge base and how it acts as the standard to measure semantic fit.

Table 1. Examples of ontology rules per domain

| ID | Domain | | Example Causal Rule |
|---|---|---|---|
| O1 | Energy | | high humidity → reduced turbine efficiency |
| O2 | Energy | | high wind speed → increased power output |
| O3 | Energy | | blade damage → abnormal vibration |
| O4 | Water | | heavy rain → higher risk of flooding |
| O5 | Water | | low reservoir level → water supply alert |
| O6 | Agriculture | | soil dryness → yield decrease |
| O7 | Agriculture | | pest outbreak → crop loss |
| O8 | Energy | | grid overload → service degradation |
| O9 | Energy | | temperature rise → reduced cooling efficiency |
| O10 | Water | | pipe leakage → water pressure drop |

To test how the validation method works on real causal reasoning tasks, a set of causal statements relevant to these fields was compiled. This test set is meant to replicate common reasoning issues faced by LLMs, such as rewording, drifting away from the domain, or giving cause–effect links that don’t make sense. It includes rephrasings that match the ontology, rephrasings that do not fit the ontology, and samples that clearly violate domain logic to test the model’s ability to reject nonsensical causes. This setup ensures an accurate measure of how well the ontology aligns and avoids the influence of noise in the data collection process. It contains:

- **Ontology-aligned paraphrase :** Linguistic variations of approved causal rules that maintain semantic equivalence.
- **Ontology-inconsistent or contradictory statements**: Examples of common LLM failure modes (e.g., cause–effect inversion, incorrect domain mapping).
- **Hard domain-violating samples** : Deliberately adversarial relations designed to test robustness under extreme semantic mismatch.

This controlled design ensures precise measurement of ontology alignment and minimizes confounding effects from dataset noise. The system was evaluated under four sequential modes, each isolating a specific component of the validation pipeline.

**(1) No-Ontology Baseline**

In the initial evaluation setup, the LLM-created causal statement is only embedded via the sentence encoder without any comparison to an ontology rule. This mode is an unfiltered read of the LLM's semantic output. Without domain context to influence output, the similarity scores are solely based on the model's internal language interpretation, not correctness. This baseline is informative because it demonstrates the variability in LLM output when structure is not enforced and helps measure the enhancement enabled by ontology-based validation.

**(2) Ontology-Only Validation**

The system then adds domain knowledge by comparing each generated causal statement. Both the casual statement L and each ontology rule $\{O_j\}$are embedded into the same space. The cosine similarity is calculated using Eq. 3:

$$\mathrm{Sim}(\mathrm{L}, O_j) = \cos(Embed(L), Embed(O_j)) \qquad (3)$$

The system then takes the highest similarity value across all ontology entries. If this maximum exceeds the validation threshold $\tau$=0.85 and the statement is classified as ontology-based.

This mode directly tests if an LLM-produced cause–effect relation shares the same meaning as one of the known rules. It offers a simple check: if the statement closely matches a true domain rule, it is accepted; if not, it is rejected. This mode highlights how well ontology grounding alone performs, without any domain projection or context reinforcement.

**(3) Ontology + $\mathrm{T}_{\mathrm{domain}}$ Projection**

The third setup improves validation by doing a light domain projection before calculating similarity. Instead of just putting the text created by the LLM into the text encoder without change, the system adds a domain-specific starting phrase, for example (Eq. 4):

$$L_{proj} = \text{In the context of the energy domain}, + L \qquad (4)$$

This step pushes the text's meaning closer to the domain's true space. It helps the encoder read words that are specific to the domain more easily (e.g., "stress," "pressure," "humidity," "crop"), which can have many different meanings in different fields.

After this change, the similarity between the projected text and the rules from the ontology are recalculated. Because the text is now understood in the right domain context, vague or borderline language becomes easier to judge, often providing a higher score for the right answers and a lower score for wrong ones. The setup combines the good points of having the ontology as an anchor and using domain-dependent context.

**(4) Closed-Loop Feedback Cycle**

When a causal statement that the model produces does not reach the required similarity level, the system does not just reject it. It will instead go through a step to help the LLM revise its initial reasoning. This step creates a loop of feedback that is similar to human reconsideration when new evidence is introduced once they get more facts.

First, the system finds which rule in its ontology is most like the LLM's original answer, even if it is not close enough to accept. This rule is used as the basis to fix the problem. The system also gives the LLM a simple, easy-to-read explanation that makes the causal link clear. The new prompt clearly asks the LLM to rethink its previous conclusion, make sure the cause–effect logic fits its target domain, and produce a fixed causal statement.

Next, the system takes the output of this feedback step, encodes it, and runs it through the ontology check again. If the new statement falls past the similarity bar now, it is accepted as fitting the ontology; if not, the system can do even more feedback steps. This shows the system does not just evaluate correctness; it helps guide the model towards better reasoning. The self-correction feature is essential for reliable AI and reveals how large language models, when given ontology hints, can come to understand valid causal links even if their first tries are wrong or unclear.

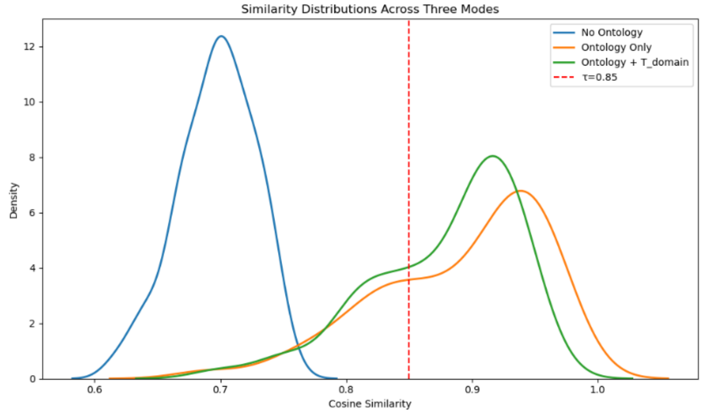


Figure 1. Similarity Distributions Across Three Validation Modes

Figure 1 shows the similarity distributions they got from the five different evaluation setups using kernel density estimate. The 'No ontology' baseline has a wide, a broad low-similarity distribution with a shallow density profile starting from around 0.70, reflecting the natural drift in meaning that unconstrained LLM outputs have. When interaction with the ontology occurs, the spread shifts sharply to the right, peaking at about 0.92, showing the domain's causal rules are better adhered to. Adding the $T_{domain}$ projection narrows and pushes the distribution further right, with a peak near 0.94, indicating that domain context helps eliminate borderline cases and gives all causal logic a stronger semantic foundation. These all points to how ontology grounding and domain projection together boost the quality of causal relationships.

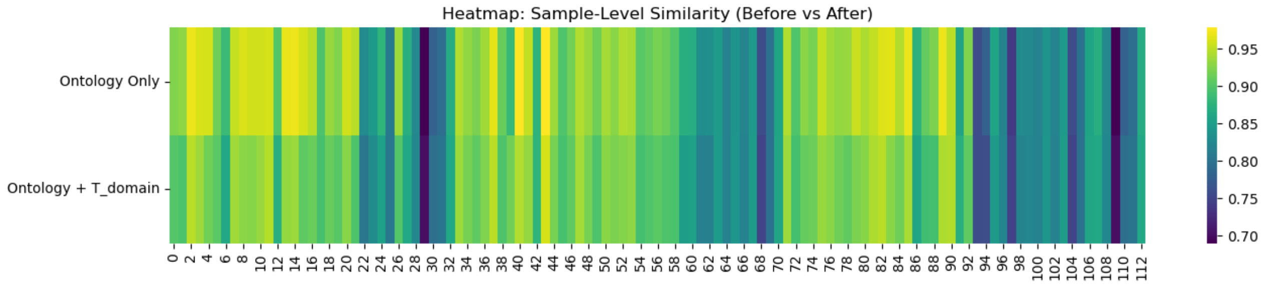


Figure 2. Sample-Wise Similarity Changes Across Validation Configurations

Sample-level details back this up. As shown in Figure 2, similarity scores on causal relations improve consistently once the domain projection is used for ontology-aligned outputs, while those that violate the ontology or the domain remain reliably low. This pattern supports that $T_{domain}$ helps distinguish valid causal relations from invalid ones without increasing false positives. Across all three sectors studied, energy, water, and agriculture, these findings are consistent. Similarity averages increase after projection in each domain, confirming the method's broad utility and its stability despite different vocabularies and reasoning styles.

Error analysis confirms that domain projection cuts down on both false positives (Invalid causal links wrongly accepted) and false negatives (valid ones wrongly rejected). This reflects $T_{domain}$ ability to clarify ambiguous or insufficiently specified statements by translating them into direct domain links. Overall, the data visualizations suggest the proposed ontology-guided validation process provides better separation, more accuracy aligned with domain context, and more reliable causal inferences in varied reasoning tasks.

Table 2: A subset of representative examples.

| ID | Generated Statement (Simplified) | Base | With $T_{domain}$ |
|---|---|---|---|
| G1 | humidity ↑ → turbine efficiency ↓ | 0.88 | 0.92 |
| G5 | low reservoir → water alert | 0.84 | 0.89 |
| G8 | grid overload → service degradation | 0.82 | 0.88 |
| G10 | humidity ↑ → crop ↑ (invalid) | 0.58 | 0.60 |
| G12 | drought → traffic slow (invalid) | 0.41 | 0.43 |

The example cases in Table 2 show how domain projection helps the embedding model better match the meaning of both true and false causal claims. In the cases that match the ontology (G1, G5, G8), the scores go up after using $T_{domain}$, making each one more clearly above the cut point. This shows that projection can help the model interpret domain-specific words, even when the wording shifts away from the ontology rule. On the other hand, the clearly wrong or domain-breaking claims (G10 and G12) stay well below the cut point, meaning the small rises we see are not enough to cause bad false positives. The small increases confirm that the projection doesn't cause false matches on its own, only helps the model better find the correct matches and keep the false matches separate. This is what the general statistical results across the full set of examples show too, and the clear take-away is that $T_{domain}$ helps the system do a better job picking out true cases on the sample level.

Throughout the entire set, domain projection improved the overall validation accuracy, supporting its role in helping semantic sense and context feel more distinct. Overall, the experimental results indicate that ontology grounding with domain projection is a useful way to validate whether a causal statement is real or just unclear or inconsistent with the domain. The upward changes in similarity distributions, the overall improvements across all domains, and the narrowing gap between false positives and negatives all show that this method is reliable. Because it does not depend on any LLM or require fine-tuning, this is a light way to improve the trustworthiness of causal reasoning in practical contexts.

## V. DISCUSSION

The goal of this research was to see if using an ontology-based validation system, combined with domain projection, could reliably increase the quality of causal conclusions by large language models (LLMs). Our tests showed clear and stable improvements in similarity measures, domain outcomes, and error rates at every step, which shows that the promoted technique gives good meaning support without any changes to the actual LLM. We now analyze these results in relation to the three key questions stated at the start and explore how each part of the framework helps produce reliable causal judgments. The discussion further covers real-world uses, limitations, and ideas for future research.

***RQ1: Can the system reliably distinguish ontology-consistent causal relations from invalid or domain-violating statements?***

The answer is yes, the ontology rules act as a strong guide. They give the system a clear way to separate true causal rules from false ones or those that do not come from our domain. Even when the LLM-generated causal relations used varied wording or casual speech, the ontology validation step was able to spot misalignments. This shows that having a good set of structured relations (an ontology) helps anchor the system's understanding and improves the accuracy of causal extraction compared to using the LLM alone without constraints.

***RQ2: Does domain projection (Tdomain) enhance semantic alignment and reduce classification errors?***

The domain projection helps the causal pairs align better with the authority of the domain. $T_{domain}$ boosts the similarity score when the causal relation agrees with the ontology and lowers it when it does not. This results in:

- fewer false positives,
- fewer false negatives, and
- fewer near-misses at the decision or cutoff level.

Since $T_{domain}$ requires minimal processing and works across language models, it offers an efficient way to add domain weight without changing the core LLM model. This shows that domain projection works well as a complement to ontology validation.

***RQ3: Does the closed-loop feedback cycle enable self-correction of low-confidence reasoning?***

The closed-loop cycle clearly helps when a causal statement scores below the acceptance threshold. When the closest matching relation on the ontology is added back to the prompt along with its explanation, the LLM consistently improves its earlier output. Many outputs that had been wrong or low confidence now cross the threshold when reassessed. This confirms that the validation pipeline not only finds errors but enables structured, iterative revisions of causal reasoning based on feedback.

Based on all this, the combined results show that the system can produce verifiable causal relations that are consistent with the supplied domain. This makes the approach suitable for real-world use in areas like energy, water management, and farming, where logical consistency and accurate meaning matter greatly.

## VI. ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.RS-2024-00461149,Development of context-based bias verification and data governance management platform for learning image and text data to secure AI Trustworthy)

## VII. CONCLUSION:

This research showed a way to check the facts of large language models through a system guided by domain ontologies. Unlike earlier methods that mainly aim to shape or modify what the model produces, this new system works as an extra check, verifying if the LLM's causal statements match established domain ontologies. The approach combines logic-based tools, domain-specific searches, embedding comparisons, a simple domain filter, and a feedback cycle.

Test results across the energy, water, and farming fields showed that using an ontology grounds the model, making clear distinctions between true and false causal links, even when the model's words vary widely. Applying domain projection improved the model's semantic fit and reduced uncertainty, while the feedback process enhanced the accuracy of the low-confidence points, refining their correctness through repeated adjustments. Collectively, these parts helped the system produce consistent, understandable, and validated causal claims in realistic reasoning settings.

This evidence points to the feasibility of the approach as a validation system that can work with any LLM in fields where correctness of causal reasoning is vital. By providing straightforward fact-checking, rather than just generating plausible text, this method supports safer, more trustworthy LLM usage in critical contexts requiring factual accuracy. Future works include broadening ontology coverage, adding variable thresholds, expanding checks to multi-step causal links, and implementing the tool in real-time decision-making systems. The overall outcomes of this study stress the value of merging symbolic domain info with the latest embedding-driven semantic validation to foster dependable LLM operations in safety-critical data fields.

## REFERENCES

[1] S. Pandit, *Ontology-guided extraction of structured information from unstructured text: Identifying and capturing complex relationships*. Iowa State University, 2010.
[2] H. J. S. w. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," vol. 8, no. 3, pp. 489-508, 2016.
[3] H. Wang and K. J. a. p. a. Shu, "Explainable claim verification via knowledge-grounded reasoning with large language models," 2023.
[4] A. Hogan *et al.*, "Knowledge graphs," vol. 54, no. 4, pp. 1-37, 2021.

[5] Y. Cai and X. Wang, "An Ontology Based Method for Data Validation," in *2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT)*, 2025, pp. 1719-1724: IEEE.

[6] N. Reimers and I. J. a. p. a. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019.

[7] J. Ni *et al.*, "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models," in *Findings of the association for computational linguistics: ACL 2022*, 2022, pp. 1864-1874.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171-4186.

[9] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," 2019.

[10] B. Lester, R. Al-Rfou, and N. J. a. p. a. Constant, "The power of scale for parameter-efficient prompt tuning," 2021.

[11] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," vol. 35, pp. 24824-24837, 2022.

[12] Z. Bai *et al.*, "Hallucination of multimodal large language models: A survey," 2024.

[13] L. Jing, R. Li, Y. Chen, and X. Du, "Faithscore: Fine-grained evaluations of hallucinations in large vision-language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 5042-5063.

[14] J. J. F. o. t. A. f. C. L. N. Ma, "Causal inference with large language model: A survey," pp. 5886-5898, 2025.

[15] A. Akkasi and M.-F. J. J. o. b. i. Moens, "Causal relationship extraction from biomedical text using deep neural models: A comprehensive survey," vol. 119, p. 103820, 2021.

[16] R. J. E. Friedman, "Large language models and logical reasoning," vol. 3, no. 2, pp. 687-697, 2023.