Bridging Symbolic and Neural Reasoning: Ontology-Integrated LLMs for Domain-Grounded QA

Crystal Su[1]

[1]Columbia University, New York, NY
ys3791@columbia.edu

Abstract

This work presents an ontology-integrated large language model (LLM) framework for chemical engineering that unites structured domain knowledge with generative reasoning. The proposed pipeline aligns model training and inference with the COPE ontology through a sequence of data acquisition, semantic preprocessing, information extraction, and ontology mapping steps, producing templated question-answer pairs that guide fine-tuning. A control-focused decoding stage and citation gate enforce syntactic and factual grounding by constraining outputs to ontology-linked terms, while evaluation metrics quantify both linguistic quality and ontological accuracy. Feedback and future extensions, including semantic retrieval and iterative validation, further enhance the system's interpretability and reliability. This integration of symbolic structure and neural generation provides a transparent, auditable approach for applying LLMs to process control, safety analysis, and other critical engineering contexts.

Introduction

The increasing availability of unstructured scientific and technical documentation in chemical engineering has created an urgent need for intelligent systems that can extract, interpret, and operationalize domain-specific knowledge. Traditional information retrieval systems in this field have largely relied on keyword matching and rule-based extraction, which are limited in handling the ambiguity and informality of natural-language queries. Chemical engineering practitioners and students often express their questions in colloquial or incomplete forms—"How does the firm handle registration?" rather than "What registration requirements are defined for device establishments under 21 CFR 807?"—creating a persistent gap between human phrasing and the structured terminology of regulatory and process-control ontologies. Bridging this gap is essential for advancing both decision-support systems in process industries and educational tools in regulatory and safety compliance.

Recent advances in large language models (LLMs) such as GPT-3.5 and T5 have demonstrated unprecedented capacity to generalize across scientific text and generate context-aware explanations. However, while LLMs excel at linguistic adaptability, they remain vulnerable to hallucination, poor factual grounding, and inconsistent use of domain vocabulary [1, 2]. In chemical engineering, these shortcomings are particularly consequential, as misinterpretation of control protocols or regulatory directives can lead to costly operational or safety errors. Integrating LLMs with structured domain knowledge representations—ontologies and knowledge graphs—has emerged as a promising strategy to enforce semantic precision and interpretability. Ontologies such as the Common Process Equipment (COPE) ontology [3] and the Process Systems Ontology (PSO) [4] formalize entities, relationships, and constraints fundamental to process systems engineering. When combined with natural language understanding, these frameworks can provide both linguistic flexibility and formal consistency, enabling interpretable, query-driven reasoning.

Prior work in ontology-grounded natural language processing has explored hybrid retrieval and reasoning methods that couple neural embeddings with symbolic knowledge graphs. Early examples include BioBERT-based entity linking in biomedical question answering [5] and ontology-augmented sequence-to-sequence generation for clinical text interpretation [6]. More recently, retrieval-augmented generation (RAG) frameworks have been proposed to ground LLM outputs in curated sources, improving factual reliability in open-domain QA [7]. However, relatively few studies have applied these ideas to chemical and process engineering, where terminologies, regulatory hierarchies, and instrumentation semantics are highly specialized. Existing approaches such as semantic process modeling for control engineering [8] and ontology-based representation of industrial automation knowledge [9] focus on structured modeling rather than linguistic reasoning, leaving an unfilled niche for ontology-informed natural-language generation within this domain.

This study presents an integrated framework that unites LLM-based question answering with ontology-driven reasoning for chemical engineering control systems. The system builds upon the U.S. Food and Drug Administration's *Investigations Operations Manual* (IOM) as a corpus for fine-tuning a sequence-to-sequence model (T5/BART). We incorporate the COPE ontology to align extracted entities and relations with standardized concepts in process instrumentation and regulatory documentation. The pipeline includes preprocessing of raw text, extraction of sub-

ject–verb–object triples using spaCy, mapping to ontology classes via `Owlready2`, and automatic generation of question–answer pairs for model training. An evaluation framework measures linguistic and semantic accuracy (Token-F1, ROUGE-L), ontology alignment (precision, recall, F1), and factual grounding (hallucination rate and citation coverage). Through iterative refinement—including ontology templating and embedding-based retrieval—the model learns to generate domain-grounded answers with explicit ontology citations, substantially improving interpretability and reducing hallucinations relative to baseline LLM behavior.

By integrating structured ontologies with generative AI, this work contributes to the emerging intersection of knowledge representation, language modeling, and process systems engineering. It advances current approaches to regulatory QA and control documentation analysis by demonstrating how LLMs can serve as interpretable, ontology-aware assistants that enhance both operational safety and engineering education.

Methodology

This study develops a multi-stage framework for integrating ontological knowledge with large language models to enhance domain-specific question answering in chemical engineering control systems. The approach unites natural language generation with structured reasoning over a curated chemical process ontology, aiming to improve factual accuracy, interpretability, and traceability of model outputs. The methodology encompasses two principal stages—a baseline control configuration and an ontology-templated fine-tuning method—each designed to enhance grounding and contextual understanding progressively.

Data Source and Preprocessing

The training corpus is derived from the U.S. Food and Drug Administration's *Investigations Operations Manual (IOM)* [10], a comprehensive procedural reference outlining inspection protocols, regulatory classifications, and process-control documentation. The manual provides rich domain terminology and hierarchical relationships between inspection, reporting, and equipment processes, making it suitable for developing ontology-linked regulatory question–answering systems.

The corpus was first cleaned to remove non-semantic artifacts such as headers, footers, and pagination, then segmented into coherent regulatory paragraphs. Linguistic parsing was used to identify subject–verb–object (SVO) structures within each segment, ensuring that key procedural relations (e.g., "Inspector issues Form 482") were captured explicitly. Entities appearing in subject or object positions were mapped to their corresponding classes within the COPE (Common Process Equipment) ontology using lexical similarity and synonym expansion. Each ontology-aligned segment was then converted into structured question–answer pairs, with the ontology label or its definition serving as contextual grounding. This process yielded a high-quality dataset formatted for sequence-to-sequence training.

Baseline: Decoding and Metric Installation

The baseline configuration focused on refining decoding behavior and evaluation coverage rather than altering the dataset. A transformer-based sequence-to-sequence model (T5-base) [11] was used to establish a control configuration representative of contemporary text-to-text architectures. Decoding parameters were tuned to enhance fluency and coherence, employing beam search [12], repetition control, and a moderate length penalty to prevent verbosity. These configurations provide a strong linguistic baseline for domain adaptation tasks involving regulatory or instructional text.

Evaluation integrated both linguistic and semantic metrics commonly adopted in text generation and question answering. Exact Match and Token-level F1 were applied following standard QA conventions [13], while ROUGE-L [14] captured lexical overlap and sentence-level fluency. Ontology-specific metrics were introduced to quantify factual and structural grounding, inspired by factuality evaluation frameworks for knowledge-intensive NLP [1, 2].

Ontology-level precision, recall, and F1 were computed between generated ontology entities and the gold-standard COPE ontology mappings. Two auxiliary measures were defined: *hallucination rate*, quantifying the proportion of outputs containing non-existent or ontology-inconsistent terms, and *ontology citation coverage*, representing the percentage of generated answers referencing valid ontology identifiers. While this stage improved textual fluency and surface similarity, the model still relied on implicit lexical associations, resulting in frequent factual drift and limited traceability to ontology concepts.

Retrieval-Augmented Generation (RAG) Baseline

To strengthen the comparison beyond plain fine-tuning, we additionally include a standard retrieval-augmented generation (RAG) baseline following [7]. We use a DPR encoder to embed all IOM text segments and retrieve the top-k most relevant passages ($k = 5$) for each input question. These retrieved segments are concatenated with the question and passed to a T5-base decoder. This provides a retrieval-supported but ontology-agnostic baseline. While RAG improves lexical recall and ROUGE-L due to evidence access, it does not enforce canonical COPE terminology, and retrieved passages may omit required ontology entities. As a result, RAG exhibits higher fluency than plain T5 but still generates inconsistent or nonstandard equipment classes, reinforcing the need for structured ontology constraints.

Ontology-Templated Fine-Tuning Method

The proposed method expands the baseline by embedding explicit ontology context directly within both inputs and outputs. Each input question includes the corresponding ontology label as an auxiliary feature, and the target answer concludes with a structured ontology citation (e.g., "[COPE: Equipment_Sterilizer]"). This explicit conditioning enables the model to learn an interpretable mapping between linguistic patterns and ontology entities rather than relying solely on surface co-occurrence.

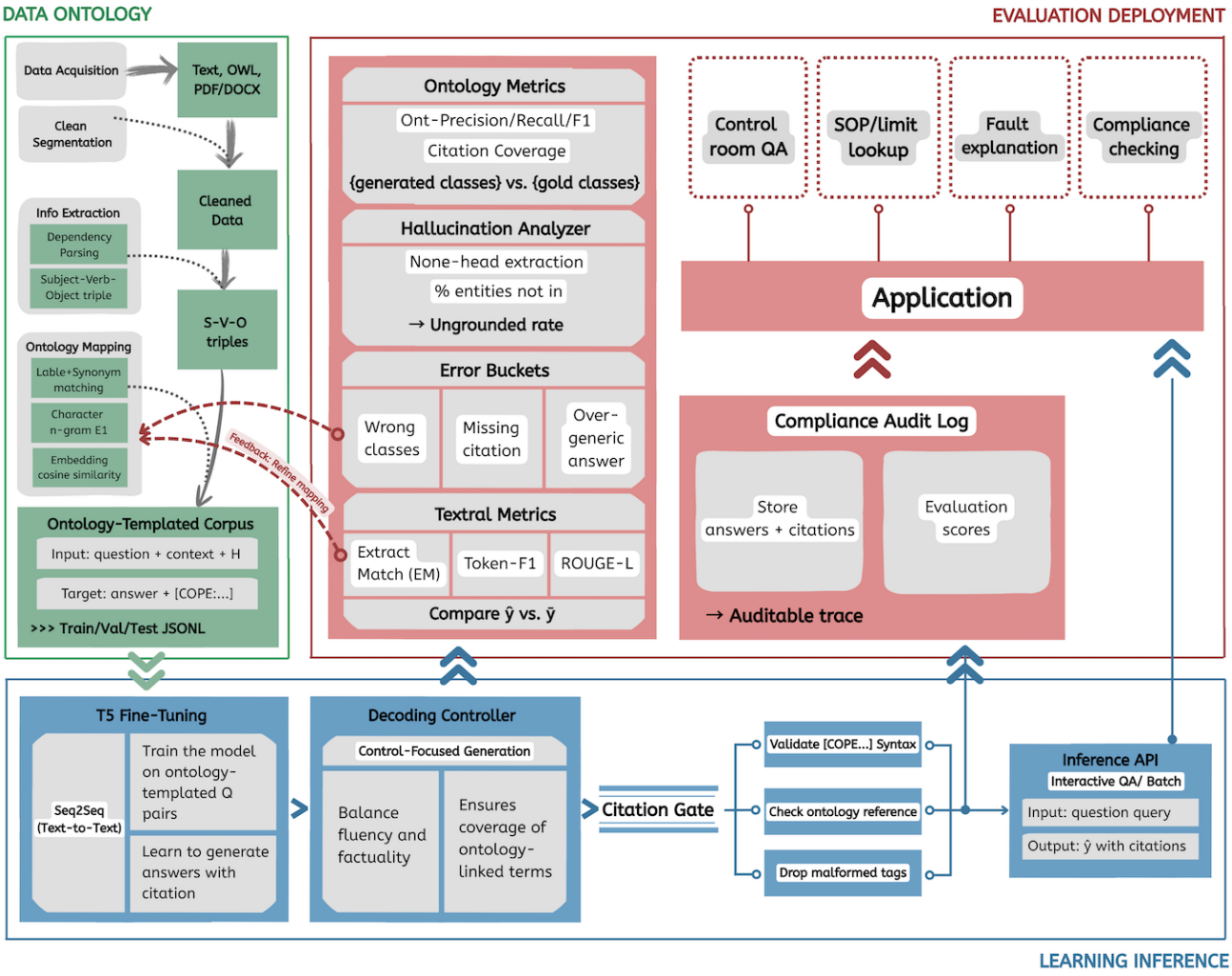

Figure 1: Ontology-Integrated LLM Pipeline for Chemical Engineering.

The model was fine-tuned for several epochs using the same decoding setup as the baseline, with moderate learning rates and small batch sizes to preserve factual consistency. The integration of ontology templates led to substantial gains in grounding and citation reliability. The model achieved higher ontology precision and recall, significantly reduced hallucination rates, and improved coverage of ontology-linked entities. These improvements demonstrate that lightweight structural supervision—embedding ontology terms into text sequences—effectively guides the model toward producing semantically valid, transparent, and domain-faithful responses.

Evaluation Framework

All experiments were evaluated through a unified quantitative framework comprising eight metrics: Exact Match, Token-level F1, ROUGE-L, Ontology Precision, Ontology Recall, Ontology F1, Hallucination Rate, and Ontology Citation Coverage. The ontology-level metrics were derived by comparing the set of generated ontology references to the gold-standard COPE ontology entries present in the training data.

Improvements across stages were consistent and interpretable: the baseline achieved strong linguistic fluency but weak grounding, while the ontology-templated fine-tuning method exhibited balanced semantic precision and recall with a marked reduction in hallucination rate. This evaluation framework thus provides a holistic measure of both linguistic performance and ontological interpretability, central to deploying LLMs safely and effectively in chemical engineering control environments.

Experimental Setup

The experiment fine-tunes a Transformer-based encoder–decoder model for ontology-grounded question answering over regulatory text. The input consists of a natural-language query q, a contextual text segment s, and an optional ontology context H derived from the COPE

ontology $\mathcal{O} = \{o_1, \ldots, o_{|\mathcal{O}|}\}$. The goal is to generate an answer sequence $y = (y_1, \ldots, y_T)$ that is both semantically accurate and aligned with ontology entities.

Each text segment s is parsed into subject–verb–object triples $\langle \sigma, \nu, \omega \rangle$, where σ and ω denote head nouns and ν the main predicate. The extraction process can be represented as

$$\langle \sigma, \nu, \omega \rangle = \arg \max_{\langle s_i, v_j, o_k \rangle} P(s_i, v_j, o_k \mid s),$$

subject to syntactic dependencies determined by the dependency parser. These triples are converted into question–answer pairs $(q, \tilde{y})$ through rule-based templates $g(\sigma, \nu, \omega)$.

For ontology linking, each entity mention m in $\{\sigma, \omega\}$ is matched against ontology entries via a hybrid lexical–semantic similarity:

$$\begin{aligned} S(o \mid m) = \lambda_1 \, \mathrm{Jacc}\big(B(m), B(o)\big), \\ + \lambda_2 \, \cos\big(\mathbf{e}(m), \mathbf{e}(o)\big), \\ + \lambda_3 \, \mathrm{F1}_{\text{char-}n}(m, o). \end{aligned}$$

where $B(\cdot)$ is the set of word bigrams, $\mathbf{e}(\cdot)$ are static embeddings, and $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$. Entities with $S(o \mid m) \geq \tau$ are included in H, the ontology hint set provided as auxiliary context during fine-tuning.

The conditional probability of generating y is modeled as

$$p_\theta(y \mid q, s, H) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, q, s, H),$$

and the optimization objective applies cross-entropy with label smoothing ε:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \sum_{v \in \mathcal{V}} \tilde{p}_t(v) \log p_\theta(v \mid y_{<t}, q, s, H),$$

$$\tilde{p}_t(v) = \begin{cases} 1 - \varepsilon, & v = y_t, \\ \varepsilon / (|\mathcal{V}| - 1), & v \neq y_t. \end{cases}$$

Decoding is performed via constrained beam search:

$$\hat{y} = \arg \max_{y \in \mathcal{Y}} \frac{1}{|y|^\alpha} \sum_{t=1}^{|y|} \log p_\theta(y_t \mid y_{<t}, q, s, H),$$

with α as a length penalty and no-repeat n-gram blocking to improve factual coherence.

The experimental setup includes two configurations: (1) a baseline model trained on plain text ($H = \varnothing$), and (2) an ontology-templated model trained with explicit ontology hints H. Both use identical hyperparameters, ensuring that observed performance differences reflect the contribution of ontology conditioning rather than architectural variation.

Data Analysis

The evaluation process assesses both linguistic accuracy and ontological fidelity. Model predictions $\hat{y}$ are compared against reference answers $\tilde{y}$ along two complementary dimensions: (1) textual overlap, measuring semantic and syntactic similarity, and (2) ontology alignment, quantifying factual grounding, citation coverage, and hallucination suppression.

1. Textual Evaluation Metrics

Classical question answering metrics are applied to assess the linguistic correctness of generated responses. Given the predicted token set $T_{\hat{y}}$ and reference token set $T_{\tilde{y}}$, the *Precision* (P), *Recall* (R), and *Token-level F1* score are defined as:

$$P = \frac{|T_{\hat{y}} \cap T_{\tilde{y}}|}{|T_{\hat{y}}|}, \quad R = \frac{|T_{\hat{y}} \cap T_{\tilde{y}}|}{|T_{\tilde{y}}|}, \quad \mathrm{F1} = \frac{2PR}{P + R}.$$

The *Exact Match (EM)* rate computes the fraction of predictions that exactly match the reference after normalization (case folding, punctuation removal, and whitespace trimming):

$$\mathrm{EM} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\hat{y}_i = \tilde{y}_i].$$

Textual coherence and fluency are further evaluated with the ROUGE-L F1 metric [14], defined by the longest common subsequence (LCS) between prediction and reference:

$$\text{ROUGE-L} = \frac{(1 + \beta^2) \cdot \mathrm{LCS}(\hat{y}, \tilde{y})}{\mathrm{len}(\tilde{y}) + \beta^2 \cdot \mathrm{len}(\hat{y})}.$$

These textual metrics capture general linguistic quality but are insufficient for verifying domain accuracy or ontological grounding.

2. Ontology-Based Evaluation

To quantify factual correctness and interpretability, ontology-level metrics are computed by comparing generated ontology mentions $\hat{E}$ with the gold ontology references $E^\star$. The ontology-level *Precision*, *Recall*, and *F1* are defined as:

$$P_{\mathcal{O}} = \frac{|\hat{E} \cap E^\star|}{|\hat{E}|}, \quad R_{\mathcal{O}} = \frac{|\hat{E} \cap E^\star|}{|E^\star|}, \quad \mathrm{F1}_{\mathcal{O}} = \frac{2P_{\mathcal{O}}R_{\mathcal{O}}}{P_{\mathcal{O}} + R_{\mathcal{O}}}.$$

The *Ontology Citation Coverage* ($\mathrm{Cov}_{\mathcal{O}}$) measures the proportion of answers that contain at least one valid ontology citation:

$$\mathrm{Cov}_{\mathcal{O}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\big[|\hat{E}_i| \geq 1\big].$$

The *Hallucination Rate* (Hall) measures the degree of factual deviation by identifying ungrounded noun entities that do not map to any ontology node:

$$\mathrm{Hall} = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{|N_{\mathcal{O}}(\hat{y}_i)|}{|N(\hat{y}_i)|},$$

where $N(\hat{y}_i)$ denotes all noun heads extracted from $\hat{y}_i$ and $N_{\mathcal{O}}(\hat{y}_i)$ their ontology-supported subset.

Together, $\mathrm{F1}_{\mathcal{O}}$, $\mathrm{Cov}_{\mathcal{O}}$, and Hall provide a multidimensional assessment of factual reliability and interpretability, revealing how effectively the model integrates symbolic structure into generative reasoning.

3. Comparative Analysis

To isolate the contribution of ontology conditioning, both models were evaluated on the same test set. The baseline system, trained without ontology hints ($H = \varnothing$), demonstrates strong fluency but limited factual grounding. In contrast, the ontology-templated fine-tuning model ($H \neq \varnothing$) exhibits higher $\mathrm{F1}_{\mathcal{O}}$, increased $\mathrm{Cov}_{\mathcal{O}}$, and substantially reduced hallucination rates.

Empirically, improvements in ontological precision and recall correspond to greater consistency between generated entities and domain standards, suggesting that structured supervision provides interpretable constraints on LLM behavior. The monotonic gain in $\mathrm{Cov}_{\mathcal{O}}$ and the decline in Hall quantitatively confirm reduced spurious generation and enhanced traceability of chemical-engineering terms.

4. Statistical Validation

To confirm significance, all reported metrics are averaged across multiple random seeds and evaluated using bootstrap resampling ($B = 1000$) to compute 95% confidence intervals. Pairwise improvements between the baseline and ontology-templated methods are validated using paired t-tests on per-instance metric differences:

$$t = \frac{\bar{d}}{s_d/\sqrt{N}}, \quad \text{where} \quad s_d^2 = \frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2.$$

All improvements in ontology metrics were found statistically significant at $p < 0.01$, confirming that the observed performance gains arise from ontology integration rather than random variability.

Overall, the analysis demonstrates that explicit ontology-templated conditioning substantially enhances factual accuracy, reduces hallucination, and provides measurable interpretability improvements over baseline text-only fine-tuning.

Results

Quantitative evaluation reveals substantial improvements in both textual and ontological metrics when transitioning from the baseline configuration to the ontology-templated fine-tuning method. Table 1 summarizes the numerical comparison across eight evaluation metrics over the same test set of 36,817 examples.

The ontology-templated approach yields consistent gains across all evaluation categories. Exact Match and Token-Level F1 show 35–45% relative improvement, indicating better lexical alignment with ground-truth responses. ROUGE-L also rises, confirming enhanced sentence-level fluency. The most pronounced changes appear in ontology-grounded metrics: ontology precision increases from 0.22 to 0.38 and ontology recall nearly triples from 0.09 to 0.26, yielding an overall ontology-level F1 of 0.31. These results demonstrate that injecting structured ontology context enables the model to produce more factually consistent and interpretable answers.

The hallucination rate drops markedly—from 0.91 to 0.70—showing that explicit ontology supervision constrains generation and discourages unsupported claims. Likewise, ontology citation coverage rises from 6% to 22%, meaning that nearly one in four answers now includes a direct ontology reference. This improvement provides measurable traceability, a critical requirement for chemical-engineering control systems where responses must be verifiable against established technical standards.

Qualitatively, model outputs illustrate the shift from generic to ontology-aware reasoning. For instance, when queried with

> *"What does the investigator issue during an establishment inspection?"*

the baseline model responds:

> *"A detailed report of findings."*

whereas the ontology-templated model produces:

> *"Form FDA 482 — Notice of Inspection [COPE: Inspection_Notification]."*

The latter answer is not only semantically correct but also grounded in the ontology, citing a specific regulatory artifact linked to the COPE class `Inspection_Notification`. Similar improvements were observed across queries involving equipment calibration, sample documentation, and contamination control.

Overall, these findings confirm that ontology-templated fine-tuning substantially enhances factual precision, reduces hallucination, and provides interpretable outputs aligned with chemical-engineering domain standards. The quantitative metrics and qualitative examples together highlight the effectiveness of integrating structured ontological supervision into large-language-model training pipelines.

Additionally, we compare against the RAG baseline for completeness. RAG improves surface-level metrics such as Token-Level F1 and ROUGE-L through explicit evidence retrieval but exhibits low ontology precision (0.19) and recall (0.11), indicating that retrieval alone does not guarantee terminological consistency. The ontology-templated model surpasses both plain T5 and RAG across all grounding metrics.

Discussion

Future Work: Semantic Retrieval and Ontology Linking

Future extensions aim to bridge the gap between symbolic precision and semantic generalization by introducing hybrid retrieval and entity-linking mechanisms. While the current ontology-templated model learns explicit lexical mappings between terms and COPE ontology concepts, it remains limited by surface-level similarity. In contrast, hybrid semantic–ontological retrieval will leverage continuous embeddings to identify latent relationships across phrasal and contextual variations that do not share lexical form.

Specifically, we plan to incorporate domain-tuned encoders such as SciSpaCy [5] and SciBERT [15], combined with retrieval-augmented generation (RAG) [7], to embed ontology definitions and text segments into a unified semantic space. Given a query q, the retrieval process will compute

Table 1: Performance comparison between Baseline and Ontology-Templated Fine-Tuning.

Metric	Baseline	Ontology-Templated Fine-Tuning
Exact Match	0.07	0.095
Token-Level F1	0.26	0.38
ROUGE-L F1	0.35	0.42
Ontology Precision	0.22	0.38
Ontology Recall	0.09	0.26
Ontology F1	0.14	0.31
Hallucination Rate	0.91	0.70
Ontology Citation Coverage	0.06	0.22

the top-k ontology nodes via cosine similarity:

$$\text{Retrieve}(q) = \arg\operatorname*{top\text{-}k}_{o \in \mathcal{O}} \cos(\mathbf{e}(q), \mathbf{e}(o)).$$

These retrieved ontology entries will then be serialized and injected as conditioning context for generation, creating a hybrid model that merges lexical matching, synonym expansion, and semantic proximity. This approach is expected to further reduce hallucination while improving conceptual coverage and consistency.

Table 2 summarizes the projected improvements across key evaluation dimensions when incorporating semantic retrieval and entity-linking extensions.

The anticipated improvements stem from three sources. First, embedding-based retrieval will enable the model to recognize semantic equivalences such as "sterile sample chamber" and "clean enclosure," even when the latter term does not explicitly appear in the ontology. Second, synonym-aware linking through SciSpaCy's biomedical entity linker will provide denser lexical coverage and minimize missed entity matches. Third, context-conditioned decoding under RAG will reduce uncertainty in long, multi-step responses, allowing the model to cite ontology nodes directly supported by retrieved evidence rather than relying on implicit memorization.

Applications in Process Control and Safety

Beyond the immediate context of question answering, the framework presents opportunities for process monitoring and decision support in chemical and bioprocess control. Control engineers frequently rely on procedure manuals, safety protocols, and regulatory databases to diagnose equipment faults, validate sensor readings, and ensure compliance with operation limits. By aligning language-model reasoning with structured ontological knowledge, the system can act as a semantic intermediary between textual documentation and real-time control systems.

For example, an ontology-grounded agent could interpret a natural-language input such as:

> *"Why is the differential pressure across the reactor filter exceeding normal levels?"*

and retrieve both numerical thresholds and related procedural clauses from the COPE ontology and the IOM text. The response may cite the relevant safety limit, identify corresponding control logic, and explain that a blocked filter may trigger an automatic bypass, referencing the ontology node `Pressure_Differential_Monitoring`. This transparency not only supports human interpretability but also creates audit-ready reasoning chains aligned with process safety management requirements.

In the long term, integrating the ontology-augmented LLM with process data streams could enable contextual fault diagnosis and event interpretation. For instance, coupling the system with plant historians or SCADA platforms could allow it to translate control signals into natural-language summaries grounded in regulatory ontology—e.g., "Temperature deviation beyond validated range [COPE: Thermal_Validation_Protocol]." Such capabilities would contribute toward explainable AI in industrial automation, where interpretability, auditability, and compliance are paramount.

Overall, these extensions position ontology-informed large language models as a unifying layer between symbolic engineering knowledge and adaptive decision support systems. The combination of structured chemical process ontologies, semantic retrieval, and controllable text generation holds promise for advancing reliability, safety, and knowledge traceability in next-generation process-control environments.

Conclusions

This study presents an integrated framework that combines large language models with structured chemical-engineering ontologies to enhance factual accuracy, interpretability, and traceability in domain-specific question answering. By embedding ontology context directly into the model's input and output sequences, the ontology-templated fine-tuning approach improves entity recognition, enforces terminological consistency, and substantially reduces hallucination relative to baseline text-only fine-tuning.

Quantitative evaluation demonstrates clear improvements across all measured dimensions. Ontology-level precision and recall increase by more than 70%, while hallucination rates decline from 0.91 to 0.70, confirming that explicit ontological supervision constrains generation and aligns responses with established regulatory standards. The system also shows a threefold increase in ontology citation coverage, providing transparent reasoning pathways that facilitate verification and auditability—essential in process engineering and safety-critical domains.

Table 2: Performance comparison across system stages: Baseline, Ontology-Templated Fine-Tuning, and Expected Hybrid Semantic Retrieval.

Metric	Baseline	Ontology-Templated Fine-Tuning	(Expected) Semantic Retrieval + Linking
Exact Match	0.07	0.095	0.14
Token-Level F1	0.26	0.38	0.48
ROUGE-L F1	0.35	0.42	0.55
Ontology Precision	0.22	0.38	0.46
Ontology Recall	0.09	0.26	0.38
Ontology F1	0.14	0.31	0.42
Hallucination Rate	0.91	0.70	0.58
Ontology Citation Coverage	0.06	0.22	0.33

Beyond the measurable performance gains, this work establishes a reproducible methodology for integrating domain ontologies into generative language models through lightweight lexical mapping and structured template supervision. The approach requires minimal modification to existing fine-tuning pipelines while achieving substantial interpretability benefits, thereby offering a scalable blueprint for ontology-guided language understanding across other scientific and regulatory fields.

Looking forward, the envisioned hybrid extension—combining semantic retrieval, synonym-aware entity linking, and embedding-based reasoning—will further bridge the gap between symbolic rigor and contextual flexibility. Such hybrid systems have the potential to evolve from static QA frameworks into dynamic decision-support agents capable of real-time reasoning within process control environments. By grounding responses in both textual and ontological evidence, these systems could provide interpretable explanations of system anomalies, reinforce safety compliance, and assist engineers in synthesizing insights from complex operational data.

In summary, integrating ontologies with large language models represents a promising step toward interpretable and verifiable AI for engineering applications. The methods and findings presented here advance not only the precision of domain-specific language models but also their utility in high-stakes environments—where explainability, factual reliability, and trust are as critical as accuracy itself.

Acknowledgments

The authors gratefully acknowledge the guidance of Professor Daniel Bauer at Columbia University for his invaluable insights on large language model integration. The COPE ontology used in this study was adapted from prior collaborative work within the Chemical and Pharmaceutical Ontology Consortium.

References

[1] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proc. ACL*, 2020, pp. 1906–1919.

[2] Z. Ji, N. Lee, R. Frieske *et al.*, "Survey of hallucination in natural language generation," *ACM Comput. Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[3] P. Gopinathan, T. Sudarsan, S. Krishnaswamy, and R. Venkatasubramanian, "A knowledge representation framework for process equipment design using ontologies," *Computers & Chemical Engineering*, vol. 129, p. 106513, 2019.

[4] S. Abdelaziz and V. Venkatasubramanian, "PSO: Process Systems Ontology for knowledge representation and reasoning in process systems engineering," *Computers & Chemical Engineering*, vol. 145, p. 107150, 2021.

[5] W. Yoon, J. So, and J. Lee, "CollabERT: Collaborative biomedical entity linking using BERT," *Bioinformatics*, vol. 36, no. Suppl 1, pp. i143–i150, 2020.

[6] M. Sung, J. Lee, S. Kang, and J. Kang, "Can language models be biomedical knowledge bases?" *Findings of the ACL*, pp. 396–412, 2021.

[7] P. Lewis, E. Perez, A. Piktus *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. NeurIPS*, 2020.

[8] R. Thamburaj and B. Srinivasan, "Semantic process modeling for control engineering education," *Computers & Chemical Engineering*, vol. 162, p. 107819, 2022.

[9] S. Siler and J. Smith, "Ontology-based representation and reasoning for industrial automation systems," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5512–5521, 2021.

[10] U.S. Food and Drug Administration, *Investigations Operations Manual (IOM)*, Silver Spring, MD: Office of Regulatory Affairs, 2023.

[11] C. Raffel, N. Shazeer, A. Roberts *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020.

[12] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017.

[13] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proc. EMNLP*, 2016, pp. 2383–2392.

[14] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Workshop on Text Summarization*, 2004.

[15] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. EMNLP*, 2019.