AN ONTOLOGY ENRICHMENT FRAMEWORK USING RETRIEVAL-AUGMENTED LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Ontology enrichment, understood as the process of extending and refining existing ontologies with new concepts, relations, and instances, has become a critical task for building robust and up-to-date knowledge bases. The exponential growth of scientific publications, datasets, and multimodal resources makes manual enrichment highly impractical, creating the need for automated or semi-automated approaches. In this work, we propose a framework that leverages multimodal large language models and retrieval-augmented generation to support ontology enrichment. Our method systematically extracts semantic knowledge units, aligns them with existing ontological structures, and generates interlinked triples, thereby enhancing both the coverage and the expressivity of the ontology. This framework addresses the knowledge acquisition bottleneck by enabling scalable integration of heterogeneous resources and fostering cross-domain semantic interoperability. To illustrate its effectiveness, we apply the framework to the domain of 4D printing, a rapidly evolving field at the intersection of materials science, manufacturing, and design. By incorporating knowledge about materials, properties, stimuli interactions, process parameters, and design strategies, the framework enriches a domain-specific ontology and supports innovation in the development of programmable and multifunctional structures.

1 INTRODUCTION

In the era of artificial intelligence (AI) and data-driven technologies, the ability to structure and interpret knowledge has become a cornerstone of intelligent systems. While vast amounts of data are continuously generated, their utility depends on transforming raw information into machine-readable semantic representations. Ontologies have emerged as a key solution to this challenge, providing a formal and explicit specification of a shared conceptualization of a domain (Gruber, 1993). They allow the definition of concepts, properties, and semantic relations, which enables reasoning, knowledge integration, and inference beyond the explicitly available information (Guarino et al., 2009). Ontologies play a central role in the development of the Semantic Web, where they serve as the backbone for annotating and linking web resources with machine-interpretable semantics (Shadbolt et al., 2006). Instead of being limited to unstructured or human-centered information, the Semantic Web envisions a knowledge-rich environment where data can be shared, reused, and reasoned upon across heterogeneous systems. This has led to a significant research focus on domain- and task-specific ontologies, which are increasingly applied in diverse fields such as biomedicine (Bodenreider, 2004), materials science (Ghedini et al., 2017), and manufacturing (Chungoora et al., 2013). Similarly, the HERMES (spatiotemporal semantics and logical knowledge description of mecHanical objEcts in the era of 4D pRinting and programmable Matter for nExt-generation of CAD systemS) domain ontology has been established to capture 4D printing knowledge at the part design level (Dimassi et al., 2021).

Despite their structured nature, traditional ontologies are limited in dynamically adapting to evolving knowledge and in processing unstructured textual data and natural language inputs. These shortcomings highlight the need for enhanced integration between ontological systems and artificial intelligence (AI), particularly through natural language processing (NLP) and machine learning (ML) techniques (Li, 2018). In such a context, large language models (LLMs), such as GPT (Radford

et al., 2019) and BERT (Kenton & Toutanova, 2019), have transformed NLP by enabling advanced capabilities in translation, question answering, and text generation (Zhao et al., 2023; Xu & Poo, 2023). Built on the Transformer architecture (Vaswani, 2017), these models leverage vast datasets to achieve high levels of contextual understanding. Scaling LLMs has led to emergent reasoning capabilities, including in-context learning (ICL) (Peng et al., 2023), chain-of-thought (CoT) (Wei et al., 2022), and retrieval-augmented generation (RAG) (Gao et al., 2023). These advances mitigate some limitations of conventional AI models by enabling real-time knowledge retrieval and contextual inference. Additionally, the recent development of multimodal LLMs (MLLMs), such as GPT-4V (Achiam et al., 2023) and Gemini (Team et al., 2023), has further expanded AI's ability to integrate textual, visual, and symbolic information (Yin et al., 2024). These models are particularly relevant for domains like 4D printing, where information must be captured across modalities to support design synthesis.

Through these advancements, LLMs remain fundamentally limited in interpretability and domain specificity. Their probabilistic nature can lead to hallucinations and unreliable outputs, particularly in highly specialized fields like 4D printing. Ontologies, by contrast, offer structured and interpretable knowledge representation but lack adaptability. The fusion of both technologies presents a promising approach to overcoming these challenges, especially in the enrichment of ontological data structures, termed as ontology learning. As manual annotation is labor-intensive and not scalable for large datasets or rapidly changing domains, semi-automatic methods, such as Phrase2Onto (Pour et al., 2023), have been developed by suggesting new concepts through phrase-based topic modeling; however, they still rely heavily on user input for validation, introducing potential subjectivity and inconsistency. Fully automated approaches using NLP and ML expedite the ontology extension process but are dependent on the quality of training data. These may introduce biases or errors if the data or models are not well-aligned with domain specifics. Advanced systems like online clustering with LLM agents (Wu et al., 2024) provide innovative ways to integrate new knowledge without extensive annotated datasets. However, they struggle with maintaining consistency and effectively integrating diverse information streams, posing challenges in ensuring the accuracy and relevance of ontology extensions.

The emergence of 4D printing – a technology combining smart materials and additive manufacturing (AM) – has opened new frontiers in fields requiring adaptive, deployable, or transformative structures (Demoly & André, 2022; Demoly & Andre, 2022). This paradigm enables objects to self-transform in response to external stimuli such as heat, light, moisture, solvent, or magnetic/electric fields (Tibbits, 2013; Ge et al., 2013). The scientific landscape of 4D printing is both rapidly evolving and inherently multidisciplinary, encompassing fields such as materials science, chemistry, mechanical engineering, process engineering, and biomimicry (Demoly et al., 2021). Since its inception in 2013, the field has experienced exponential growth, with more than 3,500 publications and an estimated annual growth rate of approximately 40%, according to the Web of Science database (Demoly & André, 2021; Demoly & and, 2021; Demoly & André, 2024). Key challenges in advancing 4D printing include improving the printability of smart materials, enhancing their mechanical and actuation performance, promoting safe and sustainable deployment, and ensuring reliability under cyclic stimuli and real-world conditions (Demoly et al., 2021). These challenges can be considered as interdependent, especially when designing and developing practical 4D-printed systems, where trade-offs between material properties, process parameters, and functional requirements must be carefully balanced (Demoly et al., 2021). To support collective and coherent progress, it becomes vital to establish a comprehensive and dynamic knowledge and data infrastructure capable of integrate both historical findings and emerging research. Such an infrastructure is crucial for consolidating the existing body of knowledge and effectively guiding future developments.

The proposed retrieval-augmented MLLMs framework aims to integrate ontology-based reasoning with the generative and retrieval capabilities of MLLMs to support knowledge discovery across diverse domains. By embedding ontological structures within LLM architectures, the framework enhances knowledge extraction, semantic reasoning, and adaptive learning from both structured and unstructured data sources, ranging from scientific literature and datasets. This active ontology enrichment approach ensures real-time alignment with emerging research and technological advancements. To demonstrate its applicability, we apply this framework to the domain of 4D printing, where it enables the integration of cross-disciplinary insights related to smart materials, processes, and programmable structures.

2 ONTOLOGY ENRICHMENT FRAMEWORK

Ontology enrichment enables the enhancement of an existing preliminary ontology by automatically adding new concepts (also considered as knowledge), relationships, and individuals (meaning information or data) to make it more comprehensive and practical for a specific domain or task. To ensure both the enrichment and population of the initial ontology, we employ an integrated framework combining information retrieval with advanced text generation capabilities (as illustrated in **Figure 1**).

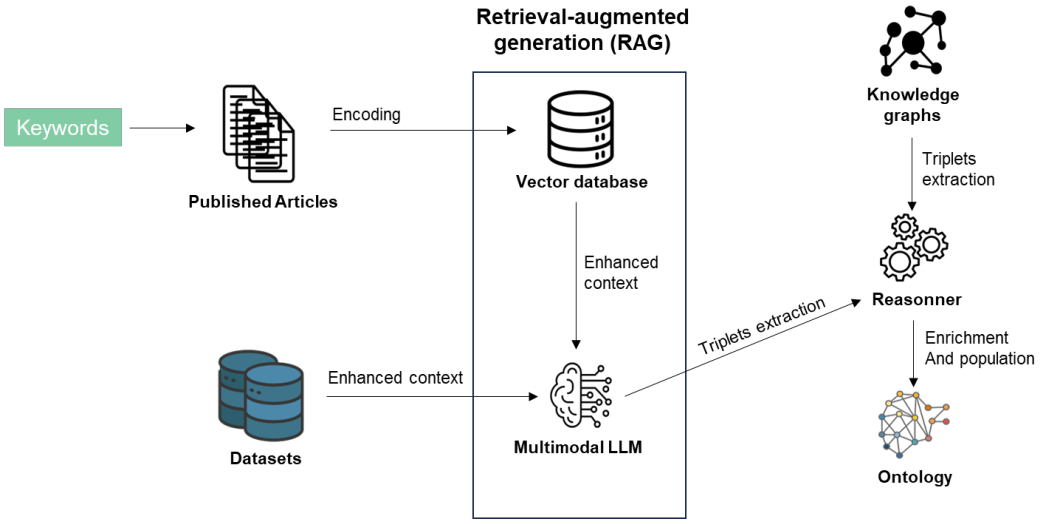

Figure 1: Retrieval-augmented ontology construction pipeline (adapted from (Bougzime et al., 2025b)).

Initially, we collected a curated corpus of published articles and domain-specific datasets using targeted keywords. Each published article is split into discrete text segments and extracted figures, while each dataset table is parsed into individual records. All text and image snippets are then encoded using a fine-tuned MLLM into dense vectors and stored in a high-throughput vector index. At inference time, the LLM issues similarity queries against this index to retrieve the top-k relevant passages or images, which it incorporates as "context windows" into its prompts. From the generated and context-aware outputs, a downstream triplet-extraction module identifies candidate [Subject-Predicate- Object] facts. These facts are merged with existing knowledge from knowledge graphs and passed to a symbolic reasoner, which enforces ontology schema constraints, checks for logical consistency, and removes duplicates. Resulted triplets are then translated into classes, properties, or instances, thereby populating and enriching the initial ontology in a continuous loop that keeps our knowledge base both up to date and semantically rigorous.

2.1 ONTOLOGY ENRICHMENT FROM SCIENTIFIC LITERATURE

To enrich the ontology, the process begins with the identification and selection of key terms relevant to the domain of interest. Using the ResearchRabbit application tool (res), an AI-supported scholarly discovery platform, the pertinent intersections among these keywords serve as the basis for collecting a large body of published research.

Then, we split these published articles using tools like LLM Sherpa (nlmatics, 2024) for robust text extraction and semantic chunking, which divided each paper into coherent chunks based on structural elements. This approach was designed to optimize both semantic completeness and computational efficiency, ensuring that each segment retained meaningful contextual information. Chunk boundaries followed the natural discourse flow (e.g., paragraphs or logical sections) rather than fixed lengths, thereby preserving local coherence throughout the segmentation process. The Aspose tool (Aspose, 2024) was used for image extraction in order to isolate each figure into standalone image files. Each token was then embedded using BERT model and CLIP for images. This process con-

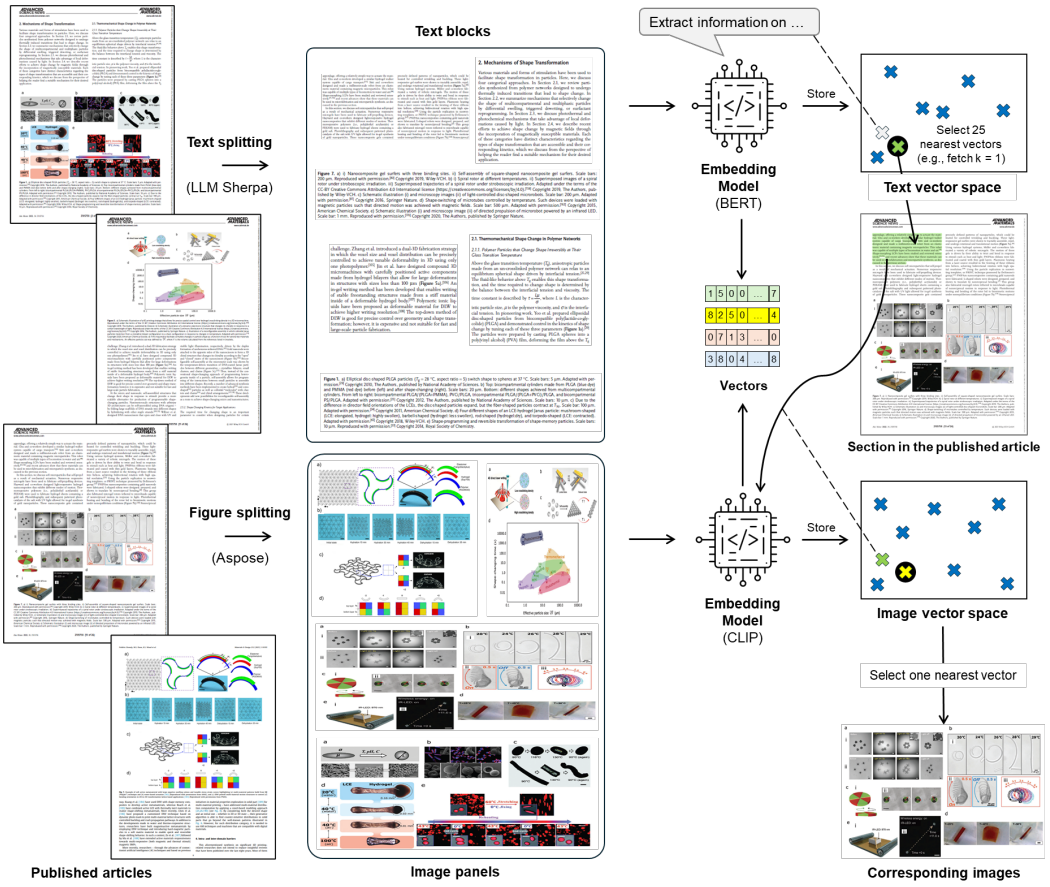

Figure 2: Pipeline for extracting textual sections and associated figures from scientific literature (adapted from (Bougzime et al., 2025b)).

verts the content into vector representations, allowing us to store them in a vector space, as shown in **Figure 2**.

Subsequently, detailed information concerning domain-specific entities, processes, and methodologies is systematically extracted from textual sources. Queries are encoded within a text vector space using BERT embeddings to identify the 25 nearest vectors. As illustrated in **Figure 2**, selecting one text vector highlights the section in green as the most relevant to the query. This section serves to identify and retrieve its corresponding relevant image through the CLIP embedding model and image vector space.

By pairing each textual section with its corresponding image, we enriched the LLaVA (Large Language-and-Vision Assistant) MLLM's input context (Parthasarathy et al., 2024; Kim et al., 2023; Jeong, 2024). This RAG process combines information retrieval with text generation, so that it helps to address challenges such as hallucination, outdated knowledge, and opaque reasoning in language models. By incorporating data from external databases, RAG ensures more accurate and credible output, particularly in knowledge-intensive tasks. This integration facilitates ongoing updates and the inclusion of specialized information, therefore making RAG a dynamic solution that combines intrinsic model knowledge with extensive external data.

Initially, we employed the few-shot learning approach (Brown, 2020; Hoffmann et al., 2022; Yang et al., 2022); however, this method was inefficient due to its time and memory demands related to the excessive length of contexts and often yielding imprecise results. To enhance efficiency, we fine-tuned the LLaVA model specifically for triplet extraction (Ghanem & Cruz, 2024; Liu et al., 2022; Zhang et al., 2024), leveraging low-rank adaptation (LoRA) (Hu et al., 2021) as a Param-

eter efficient fine-tuning (PEFT) technique (Lialin et al., 2023). We opted for LoRA because it significantly reduces the computational and memory overhead of fine-tuning. This allowed us to efficiently adapt the LLaVA model to our domain-specific task without the need for large-scale retraining. This process involved embedding domain-specific knowledge within the LLaVA model, thus configuring the output format appropriately and ensuring consistent performance without the need for additional tokens. The fine-tuning dataset comprised prompts, relevant images, and targeted responses, enabling the model to align more closely with our extraction logic and generate outputs in the specified format. Moreover, during fine-tuning, we trained the LLaVA model to distinguish ontology classes, object properties, data properties, and instances, while preserving the hierarchical relations among classes in accordance with ontology web language (OWL) formalism (Perera & Liu, 2024; Val-Calvo et al., 2025; Doumanas et al., 2025). This approach aims to refine a multimodal large language model into a tool capable of identifying ontology-relevant triplets within a specific domain.

To enable fine-tuning, a synthetic dataset is generated using a LLM. Relevant textual sections are extracted from a corpus of scientific articles, and the CLIP model is employed to retrieve the most semantically aligned image for each section. These image–text pairs are transformed into prompts, which, through a one-shot learning approach with carefully designed instructions, guide a large language model (e.g., ChatGPT-4) to generate both detailed textual descriptions and structured triplets in the form of [Subject–Predicate–Object]. The resulting dataset follows a standardized format: [prompt (combining the section and the image), triplets].

During inference, a single multimodal prompt was constructed for each target section. This prompt included: (i) the raw section text, (ii) the associated figure or schematic, and (iii) a directive stating "Extract all domain-relevant triplets". This prompt was then processed by our MLLM, which jointly attended to textual tokens and image patches to generate a set of [subject, predicate, object] assertions. For figures, the model first employed an optical character recognition (OCR) module to detect and encode text regions, and to extract key graphical elements (i.e., shapes, connectors, symbols) as visual tokens. These visual tokens interacted with text embeddings via cross-attention within the multimodal transformer. The text embeddings had been refined through our fine-tuning procedure, therefore allowing for better alignment with domain-specific semantics. This cross-modal mechanism enabled the model to infer high-level semantic relations that are not explicitly stated in the input but emerge from a combination of spatial configurations, textual cues, and prior knowledge encoded in the pretrained weights.

2.2 Ontology Enrichment From Existing Datasets

To enhance the ontology, specific datasets that align with the domain's requirements are incorporated. The selection process considers both the relevance of the datasets and their compatibility with format constraints. Integration into the ontology follows a systematic methodology involving detailed data preparation and mapping. Each dataset is decomposed into its constituent columns, which are described and cataloged, with examples provided for clarity. To categorize each attribute within the ontology, a one-shot learning approach (Li et al., 2023; Ucar et al., 2020) supported by a large language model (Jiang et al., 2023) is applied. Each cell in every row is instantiated as an individual of its corresponding ontology class, as illustrated in **Figure 3**, and resource description framework (RDF) object properties are extracted to link these instances. This end-to-end pipeline yields a richly interconnected ontology graph that faithfully captures both the structural typology and the relational semantics of the original data.

2.3 Ontology Enrichment From Knowledge Graph

Furthermore, large-scale domain knowledge graphs can be leveraged to enrich ontologies with structured knowledge. Their integration typically relies on a systematic transformation pipeline that represents information in the standard [Subject, Relation, Object] format. To ensure semantic consistency and interoperability, relationship mapping strategies are applied to align the extracted relations with the target ontology.

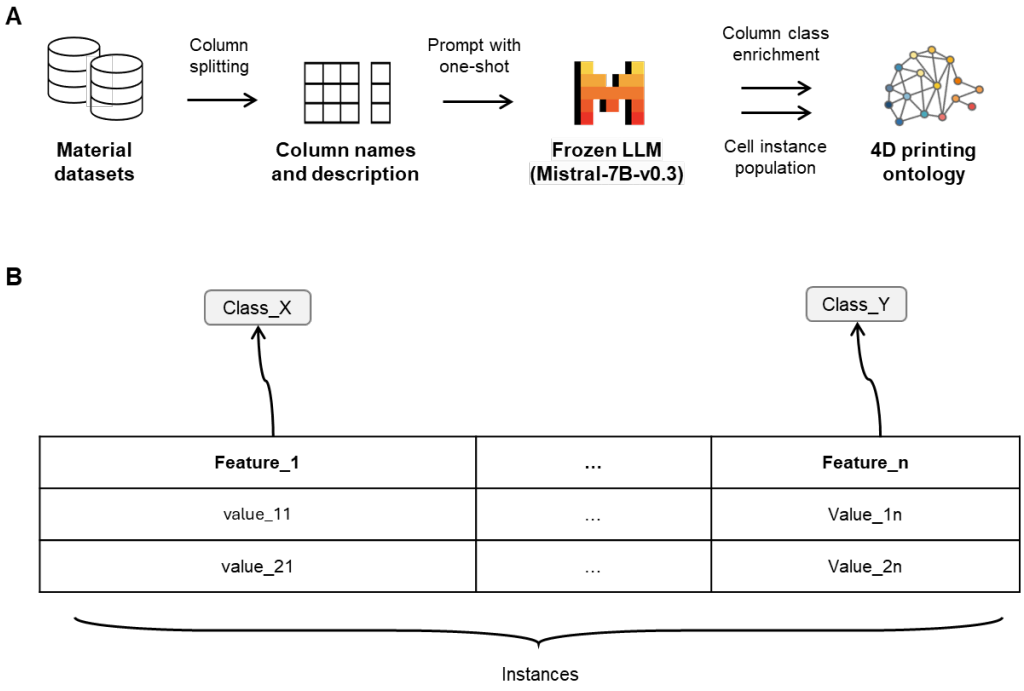

Figure 3: (A) Overview of the dataset pipeline, and (B) illustration of identified classes and intances related to a dataset representation (adapted from (Bougzime et al., 2025b)).

2.4 Preprocessing and Construction of the Ontology

The results produced by the framework are subjected to a rigorous cleaning process to ensure that only high-quality triplets are retained. In particular, the evaluation process considers the following aspects :

Domain relevance: Each triplet's subject and object are transformed into contextualized embeddings using BERT and compared against embeddings derived from a curated list of domain-specific keywords. For each triplet element, the framework computes cosine similarity scores against all domain keywords and retains the maximum similarity value as the relevance indicator. The final domain relevance assessment combines both subject and object relevance scores in the overall evaluation function. This process ensures that the data is deeply aligned with the target field.

Semantic coherence: The framework implements a comprehensive evaluation strategy to assess semantic meaningfulness. It computes direct BERT-based cosine similarity between subject and object embeddings to measure their semantic relatedness. The final coherence score integrates both the relation validity and subject-object similarity components. Additionally, predicate coherence is evaluated through template-based assessment, where the framework compares BERT embeddings of complete triplet phrases against baseline phrases and relationship templates to ensure predicate appropriateness within the semantic context.

Structural validity: The framework checks the syntactic correctness of each triplet by verifying that all elements (subject, predicate, and object) are present, of sufficient length, and follow expected formatting standards. This validation ensures data reliability for downstream applications.

Redundancy elimination: Duplicate or highly similar triplets are identified through a two-stage process. First, exact duplicates are removed through string matching of subject-predicate-object combinations. Second, semantic duplicates are detected by computing BERT-based cosine similarity between triplet embeddings, where triplets exceeding a similarity threshold are flagged as redundant. This ensures that the final dataset is concise and free from both literal and semantic redundancy.

Together, these validation steps contribute to a robust and high-fidelity cleaning process that prepares the data for subsequent ontology construction and analysis. In addition to these quality control measures, the ontology construction phase integrated several advanced techniques to further enhance the ontology. First, entity names are normalized and cleaned to create valid uniform resource identifier fragments, thereby ensuring semantic consistency across the ontology. This preprocessing step effectively mitigates errors arising from formatting discrepancies or lexical variations. Furthermore, the framework incorporates a BERT-based similarity analysis that compares new class labels with those already present in the ontology. This mechanism dynamically identifies semantically similar classes and, when a sufficient similarity threshold is met, establishes subclass relationships. In doing so, the ontology consolidates redundant entities and organizes them hierarchically in a manner that mirrors the underlying domain structure. Moreover, special attention has been given to maintaining the homogeneity of the complete ontology by enforcing uniform naming conventions and consistent semantic representations across all entities. This ensures that the entire knowledge base exhibits a high degree of internal consistency, which is critical for efficient reasoning and data integration.

3 Results: Applying the Framework to 4D Printing Ontology

The rapid advancements in 4D printing have introduced a need for a structured framework to manage and formalize the diverse knowledge involved in designing transformable systems. The HERMES ontology addresses this need by providing a semantic and logical foundation for representing the dynamic behavior of 4D-printed objects (Dimassi et al., 2021). Built upon the Basic Formal Ontology (Arp et al., 2015) and mereotopology theory (Smith, 1996), this ontology is centered on key 4D printing views, namely AM, material, transformation process, and design and engineering. Although structured around philosophical foundations and DL rules to ensure expressivity and reasoning across abstraction levels, this ontology – like most existing material ontologies – suffers from limited capabilities for automated and large-scale learning through enrichment and population. This limitation is particularly critical in emerging and rapidly evolving research domains like 4D printing, where knowledge consolidation is essential to enhance technological readiness levels and reach practical applications.

To enrich the ontology, the process starts with the selection of key terms, ie., “Additive Manufacturing”, “3D/4D Printing”, “Shape Memory Polymer”, “Shape Memory Alloy”, “Liquid Crystal Elastomer”, “Hydrogel”, “Active/Smart Material”, “Metamaterial”, and “Multi-Material Structure”. By identifying the pertinent intersections among these keywords, more than 1,810 relevant publications were retrieved. These articles are then decomposed into textual sections and extracted figures, which are encoded into dense vectors and indexed within a high-performance retrieval store. In parallel, material datasets collected from eight specialized databases (Jain et al., 2013a; Kuenneth & Ramprasad, 2022; hyd, 2023; Crews et al., 2012; of Chicago, 2023; Jain et al., 2013b; Takahashi et al., 2024; NASA) undergo a column-centric processing pipeline: column names and descriptions are parsed and mapped to ontology classes using a one-shot prompting technique with an LLM, thereby instantiating each row as an instance of its corresponding class and uncovering relationships among the fields. At inference, the MLLM retrieves the most relevant text or image snippets and generates context-aware outputs, from which a dedicated extraction module derives candidate triples. These newly extracted triples, together with pre-existing entries from the MATKG knowledge graph (Venugopal & Olivetti, 2024), are then passed to a downstream symbolic reasoner. The reasoner performs rigorous validation—ensuring coherence, semantic consistency, structural integrity, and duplicate elimination—before constructing and enriching the HERMES ontology. The quality of the extracted triplets is underpinned by a Graph BERTScore F1 (Saha et al., 2021) of 0.7, demonstrating high semantic fidelity (see Appendix A). This integrated multimodal approach thus ensures a reliable extraction of triplets from both explicit textual descriptions and implicit visual patterns.

Our framework initiates the ontology enrichment process with an initial 4D printing ontology, which comprises only 170 classes, 9 instances, 48 object properties, and 13 data properties. Through the successive integration of heterogeneous data sources and advanced validation techniques, the framework has dramatically enriched and populated the ontology. In the first phase, the system processed a corpus of scientific articles by extracting triplets that describe various domain-specific relationships. This stage resulted in the identification of 5,706 classes, 16,651 instances, 1,331 object properties, 4,390 data properties, and the establishment of 7,913 subclass relationships. The consideration of MatKG further augmented the ontology by processing additional instance-of relationships. It was

responsible for incorporating 6,629 new instances and two additional data properties with 445,370 relations. This considerable increase reflects the framework's ability to integrate detailed instance-level data from supplementary sources, thereby enhancing the granularity and applicability of the ontology. A further enrichment occurred through the automated ingestion of multiple datasets from an external directory. This step contributed 144 additional classes, 12,540,671 instances, 26 object properties, and 113 subclass relationships (see **Figure 4**). By parsing and merging these large-scale datasets, the framework ensured a comprehensive and diverse coverage of the domain knowledge, while maintaining structural validity and eliminating redundancy.

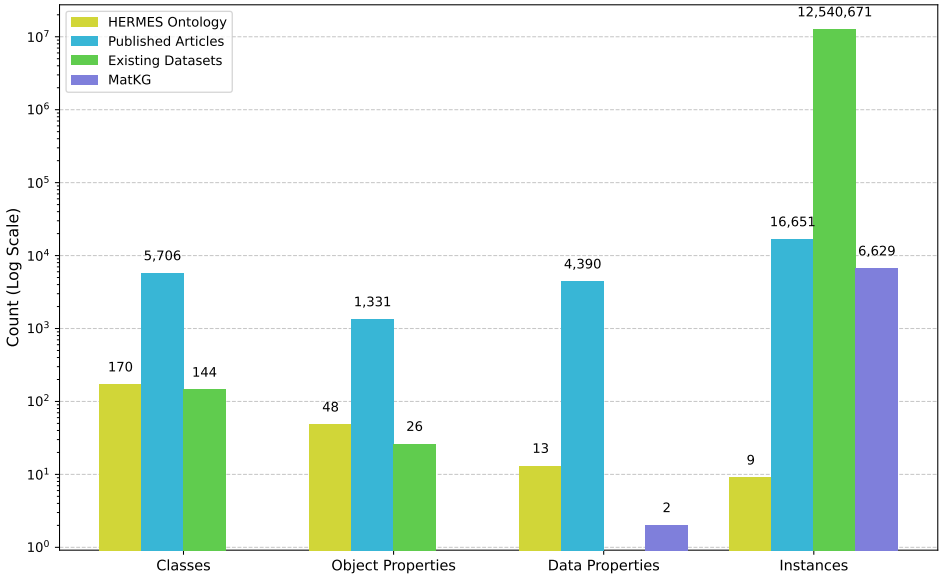

Figure 4: Comparison of ontology components between the baseline HERMES ontology and its extended counterparts derived from published articles processing, dataset parsing, and integration with MatKG triplet (adapted from (Bougzime et al., 2025b)).

After synthesizing the contributions from the scientific literature, the MatKG module, and additional datasets, the final ontology exhibits 5,849 classes, 12,563,951 instances, 1,357 object properties, 4,392 data properties, and 8,196 subclass relationships. This substantial ontology expansion demonstrates the efficacy of our multi-stage enrichment process, where material, design and engineering, and AM views have been highlighted.

In summary, the integration of multiple data sources, coupled with advanced NLP and robust validation measures, has culminated in a high-fidelity, richly structured ontology. This framework is fully domain-agnostic—while it was demonstrated on 4D printing, it can just as easily be applied to any other field. The resulting ontology not only represents a substantial expansion in scale and detail compared to its initial state but also provides a solid foundation for downstream applications such as knowledge-based reasoning, data integration, and semantic information retrieval across complex scientific and technical domains. When embedded within a neuro-symbolic AI (NSAI) framework, the ontology can be dynamically updated in real-time and reasoned over alongside neural models, thereby bridging symbolic and neural approaches for a context-aware design strategy (Bougzime et al., 2025a).

4 Conclusion

In this work, we presented an innovative framework for ontology enrichment applicable across diverse domains, integrating MLLMs and RAG to overcome the limitations of traditional ontological systems. Our approach, successfully combines the formal rigor of structured knowledge representation with the adaptive and contextual capabilities of advanced language models, which systematically captures heterogeneous information from scientific literature, databases and extensive

knowledge graphs. Experimental results demonstrate that our methodology significantly expanded an initial, rather limited ontology – starting from 170 classes and a few instances – to a comprehensive structure encompassing over 5,800 classes and more than 12.5 million instances. Future work should focus on (i) designing specialized agent architectures that integrate vision encoders and domain-specific prompt templates for materials science modalities (Bougzime et al., 2025c;d), (ii) implementing advanced verification heuristics that leverage both linguistic and visual ontological rules, (iii) developing evaluation metrics for multimodal triplet extraction that reflect the unique challenges of materials knowledge representation, and (iv) creating dedicated relation classification agents for precise typing along with specialized validation agents for ontology cohesion and triplet integrity. By embracing multimodal multi-agent systems, we can move toward adaptive ontologies that evolve seamlessly with the scientific literature, providing researchers with powerful tools for accelerated materials discovery and development.

REFERENCES

Researchrabbit. https://www.researchrabbit.ai/. Accessed: date-of-access.

Hydrogel design tools, 2023. URL https://hydrogeldesign.org/tools/.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Robert Arp, Barry Smith, and Andrew D. Spear. *Building Ontologies with Basic Formal Ontology*. The MIT Press, 2015. ISBN 9780262527811.

Aspose. Aspose.PDF for Python via .NET. https://pypi.org/project/aspose-pdf/, 2024. Accessed: date-of-access.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004.

Oualid Bougzime, Christophe Cruz, Jean-Claude André, Kun Zhou, H Jerry Qi, and Frédéric Demoly. Neuro-symbolic artificial intelligence in accelerated design for 4d printing: Status, challenges, and perspectives. *Materials & Design*, pp. 113737, 2025a.

Oualid Bougzime, Christophe Cruz, Kun Zhou, H. Jerry Qi, and Frédéric Demoly. Structuring scientific knowledge for smart materials and 4d printing: An ontology enrichment framework using retrieval-augmented large language models. *Computers in Industry*, 2025b. Under review.

Oualid Bougzime, Samir JABBAR, Christophe Cruz, and Frédéric DEMOLY. Evaluating neuro-symbolic AI architectures: Design principles, qualitative benchmark, comparative analysis and results. In *19th International Conference on Neurosymbolic Learning and Reasoning*, 2025c. URL https://openreview.net/forum?id=yCwcRijfXz.

Oualid Bougzime, Samir Jabbar, Christophe Cruz, and Frédéric Demoly. Unlocking the potential of generative ai through neuro-symbolic architectures–benefits and limitations. *arXiv preprint arXiv:2502.11269*, 2025d.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Nawaz Chungoora, Robert I. M. Young, G. Gunendran, C. Palmer, Z. Usman, and Keith Case. Towards a formal ontology for product-service systems. *Computers in Industry*, 64(4):301–309, 2013.

John H Crews, Ralph C Smith, Kyle M Pender, Jennifer C Hannen, and Gregory D Buckner. Data-driven techniques to estimate parameters in the homogenized energy model for shape memory alloys. *Journal of Intelligent Material Systems and Structures*, 23(17):1897–1920, 2012.

Frédéric Demoly and Jean-Claude André. *4D Printing, Volume 1: Between Disruptive Research and Industrial Applications*. John Wiley & Sons, 2022.

Frédéric Demoly and Jean-Claude Andre. *4D Printing, Volume 2: Between Science and Technology*. John Wiley & Sons, 2022.

Frédéric Demoly and Jean-Claude André. 4d printing: bridging the gap between fundamental research and real-world applications. *Applied Sciences*, 14(13):5669, 2024.

Frédéric Demoly, Martin L Dunn, Kristin L Wood, H Jerry Qi, and Jean-Claude André. The status, barriers, challenges, and future in design for 4d printing. *Materials & Design*, 212:110193, 2021.

Frédéric Demoly and Jean-Claude André and. Is order creation through disorder in additive manufacturing possible? *Cogent Engineering*, 8(1):1889110, 2021. doi: 10.1080/23311916.2021.1889110.

Frédéric Demoly and Jean-Claude André. Research strategy in 4d printing: Disruptive vs incremental? *Journal of Integrated Design and Process Science*, 24(2):53–73, 2021. doi: 10.3233/JID200020.

Saoussen Dimassi, Frédéric Demoly, Christophe Cruz, H Jerry Qi, Kyoung-Yun Kim, Jean-Claude André, and Samuel Gomes. An ontology-based framework to formalize and represent 4d printing knowledge in design. *Computers in Industry*, 126:103374, 2021.

Dimitrios Doumanas, Andreas Soularidis, Dimitris Spiliotopoulos, Costas Vassilakis, and Konstantinos Kotis. Fine-tuning large language models for ontology engineering: A comparative analysis of gpt-4 and mistral. *Applied Sciences*, 15(4):2146, 2025.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Qi Ge, H Jerry Qi, and Martin L Dunn. Active materials by four-dimension printing. *Applied Physics Letters*, 103(13), 2013.

Hussam Ghanem and Christophe Cruz. Fine-tuning vs. prompting: evaluating the knowledge graph construction with llms. In *3rd International Workshop on Knowledge Graph Generation from Text (Text2KG) Co-located with the Extended Semantic Web Conference (ESWC 2024)*, volume 3747, pp. 7, 2024.

Emanuele Ghedini, A Hashibon, J Friis, G Goldbeck, G Schmitz, and A De Baas. Emmo the european materials modelling ontology. In *EMMC Workshop on Interoperability in Materials Modelling*, pp. 7–8. St John's Innovation Centre Cambridge, 2017.

Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? In *Handbook on Ontologies*, pp. 1–17. Springer, 2009.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

A Jain, SP Ong, G Hautier, W Chen, WD Richards, S Dacek, S Cholia, D Gunter, D Skinner, G Ceder, et al. The materials project: a materials genome approach to accelerating materials innovation. apl mater 1: 011002, 2013a.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013b.

Cheonsu Jeong. Fine-tuning and utilization methods of domain-specific llms. *arXiv preprint arXiv:2401.02981*, 2024.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing (3rd ed. draft)*. 2025. URL https://web.stanford.edu/~jurafsky/slp3/.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, pp. 2. Minneapolis, Minnesota, 2019.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*, 2023.

Christopher Kuenneth and Rampi Ramprasad. polyone data set - 100 million hypothetical polymers including 29 properties, September 2022. URL https://doi.org/10.5281/zenodo.7124188.

Hang Li. Deep learning for natural language processing: advantages and challenges. *National Science Review*, 5(1):24–26, 2018.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Junhao Liu, Tongliang Liu, Fei Huang, et al. One shot learning as instruction data prospector for large language models. *arXiv preprint arXiv:2312.10302*, 2023.

Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*, 2023.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

Jacob Murel and Joshua Noble. What is llm temperature? https://www.ibm.com/think/topics/llm-temperature, 2024. Accessed: 2025-05-26.

NASA. Nasa shape memory repository. URL https://shapememory.grc.nasa.gov/.

nlmatics. LLM Sherpa: A Framework for Deploying and Managing Large Language Models. https://github.com/nlmatics/llmsherpa, 2024. Accessed: date-of-access.

University of Chicago. Polymer property predictor and database (pppdb), 2023. URL https://pppdb.uchicago.edu/.

Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint arXiv:2408.13296*, 2024.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Olga Perera and Jun Liu. Exploring large language models for ontology learning. 2024.

Mina Abd Nikooie Pour, Huanyu Li, Rickard Armiento, and Patrick Lambrix. Phrase2onto: A tool to support ontology extension. *Procedia Computer Science*, 225:1415–1424, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. Explagraphs: An explanation graph generation task for structured commonsense reasoning. *arXiv preprint arXiv:2104.07644*, 2021.

Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101, 2006.

Barry Smith. Mereotopology: A theory of parts and boundaries. *Data Knowledge Engineering*, 20 (3):287–303, 1996. doi: https://doi.org/10.1016/S0169-023X(96)00015-8.

Kei-ichiro Takahashi, Hiroshi Mamitsuka, Masatoshi Tosaka, Nanyi Zhu, and Shigeru Yamago. Copoldb: a copolymerization database for radical polymerization. *Polymer Chemistry*, 15(10): 965–971, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Skylar Tibbits. The emergence of "4d printing". In *TED conference*, 2013.

Talip Ucar, Adrian Gonzalez-Martin, Matthew Lee, and Adrian Daniel Szwarc. One-shot learning for language modelling. *arXiv preprint arXiv:2007.09679*, 2020.

Mikel Val-Calvo, Mikel Egaña Aranguren, Juan Mulero-Hernández, Ginés Almagro-Hernández, Prashant Deshmukh, José Antonio Bernabé-Díaz, Paola Espinoza-Arias, José Luis Sánchez-Fernández, Juergen Mueller, and Jesualdo Tomás Fernández-Breis. Ontogenix: Leveraging large language models for enhanced ontology engineering from datasets. *Information Processing & Management*, 62(3):104042, 2025.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Vineeth Venugopal and Elsa Olivetti. Matkg: An autonomously generated knowledge graph in material science. *Scientific Data*, 11(1):217, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Guanchen Wu, Chen Ling, Ilana Graetz, and Liang Zhao. Ontology extension by online clustering with large language model agents. *Frontiers in Big Data*, 7:1463543, 2024.

Bo Xu and Mu-ming Poo. Large language models and brain-inspired general intelligence. *National Science Review*, 10(10):nwad267, 2023.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 3081–3089, 2022.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, pp. nwae403, 2024.

Yujia Zhang, Tyler Sadler, Mohammad Reza Taesiri, Wenjie Xu, and Marek Reformat. Fine-tuning language models for triple extraction with data augmentation. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pp. 116–124, 2024.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

A AI MODEL ASSESSMENT

To determine the optimal temperature – a key parameter that regulates the level of randomness in the model's output during inference – we evaluated the model's performance across a range of temperature settings. Specifically, we tested two configurations: the fine-tuned model on its own, and the fine-tuned model combined with one-shot learning. Temperature plays an important role in balancing determinism and creativity in language model outputs. Lower temperatures make the model's responses more focused and predictable, while higher temperatures increase variability and originality. This trade-off impacts the accuracy and relevance of extracted triplets (Murel & Noble, 2024). As illustrated in **Figure 5**, the standard fine-tuning approach without any in-context learning demonstrated higher stability and improved performance when compared to the fine-tuning approach with one-shot across metrics which represent n-gram-based metrics encompassing precision (Bilingual Evaluation Understudy, termed as BLEU), Recall-Oriented Understudy for Gisting Evaluation (termed as ROUGE), and F1-score (combining BLEU and ROUGE metrics) (Ghanem & Cruz, 2024). These n-gram-based metrics rely on the comparison of overlapping word sequences (called n-grams) between the generated and reference texts. For instance, an *e*-gram refers to a contiguous sequence of *e* words, 1-grams are unigrams (single words), 2-grams are bigrams, and so on, thus providing nuanced evaluation of fluency and relevance in generated text (Jurafsky & Martin, 2025). Details of the metric computation are provided in the next section.

Across all four metrics, the stand-alone fine-tuned model consistently outperforms the fine-tuned with one-shot configuration, which exhibits pronounced variability and uniformly lower scores. The triplet-matching F1 peaks sharply at $T \approx 0.55$, while G-BLEU and G-ROUGE F1 scores remain optimal in the 0.55–0.70 interval. Moreover, the G-BERTScore attains its highest precision at $T \approx 0.55$, underscoring the model's fine-grained semantic alignment between predicted and reference graphs. By combining robustness – in the form of stable F1 performance – with high sensitivity afforded by both n-gram overlap and contextualized embeddings, these metrics demonstrate that $T \approx 0.55$ provides the ideal trade-off between precision and recall. Consequently, simple fine-tuning not only yields superior extractive accuracy and consistency for relational triple extraction but also avoids the added complexity and instability introduced by one-shot in-context learning.

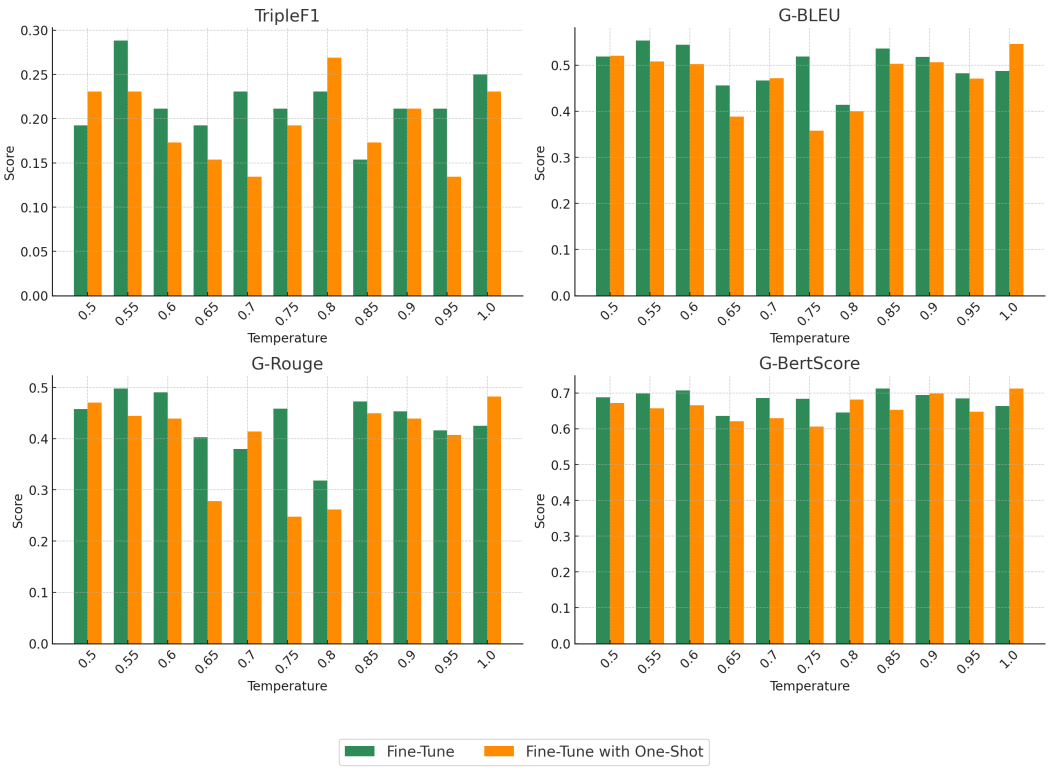

Figure 5: Comparison metrics between prompt with the fine-tuned model vs. prompt with the fine-tuned model using one-shot technique performance across various temperatures (adapted from (Bougzime et al., 2025b)).