

# Domain-Specific Ontology Construction and LLMs Fine-tuning for Procurement Knowledge

Authors: Vitalii Shevchuk, Oleksandr Kondratiuk, Daniel Hernandez De Leon and Vishwaas Narasinh

*akirolabs GmbH, Greifswalder Str. 208, 10405 Berlin, Germany*

Corresponding Author: [vitalii.shevchuk@akirolabs.com](mailto:vitalii.shevchuk@akirolabs.com)

Received: February, 2026 Published: April, 2026

## ARTICLE INFO

### Keywords:

BERT; Classification Triplet-extraction; Fine-tuning; Large Language Models Ontology; Llama; LoRA; Procurement Knowledge-extraction;

© 2026 by the Author(s). This open-access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license, making research freely available to the public and supporting a greater global exchange of knowledge and human experiments.



## ABSTRACT

This paper presents a novel semi-automated approach for creating high-quality datasets through ontology-guided knowledge extraction for domain-specific large language model fine-tuning. We address the challenge of sparse knowledge graphs (KG) generated from traditional triplet extraction methods by developing a hierarchical ontology construction framework applied to procurement domain data. Our methodology begins with procurement-specific filtering of FineWeb data using keyword-based selection, reducing the dataset size by 80%. We used Llama-3.2-3B for data annotation, achieving 3,000 positive and negative samples from 44,000 processed samples, followed by training a BERT-based classifier with an F1 score of 75%. We introduce a semi-manual ontology development approach that combines structured Resource Description Framework (RDF) with targeted large language models (LLMs) prompting for focused graph node expansion. The process involves clustering of extracted nodes to reduce complexity and enable topic-specific investigation. With procurement expert validation, we generated a dataset of 140 question-answer pairs covering key ontology nodes, while rest 460 samples were generated in automated fashion using ontology prompt. Our ontology achieves a Weighted Composite Score (WCS) of 76.42%, indicating high topic coverage across the procurement domain graph. Fine-tuning experiments on Llama-3.2-1B and Llama-3.2-3B models demonstrate improvements validated through blind A/B testing using the DeepEval framework: the fine-tuned Llama-3.2-1B model was preferred over the base model in 78.15% of comparisons for answer relevancy, 77.87% for faithfulness, and 77.95% for factual consistency rate (FCR). The fine-tuned Llama-3.2-3B model showed moderate gains, winning 68.35% for answer relevancy, 72.29% for faithfulness, and 72.36% for FCR.

## 1.0 Introduction

The rapid development of LLMs has opened new possibilities for processing specialized knowledge and supporting automated reasoning. Yet their success in domains such as procurement, healthcare, and finance is still constrained by a familiar problem: access to training data that is both accurate and contextually relevant. General-purpose datasets, while broad, often fail to capture the subtle distinctions and relationships that define these fields. As a result, even general LLMs may underperform when applied to tasks that demand precise domain understanding.

A common strategy for structuring domain knowledge relies on extracting subject-predicate-object triplets and assembling them into KGs - network-like representations of entities and their relationships. While useful, this method tends to produce sparse structures with limited semantic depth (Peng et al., 2023). The lack of connectivity makes it difficult to represent the layered relationships that exist in complex settings. Procurement offers a good example: interactions among suppliers, contracts, regulatory frameworks, and supply chain dependencies are difficult to capture with traditional extraction methods, leaving critical knowledge gaps.

One promising enhancement lies in combining semi-automated extraction with the richer scaffolding of ontologies. Ontological frameworks allow concepts and their dependencies

to be organized systematically, offering guidance that raw triplet extraction lacks. The challenge, however, is practical. Fully manual ontology construction requires significant time and expertise, while fully automated methods often fall short in precision. Bridging this gap requires a middle ground.

This study proposes a semi-automated framework that unites structured RDF-based ontology construction with targeted LLM prompting. The approach is designed to move beyond the sparsity of conventional methods by enabling systematic knowledge expansion, while still incorporating expert oversight to ensure accuracy. The result is a process that can scale without sacrificing quality, producing training and validation datasets that are both comprehensive and reliable for LLMs fine-tuning.

The research pursues three objectives. First, to design an improved methodology of domain-specific knowledge extraction that addresses the limitations of sparse triplet graphs. Second, to create a scalable, semi-automated ontology framework that blends computational efficiency with expert input. Third, to validate this approach through fine-tuning experiments on state-of-the-art models, using procurement dataset as the test case.

Procurement provides a rigorous proving ground, with its intricate web of stakeholders, contracts, and regulatory requirements. Recent surveys confirm that AI adoption in procurement spans spend analysis, supplier risk assessment, contract analytics, demand forecasting, and negotiation, yet the field remains in early adoption with most implementations at prototype stage (Guida et al., 2023). Case studies demonstrate that AI-based spend classification can meet information-processing needs that exceed human buyer capacity (Guida, Caniato, and Moretto, 2025), while experiments with LLM-based negotiation agents show that prompting strategies materially affect both pricing outcomes and supplier trust (Herold et al., 2025). Notably, LLM/GPT-based models underpin roughly 40% of supply chain AI implementations, predominantly in planning, though standardized evaluation remains limited (Bahroun et al., 2026). These characteristics - high domain complexity, growing but immature AI adoption, and a lack of standardized benchmarks - make procurement an ideal testbed for validating ontology-guided dataset construction and fine-tuning methodologies. By applying and evaluating our framework in this context, we aim to demonstrate its practical feasibility and broader potential with other domains.

## 2.0 Related Work

Parameter-efficient fine-tuning has become the dominant paradigm for adapting large language models to specialized domains. Edward J. Hu and team proposed LoRA, which reduces trainable parameters by up to 10,000× while matching or exceeding full fine-tuning performance (Hu et al., 2021). For alignment, Ouyang and team demonstrated with InstructGPT that a 1.3B RLHF-tuned model can outperform a 175B base model on instruction-following tasks (Ouyang et al., 2022). More recent preference-based methods such as DPO simplify the alignment pipeline by eliminating the need for an explicit reward model (Rafailov et al., 2023). On the knowledge integration side, Patrick Lewis and the team introduced Retrieval-Augmented Generation (RAG), which grounds LLM outputs in external document stores and achieves state-of-the-art results in open-domain QA (Lewis et al., 2020). Sanghoon Kim, together with the team, demonstrate in SOLAR 10.7B that combining continued pretraining with SFT and DPO alignment yields a highly competitive domain-adapted model, with each pipeline stage contributing distinct and complementary gains (Kim et al., 2024). For data-scarce domains Ke Wang, together with the team, survey synthetic data generation approaches across the full LLM training lifecycle, though they caution against bias and evaluation pitfalls (Wang et al., 2024).

While recent domain-adapted models such as BloombergGPT (finance), SaLLM (legal), and Med-PaLM (medical) demonstrate the value of domain fine-tuning, they typically require hundreds of thousands to millions of domain-specific samples. In contrast, our work demonstrates that ontology-guided dataset construction enables effective fine-tuning with as few as 600 curated samples, achieving 78%-win rates over base models - suggesting that

structured knowledge representation can substantially reduce data requirements for domain adaptation (Wu et al., 2023), (Colombo et al., 2024), (Singhal et al., 2023).

## 3.0 Materials and Methods

### 3.1 Experimental Design

The study objectives were to:

- Develop an improved methodology for domain-specific knowledge extraction that overcomes the sparsity limitations of traditional triplet-based KG;
- Create a semi-automated ontology construction framework for procurement domain data (Wu et al., 2025);
- Validate the effectiveness of ontology-guided dataset on LLMs fine-tuning (Doumanas et al., 2025).

The experimental design consisted of two main phases: Figure 1 and Figure 2.

### 3.2 Data Collection and Processing

Initial Data Filtering: (Penedo et al., 2024) introduce FineWeb, a 15-trillion token dataset derived from Common Crawl (free, open repository of web crawl data) snapshots. Corpus of text was filtered using procurement-related keywords and their morphological variants (including various tenses, singular/plural forms). Sample of keywords:

```
"contract": ["Contract", "contracts", "contracted", "contractor", "contractors"], "agreement": ["agreements", "agreement"], "terms and conditions": ["terms and conditions"], "supplier": ["Supplier", "suppliers", "supplied", "supplying"]
```

This keyword-based approach reduced original dataset by approximately 80%, creating a domain-focused subset.

**LLM-based Annotation:** The filtered dataset underwent annotation with AI (Karim et al., 2025) using Ollama - open source on premise LLM as a service provider and Llama-3.2-3B model that is part of the Llama 3 herd of models (Grattafiori et al., 2024) with processing times of 3 seconds per sample on Google Colab GPU and 15 seconds on Intel i9 CPU. From 44,000 samples processed from the filtered FineWeb dataset, 3,000 (approximately 7%) were identified as procurement-related positive samples with associated confidence levels and justifications with model as a judge approach. This yield rate is consistent with the nature of the source corpus: although keyword filtering retains text containing procurement-focused content from the keyword-matched pool. The annotation was performed using a model-as-a-judge approach, where the LLM serves as a labelling agent via a structured multi-output prompt. In recent research comparing LLMs and human labelling, agreement rates reached up to 90% (Baysan et al., 2025), setting a baseline and ground for this work. Specifically, the prompt is engineered to elicit three distinct outputs from a single inference pass: first - binary classification (YES/NO) if provided text related to procurement, second - confidence score of the model (0 - 100%) to access the most confident predictions, third (only during prompt enhancement) - model justification why answer was yes/no with clear reasoning. We note that this approach constitutes structured prompt engineering rather than multi-task learning in the traditional machine learning sense,

as the model is not jointly trained on multiple objectives but rather instructed to produce multiple outputs within a single prompt.

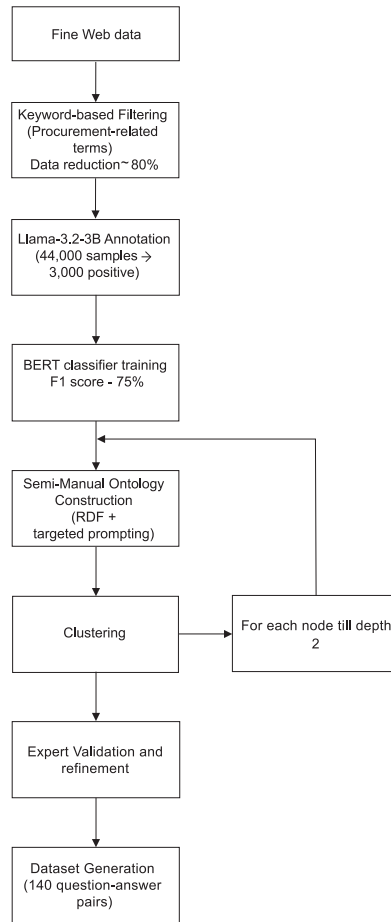


Fig. 1 Ontology construction and dataset generation scheme

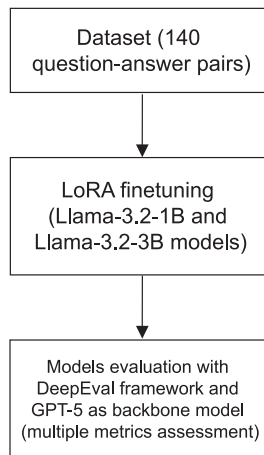


Fig. 2 Example of a figure caption

**Prompt Template used for triplets extraction:**

"You are a procurement expert. Perform a structured multi-output analysis on the text provided below. The task includes three objectives:

**Classification:** Determine whether the text is related to procurement (answer "YES" or "NO").

**Confidence Score:** Provide your confidence level (0–100%) for the classification.

**Justification (for prompt enhancement):** Explain briefly why the text was classified as procurement-related or not, with clear reasoning.

If the text is related to procurement, additionally extract up to {max\_knowledge\_triplets} knowledge triplets in the form of (subject, predicate, object), ensuring they are specific to procurement and exclude stopwords.

If the text is not related to procurement, return only the classification, confidence, and justification.

Text: {text\_placeholder}"

**Classification Model Training:** A balanced dataset was created using 3,000 positive samples and 3,000 randomly selected negative samples. Using approximately 3,000 positive and 3,000 negative samples provides a balanced and sufficiently large dataset for fine-tuning a BERT model, as this sample size typically allows the model to capture meaningful language patterns and class distinctions without overfitting, while maintaining computational efficiency during training (Devlin et al., 2019). Plus, we can't use LLMs on all data due to computational restrictions. When choosing classification model we used BERT-base-uncased model, after considering similar task done by (Safitri et al. 2025) that analyzed sentiment (that is in fact classification task) in app user reviews using the BERT-Base multilingual uncased model, achieving high accuracy across positive, neutral, and negative categories. We achieved an F1 score of 75% on positive and negative classes. Analysis of false negatives revealed that misclassified samples were predominantly from the procurement, finance, or supply chain domains. Given the nature of fine tuning, we changed embeddings space of BERT to make it better on downstream task. We note that the BERT classifier serves as a fast noise-reduction filter to scale annotation beyond the computational limits of LLM-based labeling, rather than as a core methodological contribution; accordingly, exhaustive hyperparameter tuning was not pursued.

**Triplet Extraction:** Knowledge triplets were extracted using domain-specific prompting. In low-resource settings, zero-shot relation triplet extraction has proven effective by using relational prompts to generate synthetic training examples that guide the model in identifying subject-relation-object triplets without requiring extensive annotated datasets (Halike et al., 2023).

**Prompt Template used for triplets extraction:**

"You are a procurement expert. Some text is provided below. Given the text: {text\_placeholder}, extract up to {max\_knowledge\_triplets} knowledge triplets in the form of (subject, predicate, object) related to procurement. Avoid stopwords. If the text below is not related to procurement ignore it."

In contrast to dependency parsing, triplet extraction with prompts showed less sparse data space.

**Semi-Manual Ontology Construction:** ontology construction for domain-specific information is often complex, requiring multiple systems, methodologies, and supporting technologies such as natural language processing (NLP) and LLMs. Recent research highlights that developing reliable, reusable ontologies involves structural and logical complexity, collaborative construction, and integration processes, making it a challenging,

resource-intensive task (Sattar et al., 2021). To address this, in our work we adopt a simplified yet effective approach.

**Basic Ontology Framework:** a foundational RDF based ontology was established by procurement experts with core classes. Ontologies are widely used as the backbone of the semantic web because they assign meaning, properties, and relations to data in the form of subject–predicate–object triples. This allows information to be represented in a structured, machine-readable way, typically using the RDF (Jayadianti, 2022). In our work, the RDF provides a flexible foundation for organizing domain-specific knowledge and ensures compatibility with semantic web technologies:

- Company (provides GoodsAndServices);
- Business Domain;
- GoodsAndServices;
- Procurement (canBeAppliedTo Domain);
- SupplyChain (contains Company, includes Procurement).

**Targeted Node Expansion:** Focused prompting was employed for systematic ontology expansion for the desired graph depth 2:

**System Prompt:**

"You are a procurement expert. You are provided with a text chunk and an ontology in RDF format.

{basic\_ontology\_placeholder} Your task is to use ontology and text to extract up to 3 connections related and ONLY with respect to the Procurement in the form of (subject | predicate | object). If results do not relate to Procurement do not answer."

Depth 2 was selected as the cutoff point after experimental trials and careful consideration of the following factors:

1. Coverage vs. Relevance:
  - Depth 1 nodes are too shallow and only capture high-level concepts, failing to adequately represent the structure and nuances of the topic;
  - Depth 2 provides a balanced level of detail, including meaningful sub-concepts without overwhelming the model and human expert with overly granular information.
2. Complexity Management:
  - Extending beyond Depth 2 increases the number of nodes and edges exponentially, which complicates visualization, interpretation, and downstream processing;
  - Depth 2 ensures the graph remains manageable within RDF framework while still capturing critical relationships.
3. Semantic Significance:
  - Nodes at Depth 2 represent key actionable or analyzable entities that are directly relevant to decision-making, which aligns with the goals of the graph;
  - Deeper nodes (Depth 3+) often represent very specific examples or methods, which are better treated as optional extensions rather than core nodes.
4. Computational Efficiency:
  - Limiting the depth reduces computational cost for tasks such as graph-based algorithms, LLM generation for each node;
  - Depth 2 provides sufficient context for most analytical prompts.

This design choice is further supported by the graph neural network literature, where traversals beyond 2 hops encounter two well-documented failure modes: oversmoothing, in which node representations become indistinguishable (Rusch, Bronstein and Mishra, 2023; Wu et al., 2023), and neighborhood explosion, where the extracted subgraph grows exponentially in size and computational cost (Zeng et al., 2022). Tanvir Hossain (Hossain et al., 2024) further demonstrate that even with graph sparsification to delay oversmoothing, representational quality degrades markedly at depths of 3 or more. Additionally, in our expanding-nodes dataset generation pipeline, we observed that at depth 3+ the LLM exhibits prompt forgetting: the model increasingly attends to the most recently introduced leaf nodes rather than maintaining focus on the core ontology concepts, resulting in topically drifted and less relevant extractions. Depth 2 proved the most effective balance, where the LLM consistently grounded its outputs in the central procurement concepts.

The following example illustrates the concept of graph depth levels. Note that Depth 3 nodes (e.g., "Carbon Emissions," "Community Development") are shown here solely to demonstrate what lies beyond the chosen cutoff and why such nodes were excluded from the final ontology. Only Depth 0–2 nodes were retained in our implementation:

Depth 0: Procurement

↓ is connected with:

Depth 1: Environmental and Social Governance (ESG)



**Noise reduction:** Extracted nodes underwent agglomerative hierarchical clustering to identify related topics and reduce data complexity. For each significant cluster, rather than simply pruning leaf nodes, an LLM was used to produce an abstractive summary - a single representative entity name that semantically covers the clustered words - yielding meaningful, human-readable cluster labels and improving interpretability. The filtered results were then focused on procurement-specific content and summarized using Llama-3.2-3B, which efficiently captured the most significant data signals.

We selected this two-step pipeline (clustering + LLM summarization) due to its simplicity, scalability, and the fact that it avoids the need for preparing expensive fine-tuning dataset for transformer models such as BERT. In contrast, other studies have adopted more complex approaches. For example, (Koniaris et al. 2023) applied both extractive and abstractive summarization methods for Greek legal case law, requiring fine-tuned BERT models and extensive evaluation pipelines. Similarly, (Fan et al., 2023) proposed the MFMMR-BertSum model, which integrates BERT with a Maximal Marginal Relevance component to improve extractive summarization, but at the cost of increased training complexity and higher resource demands.

**Dataset Generation:** Expert-validated question-answer pairs were generated based on keyontology nodes, resulting in 140 samples the rest 460 samples were generated in automated mode with ontology prompt. The dataset achieved a WCS of 76.42% on the created ontology, indicating strong topic coverage across the procurement domain graph.

### 3.3 Model Fine-tuning

**Model Selection:** Llama-3.2-1B and Llama-3.2-3B models were selected for comparative analysis of smaller versus medium-sized model performance;

**Training Configuration:** 600 samples were split into 500 training samples and 50 test samples and 50 validation samples. To optimize computational efficiency, we employed Low-Rank Adaptation (LoRA) for fine-tuning. LoRA freezes the pre-trained model's weights and injects small, trainable rank-decomposition matrices into each transformer layer, dramatically reducing the number of trainable parameters compared to full fine-tuning. This approach enabled us to achieve efficient task adaptation with significantly lower GPU memory requirements and training overhead, consistent with the findings of (Hu et al., 2021);

**Evaluation of models:** The final evaluation was conducted on 50 test samples. We used Qwen 3.5-35B-A3B as the primary judge from Qwen family of models (Yang et al., 2025), following recent advances in LLM-as-a-judge methodology, and validated results using the DeepEval framework. DeepEval was originally introduced as a benchmark for probing the deep semantic comprehension of large multimodal models and we used adapted evaluation methodology to textual summarization tasks to ensure more reliable alignment with human-level judgment.

### 3.4 Statistical Analysis

**Model Performance Metrics:** The following metrics were chosen for comprehensive model evaluation with DeepEval framework and Qwen 3.5-35B-A3B model as a backbone judge on a 0%-100% scale:

1. Answer Relevancy: Measures alignment between generated answers and input questions;
2. Faithfulness: Assesses accuracy and truthfulness of responses;
3. Factual Consistency Rate (FCR): is the percentage of a model's generated statements that are factually accurate and aligned with trusted sources or the provided context.

**Comparative Analysis:** Performance was assessed via blind A/B testing, in which the LLM judge compared responses from base and fine-tuned models on 50 test cases without knowing model identity. Win rates represent the proportion of cases where the fine-tuned model was preferred;

- Ontology Coverage Assessment: WCS was calculated to quantify topic coverage across and relevance the procurement domain graph;
- Validation Framework: The DeepEval framework provided automated assessment of model reliability and security across multiple dimensions. Expert validation was conducted by procurement domain specialists to ensure response quality and domain alignment.

### 3.5 Computational Resources and Reproducibility

All experiments were conducted on a single NVIDIA L4 GPU (24 GB VRAM, 21.95 GB usable) running CUDA 12.8 (Driver 550.90) on a Google Cloud Platform (GCP) VM with Linux. Both models were fine-tuned sequentially on the same GPU: Llama-3.2-1B-Instruct training completed in 13.1 minutes (785 seconds), and Llama-3.2-3B-Instruct in 30.1 minutes (1,807 seconds), for a total training time of approximately 43 minutes.

- Training VRAM peak usage was approximately 21.95 GB, fitting within the 24 GB L4 capacity with gradient offloading enabled. Inference was performed via llama.cpp using GGUF-format models, consuming approximately 17–18 GB of VRAM (model weights ~17 GB, context ~0.4 GB, compute buffer ~0.5 GB). Inference speed ranged from 23–330 tokens/s for prompt evaluation and 1000–6,000 tokens/s for generation, depending on prompt length;

- The evaluation pipeline consisted of answer generation for 50 test questions across both models (~84 minutes) and pairwise A/B evaluation (~42 minutes), totaling approximately 2.1 GPU-hours for evaluation.

The entire experiment (training + evaluation) can be reproduced for under \$2 on a single NVIDIA L4 GPU in approximately 3 hours as shown in Table 1.

Table 1. Computational cost summary

Item	GPU-hours	Est. cost (GCP L4 on-demand ~\$0.70/hr)
Training (both models)	0.72 h	~\$0.50
Evaluation pipeline	2.1 h	~\$1.47
Total	2.82 h	~\$1.97

## 4.0 Theory/Calculation

### 4.1 Knowledge Graph

Knowledge plays a fundamental role in human cognition and technological advancement. Artificial intelligence (AI) systems, in their pursuit to emulate human reasoning, rely on structured representations of knowledge to interpret and act within complex domains (Ji et al., 2022). Among various approaches to formalizing knowledge, KGs have emerged as a dominant paradigm for representing interconnected information using graph-based models. A KG encodes real-world entities as nodes and their semantic relationships as edges, enabling both human and machine understanding of how concepts interrelate (Hogan et al., 2021). This representation aligns closely with the principles of ontology theory, which seeks to model the categories, properties, and relationships that define a particular domain of discourse. Thus, KGs operate within ontological concepts in a computationally tractable manner, serving as both data structures and semantic frameworks for reasoning and inference.

The foundational role of KGs in ontology engineering lies in their ability to bridge symbolic semantics and real-world data. Ontologies provide the formal vocabulary and logical axioms that define domain meaning, while KGs instantiate these definitions through concrete data and relational structures (Kong et al., 2022). Consequently, KGs are not only instrumental in advancing AI applications - such as recommender systems, question answering, and information retrieval, or dataset creation, but also central to the development of robust, ontology-driven knowledge infrastructures. However, ongoing challenges remain in automating ontology alignment, ensuring data interoperability, and maintaining semantic coherence at scale, making the study of KGs essential for the continued progress of ontology-based AI research (Dai et al., 2020).

### 4.2 Ontology Construction

The concept of an ontology as an "explicit specification of a conceptualization" was established by Gruber (Gruber, 1993) and has since guided both manual and automated construction efforts. Practical methodologies such as the Protégé-based Ontology Development 101 guide (Noy and McGuinness, 2001) formalized iterative workflows for defining classes, properties, and constraints. While our work does not use Protégé directly, our expert-defined seed ontology and iterative expansion process follow the same principled sequence of scope definition, class enumeration, and relationship specification.

Ontology construction methodologies can be broadly categorized into four principal approaches, each with distinct trade-offs in terms of structure, flexibility, and scalability:

**Top-Down:** Ontology is constructed hierarchically from broad domain concepts toward more specific sub-concepts. This approach yields a clear, domain-centric structure but tends to be rigid when applied to evolving or heterogeneous data sources. Manual ontology construction, widely recognized as time-consuming, error-prone, and labour-intensive (Sattar et al., 2021), falls predominantly within the Top-Down spectrum.

**Bottom-Up:** Ontology is derived from granular data instances, with higher-level abstractions emerging from observed patterns. This approach offers flexibility when working with large datasets but may produce ontologies that lack coherent hierarchical organization.

**Middle-Out:** Construction begins at intermediate conceptual levels, expanding both upward toward broader categories and downward toward specific instances. This approach provides balanced scalability but requires iterative refinement cycles to achieve structural consistency.

**Semi-Automatic (LLM/NLP hybrid):** Ontology construction is supported by automated extraction and generation steps, typically leveraging NLP pipelines or LLMs to identify candidate relationships. This approach offers efficiency and consistency at scale but requires human curation to ensure domain accuracy and semantic coherence.

Recent advancements in ontology construction have sought to address the limitations inherent in traditional KG construction, particularly sparsity issues arising from dependency parsing constraints. While the Resource Description Framework (RDF) provides a foundational model for representing knowledge as triples (subject-predicate-object), this structure often yields sparse, fragmented representations when directly applied to complex, unstructured data sources. This sparsity hinders AI systems' ability to effectively reason and infer from the knowledge they represent.

Semi-automatic approaches have emerged along two evaluation paradigms. On one hand, triplet-based extraction methods (Wesslund et al., 2026) evaluate results agnostic to ground-truth, measuring the impact of generated ontologies through metrics such as faithfulness, factual consistency rate, and hallucination detection using model-as-a-judge frameworks - an evaluation strategy also adopted in this work. On the other hand, approaches such as DRAGON-AI (Toro et al., 2024) attempt to evaluate ontology quality with respect to established datasets and ground truth, reporting high precision for relationship generation, though slightly lower than logic-based reasoning, and producing definitions deemed acceptable by expert evaluators yet scoring below human-authored definitions. Both paradigms highlight the inherent tension between scalability and precision in automated ontology construction.

Furthermore, the development of domain-specific ontologies has been emphasized to capture the nuances and specificities of particular fields. By constructing ontologies that reflect the unique concepts and relationships within these domains, the resulting knowledge graphs can provide more accurate and contextually relevant information, thereby improving the effectiveness of AI applications in these areas (Hao et al., 2021). Recent studies have also explored the integration of LLMs with RDF-based knowledge representations to automate and enhance KG construction, leveraging contextual understanding to generate and refine semantic relationships (Mavridis et al., 2025).

Our methodology draws elements from multiple construction paradigms to address the limitations of any single strategy:

**Top-Down initialization:** A foundational RDF-based ontology is established by procurement experts, defining core classes and relationships (Company, Business Domain, GoodsAndServices, Procurement, SupplyChain), providing the structured hierarchical scaffold characteristic of Top-Down methods.

**Data-driven expansion:** Subsequent ontology nodes are derived through LLM-driven extraction from processed domain text, guided by the existing ontology structure. This

introduces Bottom-Up characteristics by allowing the data to surface domain-relevant relationships that were not predefined by experts.

**Depth-wise iterative refinement:** The ontology is expanded systematically across depth levels, with each new layer of nodes and edges undergoing focused prompting and noise reduction through hierarchical clustering. This layered approach shares characteristics with Middle-Out methods, where intermediate concepts drive further expansion.

**Semi-Automatic validation:** Throughout the process, LLM-as-a-judge evaluation is applied to assess generated relationships, combined with expert validation to ensure domain accuracy and semantic quality.

This hybrid strategy leverages the structural clarity of Top-Down design, the data-driven flexibility of Bottom-Up extraction, the iterative refinement of depth-wise development, and the scalability of Semi-Automatic processing. The result is an ontology construction framework that addresses sparsity limitations of traditional triplet extraction while maintaining quality standards through combined automated and expert oversight.

Unlike prior semi-automatic approaches that either evaluate ontology quality against ground-truth datasets (e.g., DRAGON-AI) or rely on end-to-end LLM generation without structural constraints, our method combines an expert-defined RDF seed ontology with depth-controlled LLM expansion and hierarchical clustering for noise reduction. This three-stage pipeline seed → guided expansion → clustering ensures both structural coherence and scalability, which no single existing method simultaneously achieves.

### 4.3 Weighted Composite Score Calculation for coverage

The WCS is a metric designed to assess the completeness of an ontology by evaluating how well its nodes are represented in each dataset. This score is calculated using the formula:

$$WCS = \frac{\sum(w_i \times c_i)}{\sum(w_i)}$$

Where:

$w_i$  is the domain-importance weight for node scored on a continuous [0.0, 1.0] scale by a procurement expert or LLM acting as a procurement domain judge. The model or/and procurement expert is given the full list of ontology nodes and prompted to rate each node's importance to the procurement domain, considering: (1) frequency of appearance in procurement workflows, (2) regulatory significance, and (3) supply chain impact. This scoring is performed once over the full node set;

$c_i$  is the coverage quality score for node  $i$ , computed via pairwise LLM-as-judge evaluation. For each ontology node the LLM rates how well the sample covers the node's topic on a [0.0, 1.0] scale. The per-node coverage score is then aggregated as the arithmetic mean across all dataset samples.

$$c_i = \frac{1}{M} \sum_{j=1}^M s_{ij}$$

where

$s_{ij}$  is the LLM's coverage score for node  $i$  on sample  $j$ .

Coverage quality ( $c_i$ ) is assessed based on LLM as a judge with the following prompt:

...

COVERAGE\_PROMPT = "You are a procurement domain expert evaluating ontology coverage.

Given the following **\*\*ontology node\*\*** (a concept from a procurement ontology):

Node: "{NODE\_NAME}"

And the following **dataset sample** (a question-answer pair from a procurement dataset):

""

{TEXT\_SAMPLE}

""

Rate how well this dataset sample **covers or addresses** the topic of the ontology node, on a scale from 0.0 to 1.0:

- 0.0 = the text has no relevant content about this node's topic
- 0.5 = the text partially or indirectly addresses this node's topic
- 1.0 = the text comprehensively covers this node's topic

Return ONLY a single number (e.g., 0.7). No extra text or explanation.

"

""

This approach provides a view of ontology completeness, emphasizing the importance of domain relevance and the quality of coverage in the evaluation process.

#### 4.4 LLMs fine-tuning

Full fine-tuning is a process in which all parameters of a pretrained LLM are updated using a task-specific dataset. This approach allows the model to adapt its representations and internal knowledge to the specific requirements of a downstream task, such as text similarity, classification, or question answering. By adjusting every parameter, full fine-tuning maximizes the model's ability to capture nuanced semantic relationships and improves task-specific performance compared to zero-shot or few-shot methods.

The theoretical foundation of full fine-tuning lies in leveraging pretrained LLM knowledge while enabling specialization through supervised learning. In practice, large datasets are required to achieve meaningful improvements, for example with Llama models full fine-tuning on 100,000 Quora Question Pair (QQP) samples improved the F1 score from 82.0% (previous S-CNN baseline) to 84.9%, a gain of only ~3%. This illustrates that while full fine-tuning can enhance performance, the gains are relatively modest compared to the computational and data resources required (Han, Shi, and Tsui, 2025). Smaller datasets often yield limited improvement and an increased risk of overfitting, motivating the use of parameter-efficient methods such as LoRA.

#### 4.5 Efficiency and Theoretical Basis of LoRA Fine-Tuning

Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning technique that modifies the weights of a pretrained model using low-rank decomposition. Instead of updating all parameters of the model, LoRA introduces trainable low-rank matrices A and B to adjust the original weight matrix W as follows:

$$W' = W + \Delta W = W + BA$$

Where:

$W \in \mathbb{R}^{d \times k}$  represents the original weight matrix;

$B \in \mathbb{R}^{d \times r}$  and  $A \in \mathbb{R}^{r \times k}$  are the low-rank matrices trained during fine-tuning;

The rank  $r$  is chosen such that  $r \ll \min(d, k)$ , ensuring that only a small fraction of parameters are updated.

By freezing the majority of the pretrained model's parameters and only training the low-rank matrices, LoRA preserves the general knowledge learned during pretraining while efficiently adapting the model to task-specific data (Dettmers et al. 2023). This enables the model to capture domain-specific patterns and distributions with limited data, improving performance metrics such as accuracy, recall, and F1 score. For instance, in domain-adaptation scenarios such as procurement, LoRA enables rapid, effective fine-tuning while significantly reducing computational cost compared to full fine-tuning.

## 5.0 Results

### 5.1 Data Processing Efficiency

The filtering process reduced the FineWeb dataset by 80%, producing a focused procurement-specific subset. Llama-3.2-3B annotation achieved fast processing speeds, averaging 3 seconds per sample on GPU and 15 seconds on CPU. The fine-tuned BERT-base-uncased classifier reached an F1 score of 75% on a balanced dataset, effectively identifying procurement-related content. Misclassifications were mainly in closely related domains, reflecting a conservative yet accurate performance suitable for ontology creation.

### 5.2 Ontology Development Outcomes

The semi-manual RDF-based ontology successfully addressed sparsity limitations, establishing core relationships among companies, procurement, supply chain, and goods/services. Hierarchical clustering and targeted LLMs-driven node expansion generated comprehensive relationship mappings, producing a high-quality, noise-free ontology. Part of generated ontology related to Procurement can be observed on Figure 3.

Expert validation resulted in a dataset of 140 samples covering key procurement topics with additional auto generated 460 samples. WCS analysis showed 76.42%, indicating strong topic coverage across the ontology nodes. The high overall score reflects both consistent dataset coverage across nodes and appropriately distributed importance weights, validating the ontology-guided dataset generation approach.

Table 2 formalizes the complete ontology structure depicted in Figure 3. The seed ontology (Appendix A) defines the expert-authored starting point: five core classes - Company, Business Domain, Goods and Services, Procurement, and Supply Chain - with four foundational relations. This seed was then expanded through LLM-guided node extraction to depth 2, yielding the full ontology (Appendix B) with 30 classes, 29 directed relations, and 24 unique predicate types.

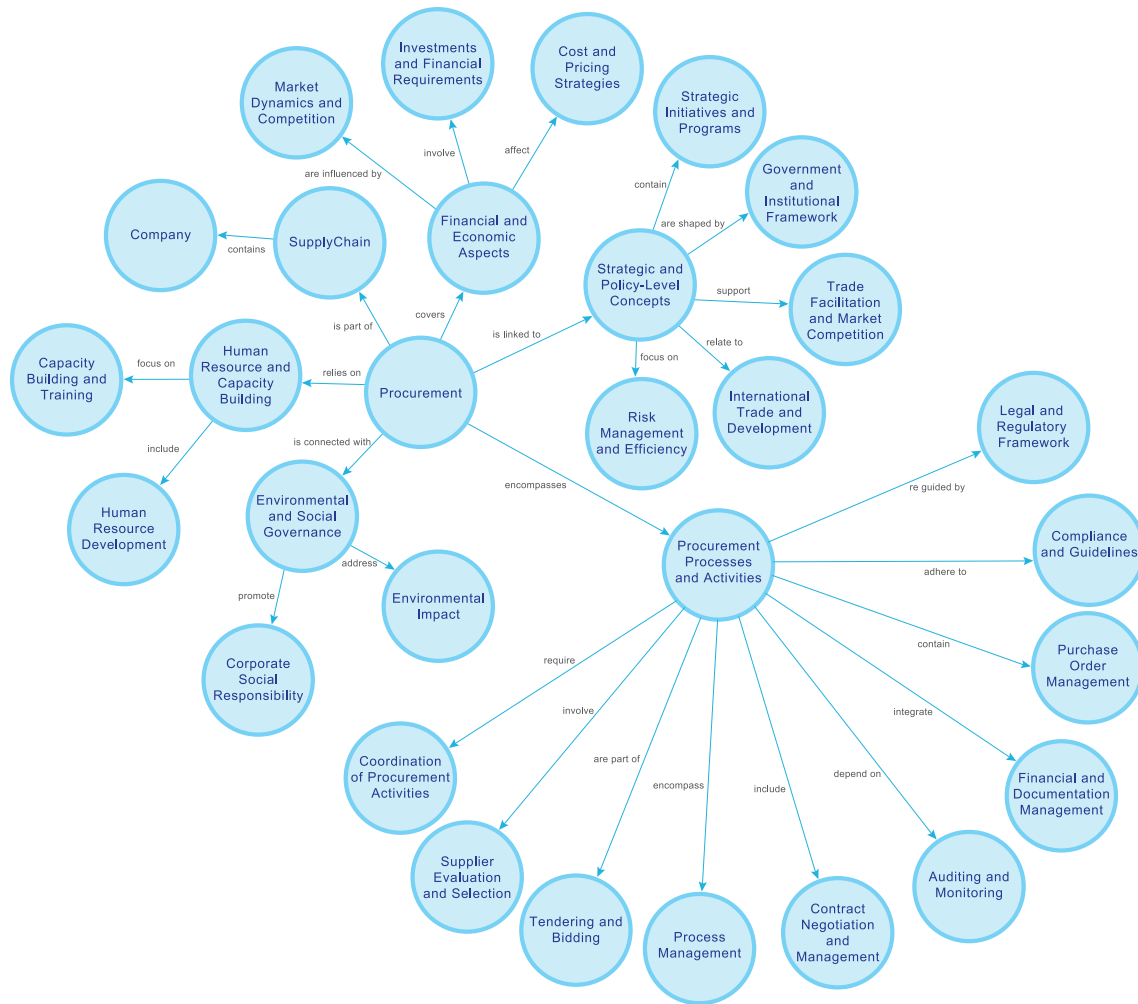


Fig. 3 Generated ontology using semi-automated ontology nodes expansion with depth

### 5.3 Fine-tuning Performance Gains

LoRA fine-tuning experiments on Llama-3.2-1B and Llama-3.2-3B models demonstrated that, when paired with our curated procurement dataset, domain-specific performance improved noticeably. The Llama-3.2-1B model showed the largest gains, indicating that smaller models benefit most from targeted domain data. The 3B model exhibited modest improvements, consistent with its larger parameter count. Qwen 3.5-35B-A3B judge evaluation on test samples confirmed that fine-tuned models trained on our dataset achieved greater domain knowledge and reduced factual inconsistencies. These results (Table 3) show that high-quality, focused datasets can effectively support parameter-efficient adaptation methods such as LoRA, enabling reliable domain specialization without full model fine-tuning. Win rates shown are based on blind A/B evaluation across 50 test samples and require larger-scale validation for statistical significance.

Our performance gains are in line with findings by (Lu et al. 2025), who showed that systematic fine-tuning strategies, including LoRA, supervised fine-tuning, and preference-based optimization can significantly enhance smaller model capabilities, often yielding improvements comparable to or exceeding those observed in larger models. This supports the conclusion that strategic low-rank adaptation is particularly effective for smaller LLMs, enabling rapid domain adaptation while maintaining computational efficiency on good quality datasets.

## 5.4 Training Convergence and Overfitting Analysis

Figure 4 presents the training and evaluation loss curves for both models with early stopping boundaries marked. For Llama-3.2-1B, training loss decreased steadily from approximately 3.0 to 1.2 over the training period, while evaluation loss converged to approximately 1.5 and plateaued from epoch 4 onward, with early stopping triggered at epoch 7. For Llama-3.2-3B, training loss decreased from approximately 2.7 to 1.0, with evaluation loss stabilizing around 1.2–1.3 and early stopping triggered at epoch 5. Critically, in neither model did the evaluation loss exhibit sustained upward divergence from the training loss, which would be the mark of overfitting. The gap between training and evaluation loss remained stable after convergence, indicating that the combination of LoRA's implicit regularization (updating <1% of parameters), dropout, and early stopping effectively prevented overfitting despite the modest training set size.



Fig. 4 Training and evaluation loss curves for Llama-3.2-1B (left) and Llama-3.2-3B (right) with early stopping boundaries marked

## 6.0 Discussion

### 6.1 Efficiency of small dataset Fine-tuning for small LLMs

The results demonstrate that even modest datasets can yield substantial performance improvements in small language models. The significant preference for the fine-tuned Llama-3.2-1B model (78.15% win rate for answer relevancy, 77.87% for faithfulness) in blind A/B testing with only 500 training pairs challenges conventional assumptions about dataset size requirements for effective fine-tuning. This finding has significant implications for resource-constrained environments where large-scale data collection is impractical or cost-prohibitive.

The differential response between Llama-3.2-1B and Llama-3.2-3B models suggests an optimal scaling relationship between dataset size and model capacity.

Table 2. Complete Ontology Structure

Depth	Parent Node	Relation	Child Node
0 → 1	Procurement	covers	Financial and Economic Aspects
0 → 1	Procurement	relies on	Human Resource and Capacity Building
0 → 1	Procurement	is connected with	Environmental and Social Governance
0 → 1	Procurement	is linked to	Strategic and Policy-Level Concepts
0 → 1	Procurement	encompasses	Procurement Processes and Activities
0 → 1	SupplyChain	is part of	Procurement
1 → 2	SupplyChain	contains	Company
1 → 2	Financial and Economic Aspects	involve	Investments and Financial Requirements
1 → 2	Financial and Economic Aspects	affect	Cost and Pricing Strategies
1 → 2	Financial and Economic Aspects	are influenced by	Market Dynamics and Competition
1 → 2	Human Resource and Capacity Building	focus on	Capacity Building and Training
1 → 2	Human Resource and Capacity Building	include	Human Resource Development
1 → 2	Environmental and Social Governance	address	Environmental Impact
1 → 2	Environmental and Social Governance	promote	Corporate Social Responsibility
1 → 2	Strategic and Policy-Level Concepts	contain	Strategic Initiatives and Programs
1 → 2	Strategic and Policy-Level Concepts	are shaped by	Government and Institutional Framework
1 → 2	Strategic and Policy-Level Concepts	support	Trade Facilitation and Market Competition
1 → 2	Strategic and Policy-Level Concepts	relate to	International Trade and Development
1 → 2	Strategic and Policy-Level Concepts	focus on	Risk Management and Efficiency
1 → 2	Procurement Processes and Activities	are guided by	Legal and Regulatory Framework
1 → 2	Procurement Processes and Activities	adhere to	Compliance and Guidelines
1 → 2	Procurement Processes and Activities	contain	Purchase Order Management
1 → 2	Procurement Processes and Activities	integrate	Financial and Documentation Management
1 → 2	Procurement Processes and Activities	depend on	Auditing and Monitoring
1 → 2	Procurement Processes and Activities	include	Contract Negotiation and Management
1 → 2	Procurement Processes and Activities	encompass	Process Management
1 → 2	Procurement Processes and Activities	are part of	Tendering and Bidding
1 → 2	Procurement Processes and Activities	involve	Supplier Evaluation and Selection
1 → 2	Procurement Processes and Activities	require	Coordination of Procurement Activities

The smaller model's higher win rate (78%) indicates more efficient parameter utilization relative to the available training data, while the 3B model's moderate win rates (68–72%) may reflect parameter underutilization given the limited dataset size.

The LoRA fine-tuning approach proves effective for compact model with small datasets, enabling significant performance enhancements without the computational overhead of full parameter updates. This efficiency becomes crucial for practical deployment scenarios where rapid iteration and resource optimization are essential.

The training convergence analysis (Figure 4) provides empirical evidence that overfitting was effectively mitigated despite the small dataset. Several complementary mechanisms contributed to this outcome: (1) LoRA's inherent regularization effect, which constrains adaptation to a low-rank subspace - updating only 0.90% and 0.75% of total parameters for

the 1B and 3B models, respectively - fundamentally limits the model's capacity to memorize training examples; (2) early stopping with patience of 2 epochs, which halted training before any sustained divergence between training and evaluation loss could develop; (3) LoRA dropout (0.05), which provided additional stochastic regularization during training. The stable evaluation loss plateaus observed for both models confirm that the learned adaptations generalize beyond the training set. Furthermore, the fact that the smaller Llama-3.2-1B model achieved higher win rates than the larger 3B model is itself evidence against overfitting: an overfitted model would perform well on training-like data but poorly on unseen test cases evaluated by an independent judge.

## 6.2 Applicability and Transferability of the Ontology Framework

The semi-automated ontology construction methodology was designed and validated within the procurement domain. While the results reported in this paper are specific to procurement, several architectural choices - RDF-based representation, hierarchical clustering for node grouping, and depth-limited LLM-guided expansion - were intentionally kept domain-agnostic. These design decisions suggest that the methodology could be adapted to other specialized domains, although such transferability remains to be empirically validated.

Table 3. Blind A/B test win rates of fine-tuned llama models over base models on 50 procurement test cases

Metric	Finetuned Llama-3.2-1B	Finetuned Llama-3.2-1B 95% CI	Finetuned Llama-3.2-3B	Finetuned Llama-3.2-3B 95% CI
Answer Relevancy	78.15%	[66%, 88%]	68.35%	[54%, 80%]
Faithfulness	77.87%	[66%, 90%]	72.29%	[60%, 84%]
FCR	77.95%	[66%, 90%]	72.36%	[60%, 84%]

The structured RDF foundation combined with expert-guided validation creates a balanced approach that leverages both automated efficiency and human domain expertise. This hybrid methodology addresses the scalability challenges inherent in purely manual ontology construction while maintaining quality standards that can be difficult to achieve with fully automated approaches.

The weighted coverage scoring system provides an objective assessment mechanism for guiding iterative ontology expansion and identifying knowledge gaps within procurement sub-domains. In principle, the same scoring approach could be applied to other domains where hierarchical concept coverage is relevant, but additional domain-specific validation would be required to confirm its effectiveness. The ability to identify and prioritize knowledge gaps through the weighted coverage scoring offers practical guidance for iterative ontology development, making the methodology particularly valuable for complex domains with extensive conceptual hierarchies.

## 6.3 Limitations

Several limitations of this work should be acknowledged. First, our fine-tuning experiments exclusively used LoRA as the parameter-efficient adaptation method. While alternative strategies exist - including full fine-tuning, QLoRA, adapter-based methods, and preference optimization techniques such as DPO and RLHF - systematic comparison across fine-tuning strategies was outside the primary scope of this work, which focused on ontology-guided dataset construction rather than fine-tuning methodology. The selection of LoRA was motivated by its established effectiveness for small-dataset domain adaptation and computational efficiency, as discussed in Section 3E. Future work could extend the evaluation to additional fine-tuning approaches to determine whether alternative strategies yield further improvements when combined with ontology-guided datasets.

Second, to our knowledge, no standardized benchmark dataset exists for procurement domain question answering that would allow direct comparison with prior work.

Existing procurement-related corpora are either proprietary, narrowly scoped to specific sub-tasks (e.g., contract clause extraction), or not publicly available. As a result, our evaluation relies on blind A/B testing with LLM-as-a-judge methodology rather than comparison against an established leaderboard. We acknowledge this as a limitation and note that developing and releasing a standardized procurement QA benchmark would be a valuable contribution to the community. As a next step, we plan to evaluate our fine-tuned models on general-purpose, non-procurement QA benchmarks to verify that domain-specific fine-tuning does not degrade general capabilities, and to provide an additional external reference point for the observed performance gains.

Third, expert validation of the 140 manually curated samples was performed by a single procurement domain specialist across three iterative review rounds. While the multi-round review process helped ensure consistency, the absence of multiple independent annotators means that formal inter-annotator agreement could not be measured. Future work should involve additional domain experts to enable agreement metrics and further strengthen validation reliability.

Fourth, the current evaluation compares fine-tuned models against base models but does not include component-level ablation studies that would isolate the contribution of individual pipeline stages. Specifically, we did not compare: (a) ontology-guided versus non-ontology-guided dataset generation, (b) expert-validated versus automatically generated QA pairs trained separately, or (c) fine-tuned models versus prompt-engineered baselines (e.g., few-shot or retrieval-augmented generation over the same procurement corpus). As the primary scope of this work was establishing the end-to-end ontology construction and dataset generation methodology, these ablations were deferred to future work. Additionally, we plan to evaluate the fine-tuned models on general-purpose QA benchmarks to assess whether procurement-specific fine-tuning preserves or degrades general knowledge capabilities, providing a broader reference point for the observed domain performance gains.

## 7.0 Conclusions

This study establishes a novel paradigm for domain-specific knowledge extraction and model fine-tuning that addresses fundamental limitations in traditional approaches like triplet extraction. The semi-automated ontology construction framework successfully bridges the gap between expert knowledge and semi-automated processing, demonstrating practical applicability in the procurement domain, with architectural choices that may support adaptation to other specialized domains pending further validation.

The procurement domain implementation validates two critical findings that reshape the understanding of domain-specific knowledge extraction. First, compact language models can achieve substantial performance improvements with remarkably small, high-quality datasets when guided by structured domain knowledge. The 600-sample dataset generated moderate improvements in key metrics for both models, demonstrating that quality is more important than quantity in specialized applications.

Second, the proposed semi-automated ontology construction methodology provides a robust and manageable way of creating high-quality knowledge representations that overcome the sparsity limitations of traditional knowledge graph approaches. The systematic nodes expansion steps, hierarchical processing mechanisms, and weighted coverage assessment enable replicable ontology development across diverse sub domains of procurement while maintaining expert validation standards.

The methodology's broader impact extends beyond immediate performance improvements, offering a scalable template for enterprise knowledge management and the development of domain specific LLMs. The integration of automated processing with expert guidance

provides practical solutions for organizations seeking to leverage specialized knowledge without prohibitive resource investments.

These findings offer evidence that small, high-quality datasets can be effective for domain-specific model fine-tuning and provide a practical framework for systematic knowledge representation that, while validated in procurement, may inform similar efforts in other specialized domains.

## References

[1] Akinci, F. S., & Tuğlular, T. (2025). A contract-driven automated unit test maintenance approach with generative artificial intelligence for backend software projects. *Journal of Smart Systems*, 4(2), 74-97.

[1] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, and et al., "The Llama 3 Herd of Models," arXiv preprint arXiv:2407.21783, 2024. <https://doi.org/10.48550/arXiv.2407.21783> .

[2] A. Halike, A. Wumaier, and T. Yibulayin, "Zero-shot relation triple extraction with prompts for low-resource languages," *Applied Sciences*, vol. 13, no. 7, p. 4636, 2023. <https://doi.org/10.3390/app13074636> .

[3] A. Hogan, E. Blomqvist, M. Cochez, C. D'Amato, G. de Melo, C. Gutiérrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann, "Knowledge graphs," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–37, 2021. <https://doi.org/10.1145/3447772> .

[4] A. Mavridis, S. Tegos, C. Anastasiou, M. Papoutsoglou, and G. Meditskos, "Large language models for intelligent RDF knowledge graph construction: Results from medical ontology mapping," *Frontiers in Artificial Intelligence*, vol. 8, p. 1546179, 2025. <https://doi.org/10.3389/frai.2025.1546179> .

[5] A. Sattar, M. N. Ahmad, E. S. M. Surin, and A. K. Mahmood, "An improved methodology for collaborative construction of reusable, localized, and shareable ontology," *IEEE Access*, vol. 9, pp. 17463–17484, 2021. <https://doi.org/10.1109/ACCESS.2021.3054412> .

[6] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, D. Liu, D. Shi, E. Wu, F. Wang, F. Li, G. Chen, G. Zhang, H. Lin, H. Zhou, H. Wang, I. Zhao, J. Chen, J. Li, J. Yang, K. Liu, K. Zhang, L. Sun, L. Wang, L. Li, M. Xu, M. Zhang, N. Li, P. Wang, Q. Zhao, R. Liu, S. Chen, S. Zhang, T. Li, T. Yang, W. Zhang, W. Xu, X. Liu, X. Wang, Y. Zhao, Y. Qiu, Z. Liu, Z. Chen, Z. Qiu, "Qwen3 Technical Report," arXiv preprint arXiv:2505.09388, 2025. <https://doi.org/10.48550/arXiv.2505.09388> .

[7] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne, "Knowledge Graphs: Opportunities and Challenges," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 13071–13102, 2023. <https://doi.org/10.1007/s10462-023-10465-9> .

[8] D. Doumanas, A. Soularidis, D. Spiliotopoulos, C. Vassilakis, and K. Kotis, "Fine-tuning large language models for ontology engineering: A comparative analysis of GPT-4 and Mistral," *Applied Sciences*, vol. 15, no. 4, p. 2146, 2025. <https://doi.org/10.3390/app15042146> .

[9] D. Wesslund, V. Stenström, P. Linde, and A. Holmberg, "LLM based triplet extraction from financial reports," arXiv preprint arXiv:2602.11886, 2026. <https://doi.org/10.48550/arXiv.2602.11886> .

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021. <https://doi.org/10.48550/arXiv.2106.09685> .

[11] G. Penedo, H. Kydlíček, L. Ben Allal, A. Lozhkov, M. Mitchell, C. Raffel, L. von Werra, and T. Wolf, "The FineWeb datasets: Decanting the web for the finest text data at scale," arXiv preprint arXiv:2406.17557, 2024. <https://doi.org/10.48550/arXiv.2406.17557> .

[12] H. Jayadianti, "Mapping relational databases to RDF using direct mapping for Indonesian movies ontology," in *MATEC Web of Conferences*, vol. 372, p. 04011, 2022. <https://doi.org/10.1051/mateconf/202237204011> .

- [13] H. Zeng, M. Zhang, Y. Xia, and R. Chen, "Decoupling the Depth and Scope of Graph Neural Networks," arXiv preprint arXiv:2201.07858, 2022. <https://doi.org/10.48550/arXiv.2201.07858> .
- [14] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre training of Deep Bidirectional Transformers for Language Understanding," in Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long and Short Papers), Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186. <https://doi.org/10.18653/v1/N19.1423> .
- [15] J. Fan, X. Tian, C. Lv, S. Zhang, Y. Wang, and J. Zhang, "Extractive social media text summarization based on MFMMR-BertSum," *Array*, vol. 18, p. 100322, 2023. <https://doi.org/10.1016/j.array.2023.100322>
- [16] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B. Agüera y Arcas, N. Tomasev, Y. Liu, R. Wong, C. Semturs, S. Sara Mahdavi, J. Barral, D. Webster, G. S. Corrado, Y. Matias, and A. Karthikesalingam, "Towards Expert-Level Medical Question Answering with Large Language Models," arXiv preprint arXiv:2305.09617, 2023. <https://doi.org/10.48550/arXiv.2305.09617> .
- [17] K. Wang, J. Zhu, M. Ren, Z. Liu, S. Li, Z. Zhang, C. Zhang, X. Wu, Q. Zhan, Q. Liu, and Y. Wang, "A Survey on Data Synthesis and Augmentation for Large Language Models," arXiv preprint arXiv:2410.12896, 2024. <https://doi.org/10.48550/arXiv.2410.12896> .
- [18] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," arXiv preprint arXiv:2203.02155, 2022. <https://doi.org/10.48550/arXiv.2203.02155> .
- [19] M. Guida, F. Caniato, A. Moretto, and S. Ronchi, "The role of artificial intelligence in the procurement process: State of the art and research agenda," *Journal of Purchasing and Supply Management*, vol. 29, no. 2, 100823, 2023. <https://doi.org/10.1016/j.pursup.2023.100823> .
- [20] M. Guida, F. Caniato, and A. Moretto, "AI meets spend classification: A new frontier in information processing," *Journal of Purchasing and Supply Management*, vol. 31, no. 3, 100993, 2025. <https://doi.org/10.1016/j.pursup.2025.100993> .
- [21] M. Koniaris, D. Galanis, E. Giannini, and P. Tsanakas, "Evaluation of automatic legal text summarization techniques for Greek case law," *Information*, vol. 14, no. 4, p. 250, 2023. <https://doi.org/10.3390/info14040250> .
- [22] M. M. Karim, S. Khan, D. H. Van, X. Liu, C. Wang, and Q. Qu, "Transforming data annotation with AI agents: A review of architectures, reasoning, applications, and impact," *Future Internet*, vol. 17, no. 8, p. 353, 2025. <https://doi.org/10.3390/fi17080353> .
- [23] M. S. Baysan, S. Uysal, İ. İşlek, Ç. Çiğ Karaman, and T. Güngör, "LLM-as-a-Judge: Automated evaluation of search query parsing using large language models," *Frontiers in Big Data*, vol. 8, p. 1611389, 2025. <https://doi.org/10.3389/fdata.2025.1611389> .
- [24] *Ontology Development 101: A Guide to Creating Your First Ontology*, N. F. Noy and D. L. McGuinness, Stanford University, Stanford, CA, USA, Technical Report, 2001. [https://protege.stanford.edu/publications/ontology\\_development/ontology101.pdf](https://protege.stanford.edu/publications/ontology_development/ontology101.pdf) .
- [25] P. Colombo, T. P. Pires, M. Boudiaf, D. Culver, R. Melo, C. Corro, A. F. T. Martins, F. Esposito, V. L. Raposo, S. Morgado, and M. Desa, "SaulLM 7B: A pioneering Large Language Model for Law," arXiv preprint arXiv:2403.03883, 2024. <https://doi.org/10.48550/arXiv.2403.03883> .
- [26] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval Augmented Generation for Knowledge Intensive NLP Tasks," arXiv preprint arXiv:2005.11401, 2020. <https://doi.org/10.48550/arXiv.2005.11401> .
- [27] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," arXiv preprint arXiv:2305.18290, 2023. <https://doi.org/10.48550/arXiv.2305.18290> .
- [28] S. E. Safitri, W. D. Yuniarti, M. R. Handayani, and K. Umam, "User opinion mining on the Maxim application reviews using BERT-Base Multilingual Uncased," *Jurnal Sisfokom*, vol. 14, no. 3, pp. 365–372, 2025. <https://doi.org/10.32736/sisfokom.v14i3.2391> .

- [29] S. Han, L. Shi, and F. R. Tsui, "Enhancing semantical text understanding with fine-tuned large language models: A case study on Quora Question Pair duplicate identification," *PLoS ONE*, vol. 20, no. 1, p. e0317042, 2025. <https://doi.org/10.1371/journal.pone.0317042>.
- [30] S. Herold, J. Heller, F. Rozemeijer, and D. Mahr, "Brave new procurement deals: An experimental study of how generative artificial intelligence reshapes buyer-supplier negotiations," *Journal of Purchasing and Supply Management*, vol. 31, no. 4, 101012, 2025. <https://doi.org/10.1016/j.pursup.2025.101012>.
- [31] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, 2022. <https://doi.org/10.1109/TNNLS.2021.3070843>.
- [32] S. Kim, D. Kim, C. Park, W. Lee, W. Song, Y. Kim, H. Kim, Y. Kim, H. Lee, J. Kim, C. Ahn, S. Yang, S. Lee, H. Park, G. Gim, M. Cha, H. Lee, and S. Kim, "SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, Mexico City, Mexico, pp. 23–35, Jun. 2024. <https://doi.org/10.18653/v1/2024.naacl-industry.3>.
- [33] S. Toro, A. V. Anagnostopoulos, S. M. Bello, K. Blumberg, R. Cameron, L. Carmody, A. D. Diehl, D. M. Dooley, W. D. Duncan, P. Fey, P. Gaudet, N. L. Harris, M. P. Joachimiak, L. Kiani, T. Lubiana, M. C. Munoz-Torres, S. O'Neil, D. Osumi-Sutherland, A. Puig-Barbe, J. T. Reese, L. Reiser, S. M. C. Robb, T. Ruemping, J. Seager, E. Sid, R. Stefancsik, M. Weber, V. Wood, M. A. Haendel, and C. J. Mungall, "Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI)," *Journal of Biomedical Semantics*, vol. 15, no. 1, pp. 1–16, 2024. <https://doi.org/10.1186/s13326-024-00320-3>.
- [34] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. S. Rosenberg, and G. Mann, "BloombergGPT: A Large Language Model for Finance," *arXiv preprint arXiv:2303.17564*, 2023. <https://doi.org/10.48550/arXiv.2303.17564>.
- [35] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," *arXiv preprint arXiv:2305.14314*, 2023. <https://doi.org/10.48550/arXiv.2305.14314>.
- [36] T. Hossain, K. M. Saifuddin, M. I. K. Islam, F. Tanvir, and E. Akbas, "Tackling oversmoothing in GNN via graph sparsification," in *Machine Learning and Knowledge Discovery in Databases: Research Track and Demo Track – European Conference, ECML PKDD 2024, Proceedings, Lecture Notes in Computer Science*, Springer, 2024, pp. 161–179. [https://doi.org/10.1007/978-3-031-70371-3\\_10](https://doi.org/10.1007/978-3-031-70371-3_10).
- [37] T. K. Rusch, M. M. Bronstein, and S. Mishra, "A survey on oversmoothing in graph neural networks," *arXiv preprint arXiv:2303.10993*, 2023. <https://doi.org/10.48550/arXiv.2303.10993>.
- [38] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993. <https://doi.org/10.1006/knac.1993.1008>.
- [39] W. Lu, R. K. Luu, and M. J. Buehler, "Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities," *npj Computational Materials*, vol. 11, p. 84, 2025. <https://doi.org/10.1038/s41524-025-01564-y>.
- [40] X. Hao, Z. Ji, X. Li, L. Yin, L. Liu, M. Sun, Q. Liu, and R. Yang, "Construction and application of a knowledge graph," *Remote Sensing*, vol. 13, no. 13, p. 2511, 2021. <https://doi.org/10.3390/rs13132511>.
- [41] X. Wu, Z. Chen, W. Wang, and A. Jadbabaie, "A non-asymptotic analysis of oversmoothing in graph neural networks," *arXiv preprint arXiv:2212.10701*, 2023. <https://doi.org/10.48550/arXiv.2212.10701>.
- [42] Y. Dai, S. Wang, N. N. Xiong, and W. Guo, "A survey on knowledge graph embedding: Approaches, applications and benchmarks," *Electronics*, vol. 9, no. 5, p. 750, 2020. <https://doi.org/10.3390/electronics9050750>.
- [43] Y. Kong, X. Liu, Z. Zhao, D. Zhang, and J. Duan, "Bolt defect classification algorithm based on knowledge graph and feature fusion," *Energy Reports*, vol. 8, suppl. 1, pp. 856–863, 2022. <https://doi.org/10.1016/j.egy.2021.11.127>.
- [44] Z. Bahroun, A. Saihi, R. Asad, and M. Tanash, "A systematic analysis of generative artificial intelligence for supply chain transformation," *Supply Chain Analytics*, vol. 13, 100188, 2026. <https://doi.org/10.1016/j.sca.2025.100188>.

[45] Z. Wu, Y. Zhang, and H. Li, "A semi-automatic ontology development framework for safety requirements," *Buildings*, vol. 15, no. 4, p. 569, 2025. <https://doi.org/10.3390/buildings15040569> .

## Appendix A. Seed Ontology (RDF Turtle)

```

@prefix : <http://example.org/procurement#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

# =====
# Listing 1 – Seed Ontology (RDF Turtle)
# The expert-defined starting point used for LLM-guided
# node expansion (Section 2.B).
# =====

# --- Core Classes (5) -----
:Company      a owl:Class ; rdfs:label "Company" .
:BusinessDomain  a owl:Class ; rdfs:label "Business Domain" .
:GoodsAndServices a owl:Class ; rdfs:label "Goods and Services" .
:Procurement    a owl:Class ; rdfs:label "Procurement" .
:SupplyChain    a owl:Class ; rdfs:label "Supply Chain" .

# --- Core Object Properties (4) -----
:provides      a owl:ObjectProperty ; rdfs:label "provides" .
:canBeAppliedTo a owl:ObjectProperty ; rdfs:label "can be applied to" .
:contains      a owl:ObjectProperty ; rdfs:label "contains" .
:includes      a owl:ObjectProperty ; rdfs:label "includes" .

# --- Core Assertions (4) -----
:Company      :provides      :GoodsAndServices .
:Procurement  :canBeAppliedTo :BusinessDomain .
:SupplyChain  :contains      :Company .
:SupplyChain  :includes      :Procurement .

```

## Appendix B. Expanded Ontology (RDF Turtle)

```

@prefix : <http://example.org/procurement#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

# =====
# Expanded Procurement Ontology (Depth 0-2)
# Generated via semi-automated LLM-guided node expansion
# from the seed ontology (Listing 1).
# 30 classes | 29 relations | 24 unique predicate types
# =====

# ===== DEPTH 0 (Root) =====

:Procurement a owl:Class ;
  rdfs:label "Procurement" ;
  :covers      :Financial_and_Economic_Aspects ;
  :relies_on   :Human_Resource_and_Capacity_Building ;
  :is_connected_with :Environmental_and_Social_Governance ;
  :is_linked_to :Strategic_and_Policy-Level_Concepts ;
  :encompasses :Procurement_Processes_and_Activities .

# ===== DEPTH 1 =====

:SupplyChain a owl:Class ;
  rdfs:label "Supply Chain" ;
  :contains :Company ;
  :is_part_of :Procurement .

:Financial_and_Economic_Aspects a owl:Class ;
  rdfs:label "Financial and Economic Aspects" ;
  :involve :Investments_and_Financial_Requirements ;
  :affect :Cost_and_Pricing_Strategies ;
  :are_influenced_by :Market_Dynamics_and_Competition .

:Human_Resource_and_Capacity_Building a owl:Class ;
  rdfs:label "Human Resource and Capacity Building" ;
  :focus_on :Capacity_Building_and_Training ;
  :include :Human_Resource_Development .

:Environmental_and_Social_Governance a owl:Class ;
  rdfs:label "Environmental and Social Governance" ;
  :address :Environmental_Impact ;
  :promote :Corporate_Social_Responsibility .

:Strategic_and_Policy-Level_Concepts a owl:Class ;
  rdfs:label "Strategic and Policy-Level Concepts" ;
  :contain :Strategic_Initiatives_and_Programs ;
  :are_shaped_by :Government_and_Institutional_Framework ;
  :support :Trade_Facilitation_and_Market_Competition ;
  :relate_to :International_Trade_and_Development ;
  :focus_on :Risk_Management_and_Efficiency .

:Procurement_Processes_and_Activities a owl:Class ;
  rdfs:label "Procurement Processes and Activities" ;
  :are_guided_by :Legal_and_Regulatory_Framework ;

```

```

:adhere_to :Compliance_and_Guidelines ;
:contain :Purchase_Order_Management ;
:integrate :Financial_and_Documentation_Management ;
:depend_on :Auditing_and_Monitoring ;
:include :Contract_Negotiation_and_Management ;
:encompass :Process_Management ;
:are_part_of :Tendering_and_Bidding ;
:involve :Supplier_Evaluation_and_Selection ;
:require :Coordination_of_Procurement_Activities .

```

```
# ===== DEPTH 2 =====
```

```
# -- Children of SupplyChain --
```

```
:Company a owl:Class ;
  rdfs:label "Company" .
```

```
# -- Children of Financial and Economic Aspects --
```

```
:Investments_and_Financial_Requirements a owl:Class ;
  rdfs:label "Investments and Financial Requirements" .
```

```
:Cost_and_Pricing_Strategies a owl:Class ;
  rdfs:label "Cost and Pricing Strategies" .
```

```
:Market_Dynamics_and_Competition a owl:Class ;
  rdfs:label "Market Dynamics and Competition" .
```

```
# -- Children of Human Resource and Capacity Building --
```

```
:Capacity_Building_and_Training a owl:Class ;
  rdfs:label "Capacity Building and Training" .
```

```
:Human_Resource_Development a owl:Class ;
  rdfs:label "Human Resource Development" .
```

```
# -- Children of Environmental and Social Governance --
```

```
:Environmental_Impact a owl:Class ;
  rdfs:label "Environmental Impact" .
```

```
:Corporate_Social_Responsibility a owl:Class ;
  rdfs:label "Corporate Social Responsibility" .
```

```
# -- Children of Strategic and Policy-Level Concepts --
```

```
:Strategic_Initiatives_and_Programs a owl:Class ;
  rdfs:label "Strategic Initiatives and Programs" .
```

```
:Government_and_Institutional_Framework a owl:Class ;
  rdfs:label "Government and Institutional Framework" .
```

```
:Trade_Facilitation_and_Market_Competition a owl:Class ;
  rdfs:label "Trade Facilitation and Market Competition" .
```

```
:International_Trade_and_Development a owl:Class ;
  rdfs:label "International Trade and Development" .
```

```
:Risk_Management_and_Efficiency a owl:Class ;
  rdfs:label "Risk Management and Efficiency" .
```

```
# -- Children of Procurement Processes and Activities --
```

```
:Legal_and_Regulatory_Framework a owl:Class ;
  rdfs:label "Legal and Regulatory Framework" .
```

:Compliance\_and\_Guidelines a owl:Class ;  
rdfs:label "Compliance and Guidelines" .

:Purchase\_Order\_Management a owl:Class ;  
rdfs:label "Purchase Order Management" .

:Financial\_and\_Documentation\_Management a owl:Class ;  
rdfs:label "Financial and Documentation Management" .

:Auditing\_and\_Monitoring a owl:Class ;  
rdfs:label "Auditing and Monitoring" .

:Contract\_Negotiation\_and\_Management a owl:Class ;  
rdfs:label "Contract Negotiation and Management" .

:Process\_Management a owl:Class ;  
rdfs:label "Process Management" .

:Tendering\_and\_Bidding a owl:Class ;  
rdfs:label "Tendering and Bidding" .

:Supplier\_Evaluation\_and\_Selection a owl:Class ;  
rdfs:label "Supplier Evaluation and Selection" .

:Coordination\_of\_Procurement\_Activities a owl:Class ;  
rdfs:label "Coordination of Procurement Activities" .