Transforming Scientific Papers into Structured Knowledge: Construction of a Knowledge Graph for BiS_2-Layered Superconductors

Author: Darío Santos Prego

Tutor: Matías Núñez

Course: Masters Degree in Computational Linguistics and LLMs

Institution: Universidad de La Rioja

Table of contents

0. ABSTRACT

The rapid expansion of materials science literature has produced a fragmentation problem in which critical experimental knowledge remains embedded in unstructured text rather than accessible data. This dissertation addresses that challenge in the domain of BiS_2-based layered superconductors by developing an automated information extraction pipeline and a structured domain knowledge graph.

The work operationalizes large language models (LLMs) to transform raw arXiv manuscripts into a queryable, topologically consistent representation of chemical compositions, synthesis conditions, and physical properties. A modular, verifiable pipeline was designed under a literate programming framework, enabling transparent data lineage and reproducibility. A curated corpus of 122 papers was constructed through iterative Boolean retrieval and rule-based segmentation. Semantic extraction proceeded in stages: claim classification, schema-constrained named entity recognition, and legality-aware relation extraction.

The modular strategy proved substantially more robust than monolithic LLM inference. Strict segmentation reduced boundary failures by 90%, enabling stable schema adherence. The resulting knowledge graph contains 2,035 unique entities and 5,743 relations. Network analysis reveals critical temperature (Tc) and $LaO_{0.5}F_{0.5}BiS_2$ as central hubs, while community detection identifies 51 coherent research subdomains.

These results demonstrate that constrained, claim-level LLM inference can construct a verifiable digital twin of a scientific field, transforming narrative literature into structured, computational knowledge and enabling reproducible, fine-grained scientific analysis.

1. INTRODUCTION

1.1 Background

The field of high-temperature superconductivity has long been driven by the discovery of new material families that challenge existing theoretical frameworks and offer novel pathways for technological application. In 2012, Mizuguchi et al. reported the discovery of superconductivity in BiS_2-based layered materials, specifically in the $Bi_4O_4S_3$ system, marking the emergence of a new family of layered superconductors. This discovery was pivotal because it introduced a structural motif distinct from the well-established copper-oxide (cuprate) and iron-based superconductors, yet sharing the essential feature of a layered crystal structure conducive to two-dimensional electronic transport.

The fundamental building block of this family is the rock-salt-type blocking layer (e.g., LnO, where Ln is a lanthanide) alternating with superconducting BiS_2 conduction layers as shown in Figure 1.1:

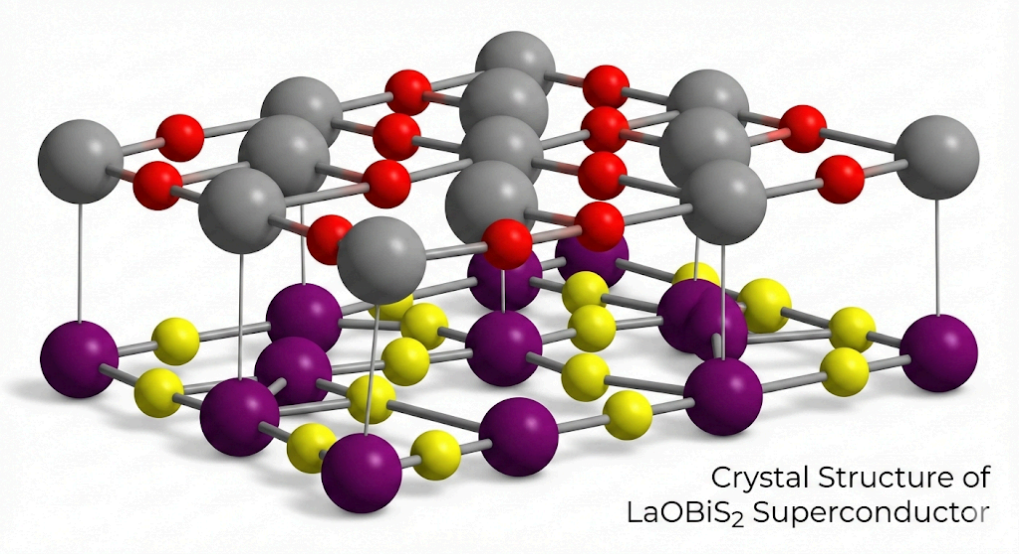


Figure 1.1 represents the structure of the LaOBiS2 superconductor, one of the most relevant compounds of the BiS2 Family

Mizuguchi (2019) elucidates that the superconductivity in these systems is induced by carrier doping into the BiS_2 layers, typically achieved through chemical substitution in the blocking layers, such as substituting Oxygen with Fluorine in $LaO_{1-x}F_xBiS_2$. This structural flexibility allows for a wide variety of chemical compositions, leading to a rich phase diagram that includes diverse phenomena such as coexistence with ferromagnetism and significant spin-orbit coupling effects.

Since the initial discovery, the family has expanded rapidly. Yazici et al. (2015) highlighted the synthesis of various analogues, including $REO_{1-x}F_xBiS_2$ (where RE = La, Ce, Pr, Nd, Yb), demonstrating that the superconducting transition temperature (T_c) is highly sensitive to the local crystal structure, particularly the in-plane chemical pressure and the bonding environment of the Bismuth and Sulfur atoms. The electronic properties of these materials are equally intriguing; Querales et al. (2016) provided an in-depth analysis of the normal state electronic properties, revealing that the conduction bands are primarily derived from Bi-6p orbitals, which is a significant departure from the d-orbital dominance in iron-based superconductors.

Recent research has focused heavily on elucidating the pairing mechanism governing these materials. While early consensus pointed towards conventional phonon-mediated superconductivity, Hoshi & Mizuguchi (2021) have discussed experimental evidence suggesting that weak coupling Bardeen-Cooper-Schrieffer (BCS) theory may not fully explain the behavior observed in all variants, particularly those with higher T_c values achieved under high pressure. Most recently, Romano et al. (2024) reviewed point-contact spectroscopy data, reinforcing the complexity of the superconducting gap structure and suggesting that while the BiS_2 family shares traits with conventional superconductors, the interplay of spin-orbit coupling and specific structural parameters introduces unique physical regimes that remain a subject of active debate.

1.2 Problem Statement

The rapid expansion of the BiS_2 research field, while scientifically fruitful, has created a significant challenge in data management and synthesis. The volume of scientific literature has

grown exponentially, dispersing critical experimental data—such as T_c values, doping concentrations, and synthesis protocols—across hundreds of isolated PDF documents. This phenomenon, often termed "information fragmentation," severely hampers the ability of researchers to identify holistic trends or perform meta-analyses.

Traditional methods of literature review are inherently limited by human cognitive bandwidth. A researcher attempting to correlate specific synthesis parameters (e.g., annealing temperature) with superconducting properties across the entire BiS_2 corpus faces a daunting task. As noted in the comprehensive review by Romano et al. (2024), the sheer diversity of experimental setups and sample qualities makes manual comparison prone to error and omission. The "manual synthesis" approach is not only time-consuming but also static; a review paper is often outdated by the time it is published.

Furthermore, valuable negative results or subtle correlations—such as the precise impact of varying the RE ionic radius on the Fermi surface topology described by Querales et al. (2016) —are often buried in the text of experimental sections, inaccessible to standard keyword searches. This lack of machine-readable structured data creates a bottleneck in materials informatics, preventing the application of data-driven discovery methods that have proven successful in other domains.

1.3 Objectives

To address the challenge of information fragmentation in the BiS_2 domain, this dissertation aims to develop an automated Information Extraction (IE) pipeline capable of transforming unstructured scientific text into a structured Knowledge Graph (KG). The primary objective is to leverage the capabilities of modern Large Language Models (LLMs) to perform tasks that were previously reliant on rigid rule-based systems or manual curation.

While foundational work by Tshitoyan et al. (2019) demonstrated that unsupervised word embeddings could capture latent materials knowledge from large corpora, such methods often lack the granular precision required to extract complex relationships, such as specific

property-condition dependencies. Similarly, Kononova et al. (2019) successfully text-mined synthesis recipes, yet their approach relied heavily on complex, domain-specific rule sets.

Recent perspectives in AI for science propose the notion of a "Materials Science GPT"—a system that develops scientific reasoning capabilities through large-scale exposure to domain literature, analogous to how human researchers accumulate knowledge through reading. Within this framework, the scientific corpus functions as the collective memory of a research community. The present work contributes to this paradigm not by training such a model directly, but by constructing the structured knowledge substrate required to support it. By transforming unstructured literature on BiS_2-based superconductors into an explicit knowledge graph of materials, methods, conditions, and properties, this dissertation builds a machine-interpretable representation of the domain's accumulated scientific understanding. Operationalising this vision requires concrete steps in corpus construction, semantic extraction, and structured knowledge integration.

This research adopts an emerging paradigm in materials informatics by utilising the generative and contextual modelling capabilities of Transformer-based models (Vaswani et al., 2017). Specifically, the project objectives are threefold. First, it seeks to construct a domain-specific corpus by systematically retrieving and curating a high-fidelity dataset of peer-reviewed papers and preprints strictly related to BiS_2-based layered superconductors. Second, it aims to develop an LLM-driven extraction pipeline by implementing and optimising open-weights LLMs—specifically Gemma 2 and Qwen 3—to perform Named Entity Recognition (NER) and Relation Extraction (RE) without the need for extensive supervised training data. Finally, the research intends to synthesise a Knowledge Graph by aggregating extracted entities, such as materials, transition temperatures (T_c), and synthesis methods, into a graph database. This structure will facilitate the visualisation of the research ecosystem, the identification of influential works, and the revelation of hidden connections between structural parameters and physical properties.

1.4 Scope and Limitations

The scope of this research is delineated by the specific material domain of interest and the computational parameters of the experimental environment. Geographically and scientifically, the study is strictly confined to the BiS_{2}-based family of superconductors and their direct structural analogues. While the methodology is potentially generalisable, this dissertation does not extend its analysis to broader classes such as Iron-based or Cuprate superconductors, unless they serve as explicit comparative subjects within the BiS_2 literature. Furthermore, the text corpus is derived exclusively from the arXiv repository. While this approach facilitates open-access retrieval of full-text data, it inherently introduces a potential selection bias by excluding older publications or research restricted to journals that do not permit preprint archiving. Additionally, the linguistic scope is restricted entirely to English-language scientific discourse.

Computational implementation is similarly bounded by the use of "open-weights" models, specifically the Gemma 2 and Qwen 3 architectures, executed on research-grade hardware such as the NVIDIA T4 GPU. This necessitated a focus on parameter-efficient inference techniques, most notably 4-bit quantisation, rather than the computationally prohibitive task of training foundational models from scratch. It is important to note that the system is architected to extract and categorise explicit claims as stated within the source text; it does not attempt to adjudicate the scientific validity of these claims, nor does it perform ab initio physics simulations to verify reported values. The resulting Knowledge Graph is therefore a reflection of the published literature's assertions rather than a primary source of physical verification.

2. LITERATURE REVIEW

2.1 Layered Superconductors: Crystal Structures, Critical Temperatures, and the *BiS_2* Family

Layered superconductors represent a broad and significant class of materials in condensed matter physics, characterised by anisotropic crystal structures that critically influence their superconducting properties. In these systems, superconductivity typically arises within specific conduction layers separated by insulating or "blocking" layers, leading to two-dimensional (2D) electronic behaviour and often enhanced anisotropy in superconducting parameters (Romano et al., 2024). The general structure of these materials—encompassing cuprates, iron-based superconductors, and the more recently discovered BiS_2 family—consists of repeated units of active superconducting planes interleaved by spacer layers that can be chemically tuned to modify electronic properties.

The crystal structure plays a pivotal role in determining pairing interactions and the resulting critical temperature (T_c). For instance, cuprate superconductors such as $Bi_2Sr_2CaCu_2O_{8+\delta}$ exhibit copper-oxide planes where superconductivity emerges at relatively high temperatures (above 80 K) due to strong electron correlations. Similarly, iron-based superconductors with layered $FeAs$ or $FeSe$ structures display strong anisotropy and significant tunability in T_c under chemical substitution or pressure (Romano et al., 2024). In both instances, the layered topology facilitates enhanced electronic complexity and unconventional pairing mechanisms.

The BiS_2 superconductors represent a relatively recent addition to these families, first identified in 2012 with the discovery of superconductivity in $Bi_4O_4S_3$. This compound exhibits a distinctive layered crystal structure composed of superconducting BiS_2 layers separated by insulating $Bi_4O_4(SO_4)_{1-x}$ spacer layers, with a transition temperature of approximately 8.6 K at ambient pressure (Romano et al., 2024; Wakimoto et al., 2012). While these tetragonal structures share a layered motif with Fe-based and cuprate superconductors, their active electronic layers differ, resulting in distinct superconducting behaviours (Romano et

al., 2024). Although T_c values in the BiS_2 family typically lie in the few-Kelvin range at ambient pressure—considerably lower than high-T_c cuprates—their structural flexibility and sensitivity to external perturbations render them a compelling subject for research.

Subsequent materials in the family, such as $LnO_{1-x}F_xBiS_2$ (where Ln = La, Ce, Pr, Nd), adopt structures consisting of alternating BiS_2 conduction layers and $Ln(O,F)$ blocking layers. In these compounds, superconductivity is induced by electron doping of the BiS_2 layers. Figure 2.1 provides a schematic visualisation of this mechanism, illustrating how the upper insulating layer acts as a charge reservoir where dopant substitution (e.g., Oxygen for Fluorine) occurs. This process injects electrons into the adjacent BiS_2 conduction plane, driving the transition from a semiconducting to a metallic, superconducting state (Yazici et al., 2015).

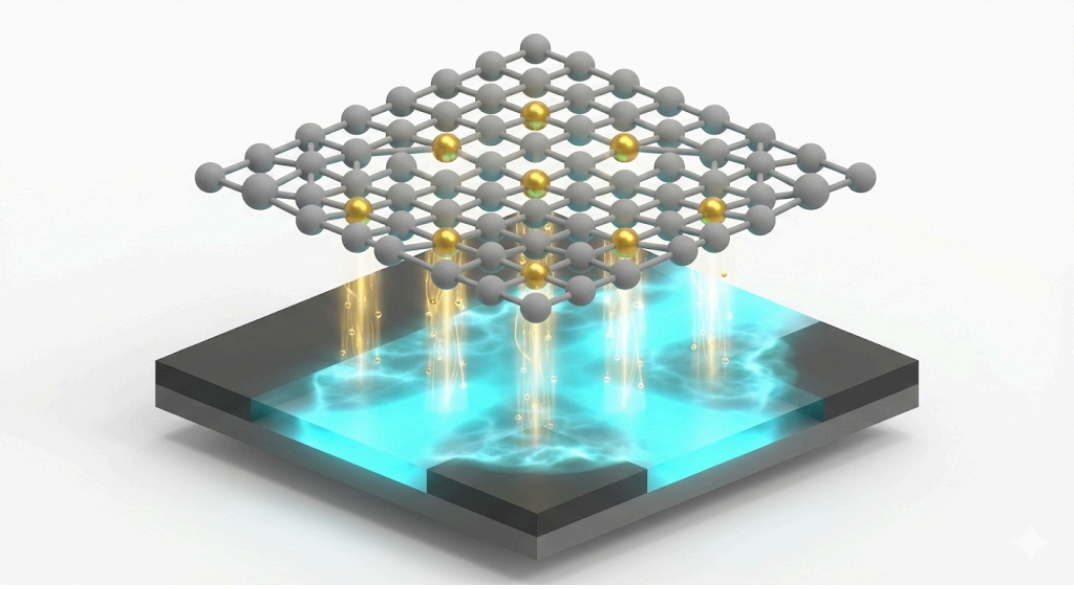


Figure 2.1: Visually enhanced representation of the carrier doping mechanism in a BiS_2-based superconductor. The upper insulating layer acts as a charge reservoir where dopant substitution occurs (indicated by gold spheres). This process injects electrons into the adjacent BiS_2 conduction plane below, driving the transition from a semiconducting to a superconducting state (represented by the cyan glow).

The actual T_c is sensitive to both chemical composition and the application of pressure; hydrostatic pressure and high-pressure synthesis often increase T_c via structural modifications that enhance the electronic bandwidth and carrier density (Romano et al., 2024; Wolowiec et al., 2013). The crystal chemistry of BiS_2-based superconductors further reveals that the emergence of superconductivity is intimately tied to the local geometry of the $Bi-Ch$ (Ch = S, Se) layers and the in-plane chemical pressure exerted by the surrounding lattice (In-plane chemical pressure study, 2015). Structural parameters, such as bond lengths and angles, influence the electron band structure and phonon coupling, which in turn dictate the pairing mechanism.

One of the most intriguing aspects of this family is the interplay between structural instability, lattice degrees of freedom, and superconductivity. For example, structural studies indicate that oxygen vacancy or lanthanide substitution can induce superconductivity by suppressing semiconducting behaviour and enabling metallic conduction within the BiS_2 layers (Chen et al., 2019). This suggests that structural tuning offers a viable route to optimising superconducting behaviour. In summary, the BiS_2 family demonstrates that crystal structure and electronic topology are inextricably linked to superconducting properties. While they may not rival high-T_c superconductors in absolute magnitude, the tunability of their superconducting states and their structural similarities to widely studied layered systems make them a vital focus for theoretical and experimental investigation.

2.2 Natural Language Processing in Materials Science

The application of Natural Language Processing (NLP) within the domain of materials science has undergone a radical transformation, evolving from rudimentary statistical methods to sophisticated, semantics-aware architectures. This trajectory is not merely a technical upgrade but represents a fundamental **Paradigm Shift** in how scientific knowledge is codified, retrieved, and synthesized. As the volume of unstructured text data grows exponentially, the field has moved away from rigid, rule-based pipelines toward models where natural language itself functions as a computational interface for physical properties.

2.2.1 Historical Evolution: From Statistical Counts to Vector Spaces

The early era of materials NLP was characterised by non-neural approaches that relied heavily on statistical frequency and surface-level syntax. Foundational methodologies, such as the n-gram models utilised by Brown et al. (1992) and the structural learning frameworks proposed by Ando and Zhang (2005), focused primarily on discrete token matching. While effective for simple keyword retrieval, these systems lacked the capacity to model semantic relationships or resolve the ambiguity inherent in scientific prose.

The subsequent "Embedding Era" marked a significant leap in capability with the advent of static vectorisation techniques like Word2Vec and GloVe. These models allowed for the projection of words into continuous vector spaces, enabling algebraic operations on concepts—for example, capturing the latent relationship between "doping" and "conductivity." However, as noted by Gupta et al. (2022), these static embeddings suffered from a critical limitation: the lack of domain-specific context. A term such as "phase" could refer to a thermodynamic state, a wave property, or a procedural step; static embeddings struggled to disambiguate these meanings without extensive, manually curated feature engineering.

Furthermore, the rapid pace of advancement in Artificial Intelligence has created what Gupta et al. (2022) identify as a "Velocity Gap." This phenomenon describes the increasing disparity between AI development cycles—which now occur on a weekly basis—and the traditional academic publication cycle, which operates on a yearly timeline. Methodologies such as BERT (Bidirectional Encoder Representations from Transformers), once considered the gold standard for context-aware embedding, are often rendered obsolete by newer architectures before their comprehensive reviews are published. This necessitates a research methodology that is agile and decoupled from rigid, model-specific architectures.

2.2.2 The Paradigm Shift: LLM-Based NLP

Recent scholarship indicates a decisive move away from "Traditional NLP"—characterised by fragmented pipelines requiring separate, complex models for Named Entity Recognition (NER) and Relation Extraction (RE)—toward LLM-based NLP. In this modern framework, described by Jiang et al. (2025), a single Large Language Model (LLM) handles multiple extraction tasks

simultaneously through prompting. This shift eliminates the cascading errors typical of multi-stage pipelines, where a failure in entity boundary detection would inevitably corrupt downstream relation classification.

The theoretical anchor for this transition is the concept of Natural Language as a "Universal Descriptor." Xie et al. (2025) argue that unlike specialised, rigid descriptors such as Crystallographic Information Files (CIF), natural language possesses the unique ability to capture the "messy" realities of experimental physics. Standardised formats often fail to account for synthesis nuances, defect structures, or unexpected phase behaviours that are crucial for reproducibility. Natural language, conversely, provides a high-dimensional descriptive space capable of encoding these complexities. This capability justifies the adoption of a Knowledge Graph approach populated by LLM extraction, as it allows for the structured retention of subtle, qualitative details that purely numerical simulations or rigid databases inevitably miss.

2.3 Large Language Models & Prompting Strategy

The implementation of Natural Language Processing (NLP) in this dissertation is grounded in the strategic utilisation of Large Language Models (LLMs) via Prompt Architecture. Contrary to earlier assumptions that domain adaptation required expensive computational retraining, current State-of-the-Art (SOTA) extraction methodologies demonstrate that inference-time optimisation is often superior to parameter updates.

2.3.1 Strategic Justification: The Case Against Fine-Tuning

A critical decision in the design of the extraction pipeline is the choice between fine-tuning a model on a domain-specific corpus or utilising general-purpose models with advanced prompting. The work of da Silva et al. (2024) provides empirical justification for the latter approach. Their research demonstrates that generic, open-source LLMs (such as Llama-3 or Mistral), when paired with well-designed In-Context Learning (ICL) protocols, achieve extraction performance comparable to specialised workflows. This effectively removes the need for resource-intensive fine-tuning.

This finding has profound implications for resource efficiency. Fine-tuning requires significant GPU memory and curated training datasets, which are often unavailable for niche fields like BiS_2 superconductivity. By validating the efficacy of off-the-shelf models, da Silva et al. establish that high-fidelity extraction (achieving F1 scores of approximately 0.98) is attainable without the prohibitive costs of model retraining. This "Prompt-First" strategy shifts the engineering burden from model training to prompt design, allowing for rapid iteration and adaptability.

2.3.2 Prompt Engineering Pillars

To maximise the reasoning capabilities of these models, the methodology incorporates four core pillars of prompt engineering, as categorised by Lei et al. (2024) and illustrated in Figure 2.2.

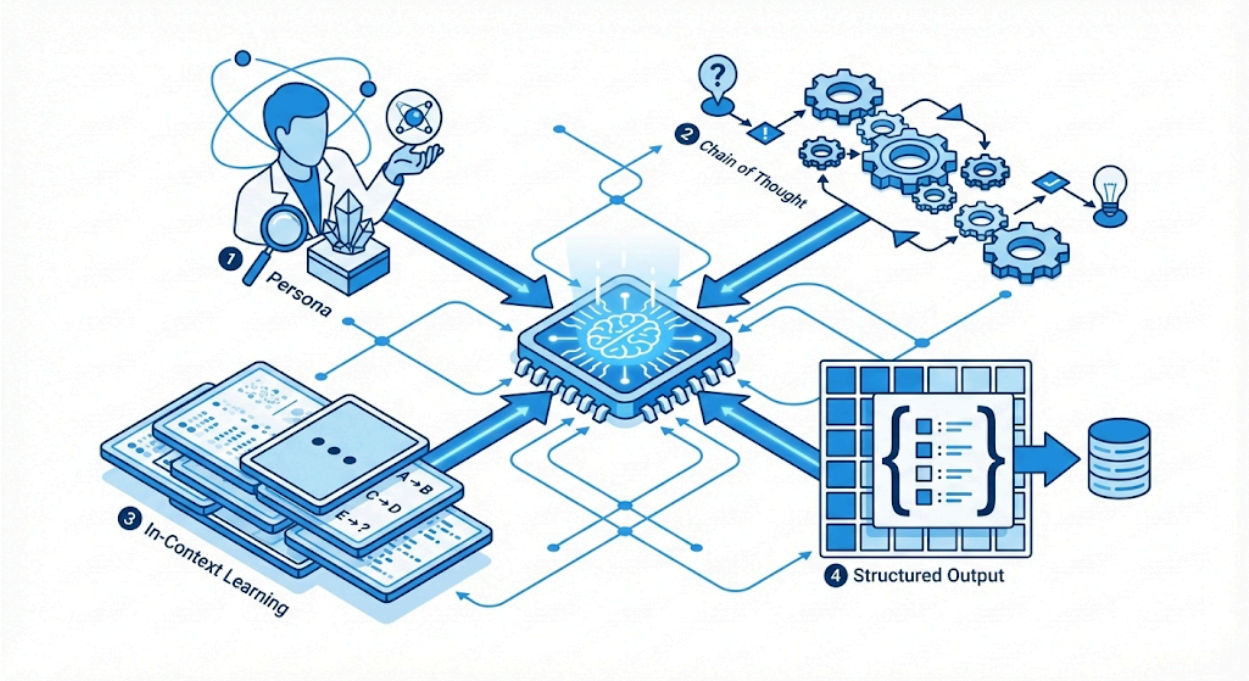


Figure 2.2: Representation of the 4 pillars converging into the LLM reasoning capability

First, the Persona Pattern involves explicitly assigning a role to the model—in this case, that of a "Material Scientist." This primes the model's latent space to prioritise domain-specific terminology and logic, aligning its linguistic expectations with the academic tone of the target

text. Second, Chain-of-Thought (CoT) prompting is utilised to force the model to decompose the extraction task into intermediate reasoning steps rather than attempting a direct output.

This is particularly vital for resolving complex dependencies, such as linking a T_c value to a specific doping concentration across multiple sentences.Third, In-Context Learning (ICL) serves as a powerful tool in the prompt engineering arsenal. By providing "few-shot" examples—complete instances of input text paired with the desired JSON output—the prompt defines the extraction schema and formatting rules without requiring any weight updates. Finally, Structured Prompting employs syntactic delimitations, such as explicit whitespace, JSON code blocks, and clear separators, to reduce parsing errors and ensure that the output remains machine-readable.

2.3.3 Advanced Tactics for Reliability

To further mitigate the risks of hallucination and improve the granularity of the extracted data, advanced tactics proposed by Xie et al. (2025) are employed. Key among these is the use of Counter-examples. By explicitly teaching the model when not to extract information (i.e., when to return NULL), the system reduces false positives—a common failure mode where models attempt to force irrelevant text into a valid schema.

Additionally, Multi-task Synergy is leveraged to extract related attributes—such as Temperature, Pressure, and Material Composition—in a single inference pass. This not only improves computational efficiency but also preserves the semantic linkage between variables that might be severed if extracted independently. Finally, regarding token choice, Xie et al. advise the use of Chemical Formulas (e.g., $LaO_{0.5}F_{0.5}BiS_2$) over string notations like SMILES. LLMs, having been trained on a vast corpora of scientific literature, demonstrate superior reasoning capabilities on human-readable compositional representations, allowing for a more accurate mapping of stoichiometric relationships.

2.4 Knowledge Graphs: Ontologies and Topology Analysis in Scientific Discovery

The transition from isolated entity extraction (NER) to a holistic understanding of scientific literature requires a structured framework capable of preserving semantic relationships. Knowledge Graphs (KGs) have emerged as the standard solution for this challenge, offering a graph-structured data model where entities (nodes) are interlinked by specific relationships (edges). In the context of materials informatics, KGs provide a "semantic layer" that transforms unstructured text into a machine-readable network, enabling the discovery of complex material-property-synthesis correlations that remain invisible to standard keyword-based search engines.

2.4.1 From Text to Graph: The Ontological Imperative

The construction of a valid KG relies fundamentally on its ontology—the formal schema defining the types of entities and permissible relationships. As noted by Olivetti et al. (2020), the primary challenge in data-driven materials research is not merely identifying terms, but mapping the complex interdependence between synthesis parameters (e.g., annealing temperature), structural features (e.g., lattice constants), and functional properties (e.g., T_c).

A well-defined ontology acts as the grammatical rule set for the graph. For instance, linking a superconductor like BiS_2 to a generic concept like "Temperature" is scientifically ambiguous; the ontology must distinguish between Synthesis Temperature and Critical Temperature (T_c). Olivetti et al. (2020) highlight that information extraction pipelines must move beyond simple co-occurrence to capture these specific semantic roles, ensuring that the resulting database supports high-fidelity inverse design.

2.4.2 State-of-the-Art: MatKG and Autonomous Generation

The field has recently advanced from manually curated databases to large-scale, autonomously generated graphs. A seminal precursor to the methodology proposed in this dissertation is MatKG (Materials Knowledge Graph), developed by Venugopal and Olivetti (2024). MatKG represents a massive, autonomously constructed graph containing over 28 million triples extracted from millions of scientific abstracts.

Venugopal and Olivetti's work demonstrates the scalability of automated pipelines, proving that unsupervised or semi-supervised methods can successfully cluster materials, characterise functional relationships, and even predict identifying descriptors for unknown materials. However, while MatKG provides a macroscopic view of the entire materials domain, it acknowledges the trade-off between breadth and depth. Niche material families—such as the BiS_2-based superconductors—often require higher-resolution ontologies to capture subtle doping-dependent behaviours that broad-spectrum graphs might obscure. While this work aims for a similar scope in terms of methodology, the goal is not to build a static encyclopaedia. Rather, it seeks to construct an interactive, dynamic database that serves as a comprehensive repository and analytical engine for this specific field of study.

2.4.3 The LLM Revolution in Graph Construction

While MatKG relied on established NLP techniques (such as LSTM and BERT-based models), the integration of Large Language Models (LLMs) represents the current frontier in KG construction. Zheng et al. (2023) validated this paradigm shift through their development of a "ChatGPT Chemistry Assistant," demonstrating that LLMs could effectively perform text mining to predict Metal-Organic Framework (MOF) synthesis conditions with high accuracy.

Crucially, Zheng et al. (2023) showed that LLMs possess an inherent reasoning capability that allows them to navigate complex syntactic structures far more effectively than rigid rule-based systems. By utilising an LLM to populate the KG, the system can parse intricate experimental procedures and infer implicit relationships—for example, deducing a "successful synthesis" claim from the context of a measured property. This aligns with the "Natural Language as Universal Descriptor" theory discussed in Section 2.2, positioning the LLM not merely as a parser, but as a reasoning agent that structures knowledge directly into the graph topology.

2.4.4 Topological Analysis and Hypothesis Generation

Once constructed, the KG becomes a substrate for topological analysis, where the structure of the graph itself reveals hidden scientific insights. By applying centrality measures, one can identify "hub" nodes—such as highly connected synthesis methods or precursor materials—to reveal

standard practices and dominant paradigms within the BiS_2 field. Furthermore, community detection algorithms, such as Louvain clustering, can be employed to partition the graph into thematic sub-regions, automatically segregating theoretical Density Functional Theory (DFT) studies from experimental synthesis papers. Perhaps the most powerful application, however, is link prediction for identifying "missing links." If a subgraph pattern connects Material A to Property X, and Material B is topologically similar to A, the graph suggests a high probability that Material B also exhibits Property X, thereby guiding researchers toward promising candidates for experimental verification.

3.METHODOLOGY

3.1 Computational Framework and Design Principles

To ensure the transparency and reproducibility of this research, the computational workflow was designed according to the principles of Literate Programming. Rather than treating the software merely as a data processing tool, the codebase was constructed as a "verifiable technical appendix" that documents the granular evolution of the analytical model.

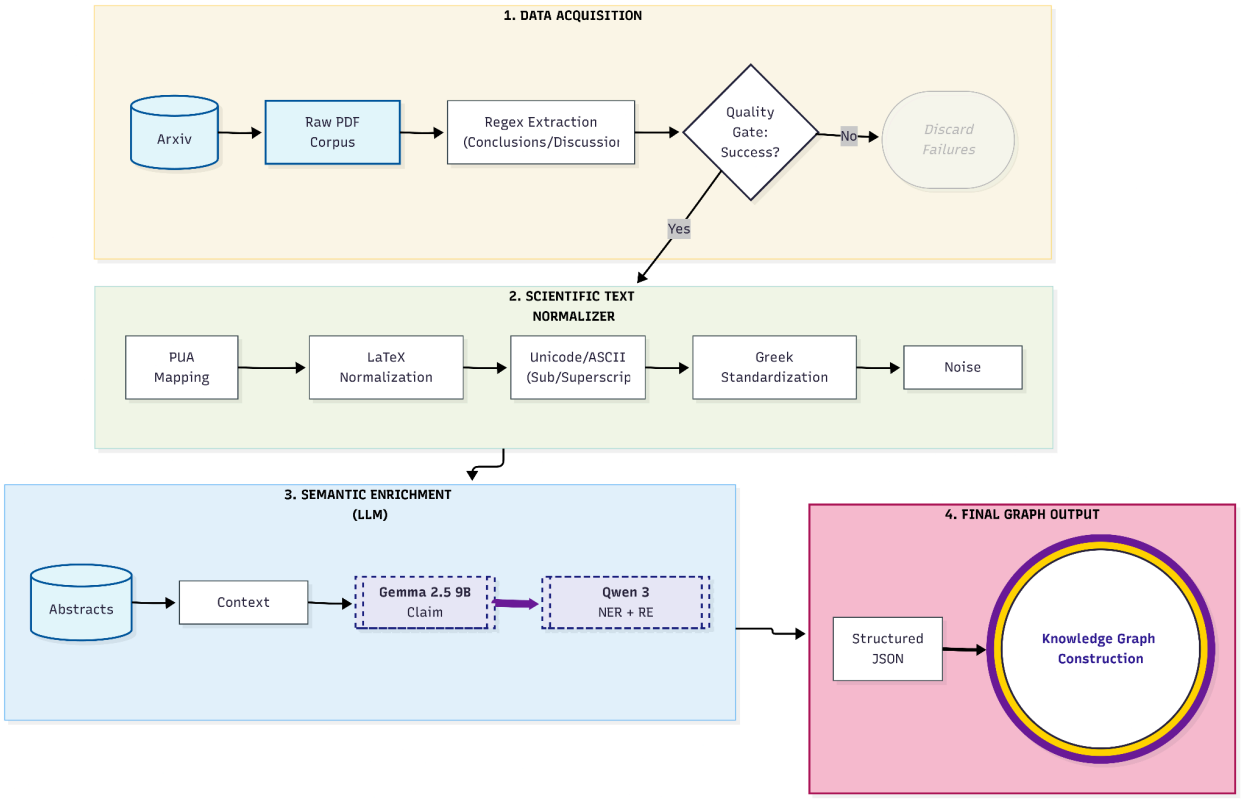


Figure 3.1 Project General Pipeline.

The implementation is encapsulated in a series of sequential Jupyter Notebooks (Notebooks 1–5.3), which are detailed in Table 3.2. These artifacts systematically interleave executable Python code with narrative context, providing immediate justification for algorithmic

choices, data transformations, and parameter selection. The notebooks adhere to PEP 8 standards for code legibility and utilise hierarchical Markdown headers that mirror the structure of this dissertation, ensuring a logical flow from Data Acquisition to Knowledge Synthesis. The general pipeline is illustrated in **Figure 3.1**, while **Table 3.2** provides a comprehensive mapping of the specific methodological functions, underlying algorithms, and their roles within the wider pipeline.

Table 3.2: Computational Workflow and Notebook Organization

This table provides a comprehensive overview of this workflow, mapping each notebook to its specific methodological function, underlying algorithms, and role within the wider pipeline

Phase	Notebook	Primary Function & Methodology
Phase I: Data Acquisition	Notebook 1	**Corpus Acquisition & Curation** Implementation of the iterative query architecture (v1 $\rightarrow$ v3) to retrieve metadata from the arXiv API. Applies Boolean exclusions and quality filters to generate the validated metadata corpus.
	Notebook 2	**Binary Retrieval & Extraction** Automated downloading of full-text PDFs and conversion to raw text using **PyMuPDF**. Handles layout parsing to preserve reading order in multi-column scientific papers.
Phase II: Text Engineering	Notebook 3	**Section Segmentation** Application of the "Strict Mode" regex pipeline (v3.1) to isolate *Conclusion* and *Summary* sections, removing noise (references, headers) to ensure high-density input for NLP.
	Notebook 4	**Scientific Normalization** Execution of the ScientificTextNormalizer. Standardizes chemical formulas (e.g., flattening unicode subscripts), resolves PUA artifacts, and fixes LaTeX formatting to ensure entity consistency.

Phase III: Semantic Extraction	Notebook 5.1	**Epistemological Claim Extraction** Inference using **Gemma 2 9B**. Extracts atomic scientific assertions and classifies them by epistemic type (*Observation* vs. *Speculation*) while resolving pronominal ambiguities.
	Notebook 5.2	**Named Entity Recognition (NER)** Inference using **Qwen 3 8B**. Identifies and categorizes specific graph nodes (*Material, Property, Condition, etc.*) based on a strictly defined ontology.
Phase IV: Knowledge Synthesis	Notebook 5.3	**Relation Extraction & Graph Assembly** Extracts semantic triplets (*Subject $\rightarrow$ Predicate $\rightarrow$ Object*) using a "Strict Legality" prompt. Constructs the final **NetworkX** graph, performs Entity Fuzzy Matching, and executes topological analysis (Louvain Clustering).

3.1.1 Project Folder Structure and Data Persistence

A critical component of the system architecture is the management of data provenance. The workflow is organised within a hierarchical directory structure hosted on Google Drive, ensuring persistent storage and accessibility across diverse computational environments (e.g., Google Colab, Kaggle, and local machines). This structure enforces a strict Separation of Concerns, isolating raw inputs from intermediate processed states and final outputs. The directory tree is organised as follows:

```
TFM/                          <- Root directory
├── notebooks/                <- Jupyter notebooks for experimentation
├── code/                     <- Supporting scripts and utility functions
├── data/
│   ├── raw/                  <- Original, unprocessed datasets (PDF binaries)
│   ├── processed/            <- Intermediate processed datasets
│   ├── output/               <- Final outputs from analysis or models
│   └── corpora/              <- Corpus snapshots at different processing stages
│       ├── 01_raw/           <- Raw corpus (original metadata + JSON snapshots)
│       ├── 02_extracted/     <- Extracted textual content from raw files
│       ├── 03_normalized/    <- Normalized text (cleaned, standardized)
│       └── 04_enriched/      <- Enriched corpus (ontology-aligned with Claims/Entities)
└── results/                  <- Figures, tables, and other outputs
```

The data/corpora/ directory serves as the backbone of the pipeline. Each subdirectory (from 01_raw $\rightarrow$ 04_enriched) corresponds to a distinct processing stage. This "snapshot" approach facilitates the inspection of data at any point in the transformation lifecycle, ensuring that errors in later stages—such as Entity Recognition—can be debugged without the necessity of re-running costly retrieval or extraction tasks.

This organised foundation supports the complex requirements of the project by facilitating the seamless integration of large binary files (PDFs), lightweight JSON metadata, and the iterative versioning of the Knowledge Graph. By standardising relative paths and mounting the drive programmatically, the structure ensures that the inputs for Notebook N are invariably the outputs of Notebook N-1. This creates a deterministic execution chain, allowing the entire experimental setup to be reproduced reliably by other researchers.

3.2 Phase I: Data Acquisition (Corpus Construction): Notebooks 1 & 2

The initial phase of this research focused on the construction of a domain-specific corpus. Given the niche nature of BiS_2-based superconductivity, generic web scraping would likely yield high noise ratios. Therefore, a specialised pipeline was engineered to interact with the arXiv API, transitioning from broad metadata harvesting to targeted binary acquisition and text extraction. This process is encapsulated in the first two computational notebooks of the project's codebase.

3.2.1 Query Engineering: Iterative Refinement and Noise Reduction

The foundation of the corpus rests on a reproducible query strategy designed to maximise recall while suppressing false positives from adjacent physics domains. The implementation utilises the arxiv Python client to interface with the repository's metadata endpoints. The query design followed an evolutionary trajectory, moving from an initial strategy (*focused_queries_1*) that employed broad descriptive phrases targeting the general BiS_2 and $BiCh_2$ family. While effective for recall, this early approach risked retrieving tangential literature.

To enhance precision, the logic was refined into an optimised, multi-tiered architecture (*focused_queries_1_3*).

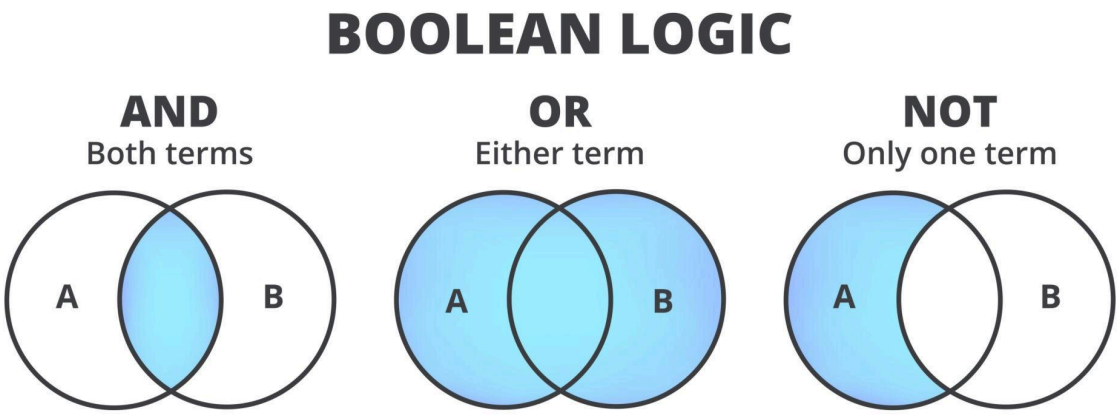


Figure 2.3: The following diagram shows how Boolean logic and its basic operators can be represented visually using Venn diagrams. Boolean logic is a system of logic in which the values of variables are either true or false.

This strategy incorporated specific chemical formulas and structural keywords, executing multiple specific queries in parallel. A custom orchestration function, execute_multiple_searches, managed the retrieval process by aggregating results from these disjoint queries and enforcing strict deduplication based on the unique arxiv_id. This ensured that the final dataset represented a set of unique research artifacts rather than a collection of redundant search hits. This iterative refinement was critical to isolating the specific research domain of interest, distinguishing it from broader condensed matter physics literature.

3.2.2 Filtering Heuristics: Addressing the "Acronym Trap" and "Legacy Family Trap"

Raw search results inevitably contain semantic ambiguities. To ensure the conceptual coherence of the corpus, a robust exclusion layer was implemented to address two specific domain challenges identified during exploratory analysis (p.38).

The first challenge, termed the "Acronym Trap," arises from the overloading of acronyms in condensed matter physics. For instance, the acronym "BIS" can denote "Band Inversion Surface" in the context of topological insulators, rather than Bismuth Sulphide systems. Consequently, papers containing terms such as "band inversion surface" or topological insulators like Bi_2Se_3 were explicitly blacklisted to prevent semantic drift. The second challenge, the "Legacy Family Trap," accounts for the fact that Bismuth-based superconductivity predates the 2012 discovery of the BiS_2 layered family. To avoid contaminating the corpus with chemically distinct legacy materials, specific exclusion criteria were applied to "bismuthate" compounds such as $BaBiO_3$.

Following the application of these filters, the corpus underwent a comprehensive validation phase. This included column verification and statistical analysis of abstract lengths to confirm that the filtered dataset maintained high data quality and completeness before storage.

3.2.3 Ingestion: Automated Retrieval and Layout-Aware Extraction

Once the metadata corpus was validated, the workflow transitioned to the acquisition of primary source data. This stage, handled by the second computational notebook, focused on transforming metadata references into machine-processable text.

The ingestion pipeline was architected with several key technical specifications. First, it employed idempotent retrieval, orchestrated by the build_local_corpus function to manage the batch download of full-text PDFs. To ensure robustness and efficiency, the system normalised arXiv URLs to direct binary links and implemented checks to prevent redundant downloads of existing files. Data integrity was strictly enforced via MD5 checksum verification for every downloaded binary. Additionally, a JSON-based corpus manifest was generated alongside the physical files, logging metadata, file sizes, and cryptographic hashes to ensure full traceability of the raw data.

Additionally, the pipeline addressed the challenge of layout-aware extraction. Standard text extraction often fails on scientific multi-column layouts. To mitigate this, the pipeline utilised PyMuPDF (fitz), a high-fidelity PDF processing library. The extract_text_from_pdf function was deployed to parse the visual geometry of the documents, extracting content into

plain text streams while calculating metadata such as character and word counts. The final output of this phase was a dual-layer dataset: a repository of raw PDF binaries for archival purposes, and a parallel directory of processed .txt files ready for the Natural Language Processing tasks described in subsequent chapters.

3.3 Phase II: Text Engineering and Normalization (Notebooks 3 & 4)

Subsequently to the acquisition of raw textual data, the research pipeline advanced to Phase II: Text Engineering. The raw output from PDF extraction tools often contains structural noise—ranging from hyphenation artifacts to non-semantic headers—that can severely compromise the performance of downstream Natural Language Processing (NLP) models. Consequently, a dedicated transformation layer was implemented to convert these "noisy" dumps into a clean, targeted, and chemically canonicalized dataset suitable for Knowledge Graph construction.

3.3.1 Section Segmentation and Targeted Extraction

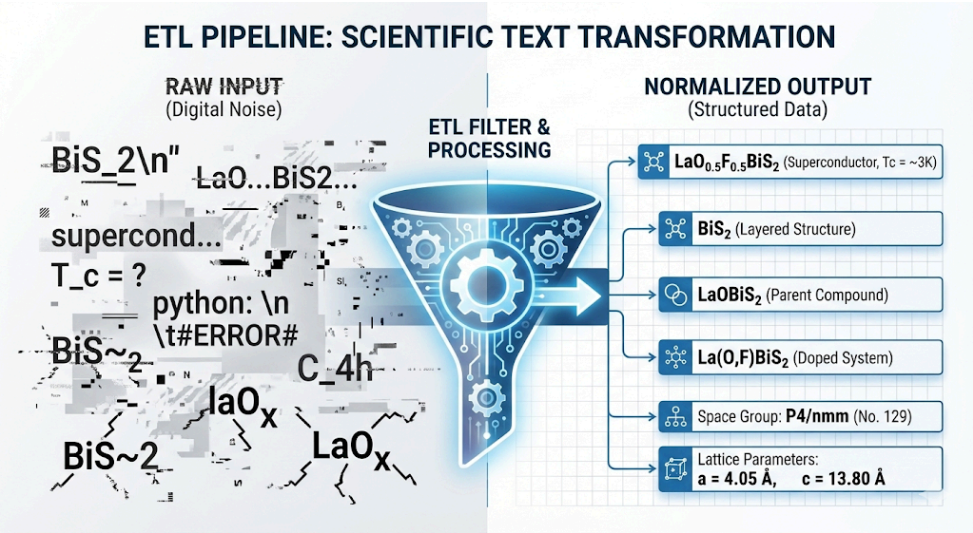


Figure 3.3 This image represents the normalization process of the extracted raw text

The primary objective of Notebook 03 was to isolate the "Conclusion" or “Discussion” sections from the full-text corpus. These sections typically contain the highest density of high-confidence scientific claims, free from the methodological minutiae found in experimental sections. The normalisation process applied to this extracted text is illustrated in Figure 3.3.

Before extraction, a preprocessing function, clean_text_v2, was applied to the raw streams to address specific PDF artefacts that hinder semantic analysis. This function handles ligature resolution by mapping artefacts (e.g., fi $\rightarrow$ fi) back to standard ASCII. Crucially, it performed de-hyphenation, rejoining words split across line breaks (e.g., super- + conductivity $\rightarrow$ superconductivity) while intelligently preserving essential chemical formulas (e.g., Bi-S). Finally, whitespace normalisation compressed vertical and horizontal spacing to produce continuous, machine-readable prose.

The extraction mechanism evolved through a strictly versioned iterative process, moving from rigid keyword matching to semantic boundary detection. The baseline model (V1) utilised simple regex patterns targeting explicit "Conclusion" headers but suffered from low recall (~40%) due to its inability to handle diverse section titles. The subsequent iteration (V2) expanded these patterns to include variants like "Concluding Remarks" and "Summary," significantly improving recall to approximately 80%.The final production logic (V3.1, "Strict Mode") focused on maximising precision and preventing "content bleed," such as the accidental inclusion of references.

This version implemented strict case-sensitivity for headers to avoid false positives in the body text and utilised implicit stop-patterns—identifying termination signals such as "We thank" or "This work is supported by"—to strictly demarcate the end of the scientific narrative. The performance of these iterations was benchmarked against a manually curated "Gold Standard" of 10 papers. The final V3.1 logic achieved a significant reduction in boundary errors, decreasing instances of missing stop patterns by 90% compared to previous versions.

3.3.2 Scientific Text Normalization

While extraction isolated the correct text spans, the raw strings remained chemically inconsistent. Notebook 04 introduced the ScientificTextNormalizer class, designed to standardise

entities into a canonical format essential for the Entity Resolution phase of the Knowledge Graph.

To ensure that the Knowledge Graph treats variants like BiS_2 and BiS2 as identical nodes, the pipeline applied several key transformations. PUA Character Mapping restored characters from the Unicode Private Use Area (e.g., \uf02d) to their standard equivalents, repairing broken numerical ranges and formulas. LaTeX Parsing stripped formatting commands (e.g., \textit) and converted mathematical macros (e.g., \rho) into standard Unicode symbols (ρ). Finally, Subscript/Superscript Flattening converted Unicode subscripts (e.g., ₂) to ASCII digits. This "flattening" ensures that complex formulas like $LaO_{0.5}F_{0.5}BiS_2$ are consistent across the entire corpus, regardless of the original paper's formatting style.

The result of this phase was the bis2_corpus_v1_normalized.json, a chemically accurate and topologically consistent dataset ready for semantic injection into the Large Language Models.

3.4 Phase III: Semantic Information Extraction (Notebooks 5.1 & 5.2)

Following the normalization of the corpus, the research transitioned from structural text engineering to **Semantic Information Extraction (IE)**. This phase constituted the core intelligence layer of the pipeline, designed to transmute raw scientific prose into structured data suitable for graph composition. Unlike traditional keyword-based approaches, this methodology employed a two-stage Large Language Model (LLM) pipeline to first isolate scientific assertions (Claims) and then identify the constituent variables (Entities) within them.

3.4.1 Stage 1: Epistemological Claim Extraction

The first stage of the extraction process addressed the ambiguity inherent in scientific writing, where direct experimental results are often interwoven with theoretical speculation and citations of previous work. To resolve this, a specialised extraction engine was constructed using Gemma 2 9B-IT, a model selected for its high performance in instruction-following and reasoning tasks.

Unlike standard extraction pipelines that treat all text as uniform fact, this methodology enforced a epistemic classification schema. The model was prompted to categorise every extracted claim into one of three distinct types: Observation, representing direct experimental findings (e.g., "We observed a T_c of 4.5 K"); Inference, capturing logical deductions derived from data (e.g., "This suggests a phonon-mediated mechanism"); and Speculation, denoting theoretical proposals or future outlooks (e.g., "Higher pressure may enhance T_c").

To ensure high-fidelity extraction, the system prompt established a "Materials Physicist" persona. A critical component of this prompt was Mandatory Pronoun Resolution. Since scientific abstracts frequently employ anaphora (e.g., "It exhibits superconductivity"), the model was explicitly instructed to resolve these pronouns, replacing ambiguous terms like "it" or "the sample" with the specific chemical formula found in the preceding context (e.g., $LaO_{0.5}F_{0.5}BiS_2$).

To accommodate the memory constraints of the NVIDIA T4 GPU environment, the model utilised 4-bit quantisation (NF4) via BitsAndBytesConfig. Furthermore, inference was executed using Greedy Decoding (temperature = 0.0). This setting was chosen to enforce determinism and reproducibility, thereby minimising the risk of "hallucinations" common in generative models.

3.4.2 Stage 2: Named Entity Recognition

The second stage of the extraction pipeline focused on granular parsing. Once the epistemological claims were isolated, they served as the direct input for Named Entity Recognition (NER). This process utilised the Qwen 3 (8B) model, selected for its superior performance in structured data formatting, to identify the nodes of the eventual Knowledge Graph.

To ensure the resulting graph remained queryable and standardised, a strict ontology was defined. As illustrated in Figure 3.2, the model was restricted to classifying entities into six exclusive classes connected by nine allowed relations (detailed fully in Appendix E).

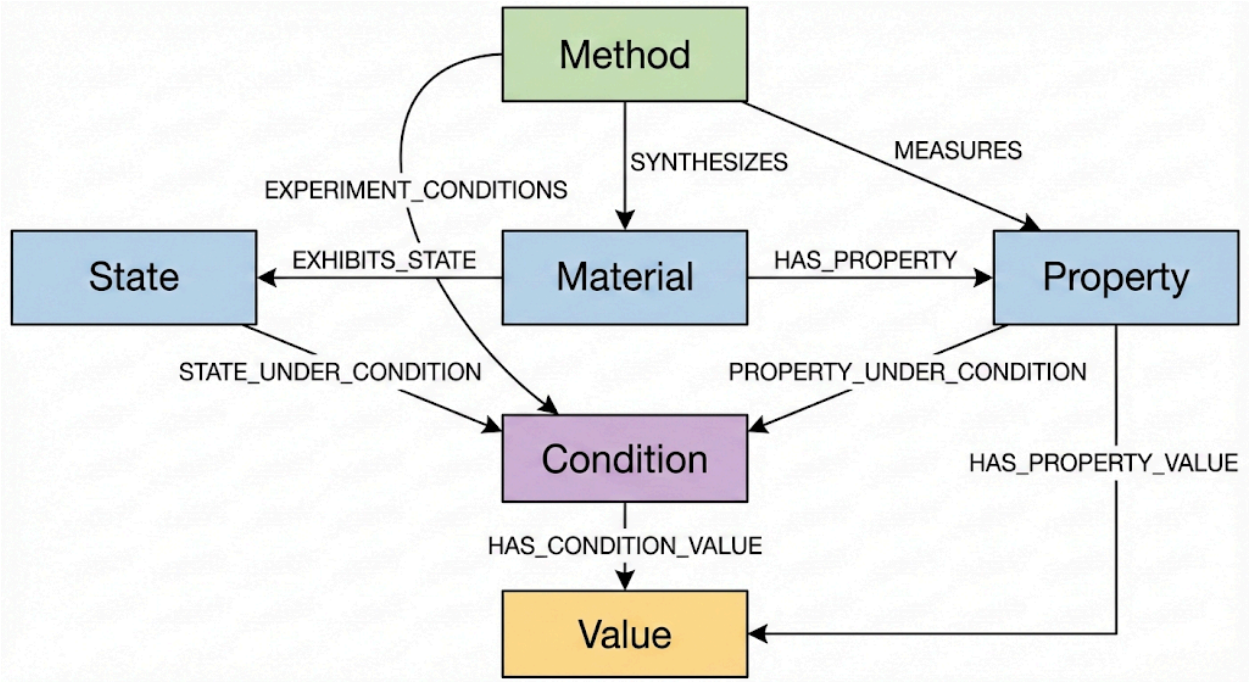


Figure 3.2: Entities and relations Ontology Diagram

The extraction logic employed a Few-Shot Prompting strategy, embedding a "Gold Standard" subset of manually annotated claims directly into the context window to guide the model's reasoning. To ensure operational robustness, the pipeline featured a fault-tolerant inference loop that serialised results to .jsonl files in real-time, protecting the workflow against runtime crashes. The quality of the extraction was validated quantitatively through a performance assessment using Precision, Recall, and F1-Score metrics against the Gold Standard dataset. This testing ensured the model accurately distinguished between subtle categories, such as the nuance between a Condition (e.g., "under pressure") and a Method (e.g., "high-pressure synthesis").

3.5 Phase IV: Knowledge Synthesis and Graph Construction (Notebook 5.3)

The final phase of the methodology focused on synthesizing the isolated semantic components—claims, entities, and extracted relations—into a cohesive topological structure. This process, encapsulated in Notebook 5.3, transformed the linear dataset into a

multidimensional Knowledge Graph (KG), enabling the discovery of latent patterns through network theory and interactive visualization.

3.5.1 Entity Resolution and Canonicalization

A critical challenge in automated graph construction is the linguistic variance of scientific terms. Different authors frequently refer to identical concepts using divergent nomenclature—for example, denoting the transition temperature as "T_c," "critical temperature," or "superconducting transition temperature." Without resolution, these variations would result in fragmented, disconnected nodes, severely degrading the graph's analytical utility.

To address this, an Entity Fuzzy Matching layer was implemented using the EntityMatcher class. This module utilised the Levenshtein distance algorithm (via the fuzzywuzzy library) to calculate string similarity scores between entities. Entities exceeding a predefined similarity threshold were automatically mapped to a single Canonical Form. For instance, various lexical permutations of "Critical Temperature" were unified under the single node T_{c0}. This normalisation process ensured topological consistency, allowing distinct papers to contribute to shared nodes and thereby revealing the true connectivity of the research landscape.

3.5.2 Graph Topology and Assembly

Following normalisation, the enriched data—comprising normalised entities, semantic triplets, and their provenance metadata—was merged back into the corpus structure. The Knowledge Graph was then instantiated as a Directed Multigraph (MultiDiGraph) using the NetworkX library.

The graph topology was designed with a strict hierarchical schema to preserve scientific context. At the Provenance Layer, Article nodes are linked to Claim nodes via CONTAINS edges, maintaining the traceability of every data point. The Semantic Layer then links Claim nodes to Entity nodes, which are further interconnected via specific physical relations (e.g., $Material \xrightarrow{\text{HAS\_PROPERTY}} Value$). Crucially, a Material Anchoring constraint was enforced, requiring that all property and condition nodes eventually link back to a

Material node. This logic prevents "floating facts" and ensures every data point remains chemically contextualised.

3.5.3 Network Analytics and Community Detection

Following the assembly of the Knowledge Graph, the topological structure itself was subjected to analytical scrutiny to reveal the latent semantic organisation of the BiS_{2} field. To elucidate the hierarchy of influence within the network, Degree Centrality algorithms were employed to identify "Hub" nodes. This process metricated the epistemic weight of specific entities, quantitatively confirming that materials such as $LaO_{0.5}F_{0.5}BiS_{2}$ and their associated physical properties function as the central pillars of the research domain.

Complementing this centrality analysis, the Louvain algorithm was leveraged to perform unsupervised clustering on the entity subgraph. By optimising for modularity, this algorithmic approach partitioned the graph into densely connected subgraphs, effectively delineating distinct "Scientific Communities." This automated segmentation successfully revealed thematic clusters—ranging from High-Pressure Synthesis to Theoretical Band Structure—independent of manual labelling, thereby validating the graph's ability to preserve the inherent distinctness of research sub-fields.

4. RESULTS AND ANALYSIS

4.1 Data Retrieval Metrics and Corpus Construction

This section details the results of the automated data acquisition phase, focusing on the construction of the primary corpus for BiS_2-based layered superconductors. The objective was to transform a high-volume, noisy data stream from the arXiv API into a high-fidelity, domain-specific dataset suitable for downstream Knowledge Graph construction. The process involved a systematic evolution of query strategies (v1.0 $\rightarrow$ v3.0) and rigorous quality control mechanisms.

4.1.1 Iterative Query Strategy Performance

The development of the search strategy followed an iterative optimisation path, evolving from broad recall-oriented queries to high-precision, exclusion-based architectures.

V1.0 (Initial Baseline)

The initial execution utilised broad queries targeting general terms such as BiS2_general and descriptive keywords. While this successfully retrieved 121 unique papers, the "descriptive" module exhibited 100% duplication, as illustrated in Figure 4.2. This outcome indicated extreme redundancy in broad keyword searches, prompting an immediate refinement of the boolean logic.

Table 4.1: Query Strategy v1.0

The following table outlines the Boolean search designed to capture literature regarding BiS2 and BiCh2-based superconducting materials.

Query Name	Search String	Explanation
BiS2 General	all:BiS2 AND all:superconductor	**Broad Capture:** Retrieves all records containing both "BiS2" and "superconductor" to ensure context relevance.
BiCh2 General	all:BiCh2	**Structure Specific:** Targets the general "BiCh2" (Bismuth Chalcogenide) structural unit without restricting to superconductivity immediately.
REOBiS2 Series	all:LaOBiS2 OR all:CeOBiS2 OR all:NdOBiS2 OR all:PrOBiS2	**Rare Earth Series:** specifically targets the prominent Rare Earth Oxide BiS2-based systems (Lanthanum, Cerium, Neodymium, Praseodymium).
Bi4O4S3	all:Bi4O4S3	**Key Compound:** A targeted search for the specific layered superconductor $Bi_4O_4S_3$.
Sr/Eu Compounds	all:SrFBiS2 OR all:EuFBiS2 OR all:EuBiS2F	**Fluoride Doping:** Focuses on Strontium and Europium fluoride-doped BiS2 compounds.
Descriptive	ti:"BiS2-based" OR abs:"BiS2-based" OR ti:"BiCh2-based" OR abs:"BiCh2-based"	**High Precision:** Limits the search to Titles (ti) and Abstracts (abs) to find papers explicitly defining the material class as "BiS2-based" or "BiCh2-based".
Bismuth Chalcogenide	all:bismuth AND all:chalcogenide AND all:superconductor	**Keyword Fallback:** A broad semantic search for documents mentioning bismuth, chalcogenides, and superconductors that might miss specific formula notations.

Notes: Field Codes: all: (All Fields), ti: (Title), and abs: (Abstract). The logic uses standard AND / OR operators. Ensure these are capitalized in your search engine.

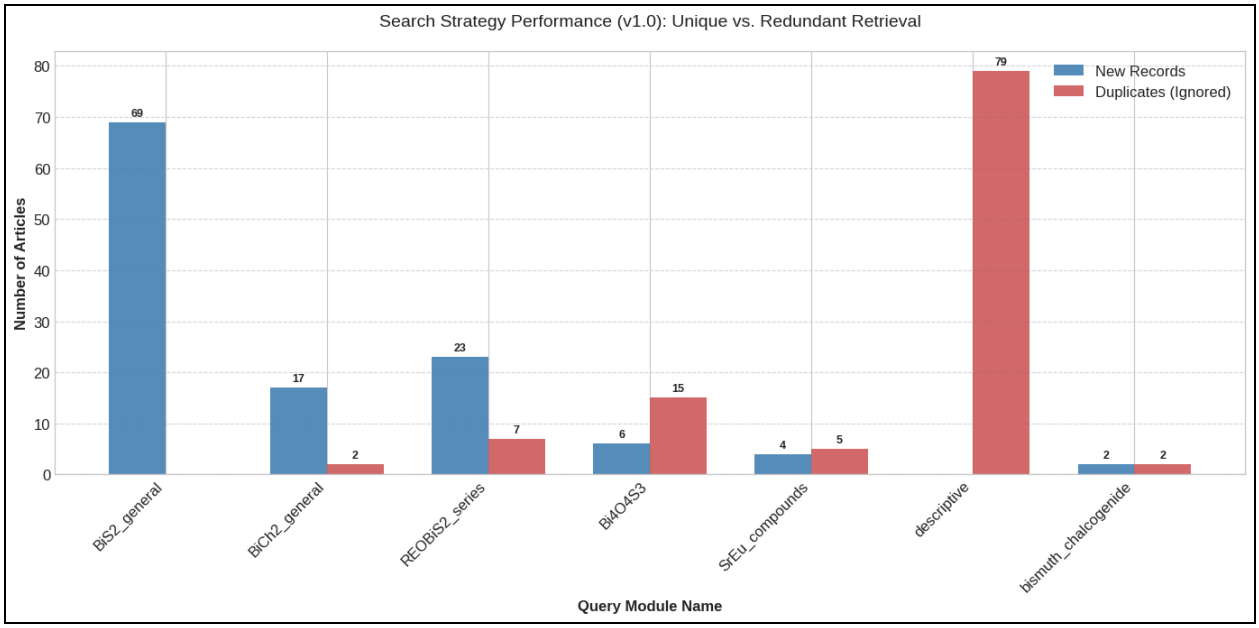


Figure 4.2. Retrieval efficiency per query module with the first version of the search query. Pale red columns describe the total of previously found articles by the query.

V1.2 (Recall Expansion)

To maximize coverage, the strategy was expanded to include rare-earth series, fluorine-doped variants, and structural keywords (layered_bismuth). This increased the unique yield to 176 papers. However, deep analysis revealed a critical "Noise Pollution" issue: the BiSSe_mixed module generated 67.6% noise, and layered_bismuth contributed 12.9% noise. These false positives were primarily driven by topologically insulating materials (e.g., Bi_2Se_3) and legacy bismuth compounds unrelated to the specific BiS_2 superconducting family.

Table 4.3: Query Strategy v1.2: BiS2 and BiCh2 Superconducting Families

The following table details the updated Boolean search strings. This version (v1.2) expands on the Rare Earth Oxide series, introduces Fluoride-doped variants, and adds specific searches for mixed Selenium/Sulfur systems and layered bismuth structures.

Query Name	Search String	Explanation
BiS2 Core	all:BiS2 AND all:superconductor	**Core Search:** Retrieves the fundamental BiS2 superconductor literature.
BiCh2 Family	all:BiCh2 AND all:superconductor	**General Structure:** Targets the broader class of Bismuth Chalcogenide (BiCh2) superconductors.
Ln/RE OBiS2 Series	all:LaOBiS2 OR all:CeOBiS2 OR all:NdOBiS2 OR all:PrOBiS2 OR all:YbOBiS2	**Expanded RE Series:** Covers the primary Lanthanide (Ln) Oxide BiS2 systems, now including Ytterbium (Yb).
F-doped LnOBiS2	all:LaOFBiS2 OR all:REOFBiS2 OR all:LnOFBiS2	**Doping Variants:** Specifically targets Fluorine-doped variations of the Lanthanide Oxide BiS2 layer.
Bi4O4S3 Parent	all:Bi4O4S3	**Prototype Compound:** The specific formula for the first reported BiS2-based superconductor ($Bi_4O_4S_3$).
Sr/Eu Blocking Layers	all:SrFBiS2 OR all:EuFBiS2	**Spacer Layers:** Focuses on Strontium and Europium Fluoride blocking layers.
BiSSe Mixed	all:BiSSe OR all:LaOBiSSe	**Solid Solutions:** Targets mixed anion systems ($S_{1-x}Se_x$) often investigated for T_c enhancement.
Descriptive	all:bismuth AND all:chalcogenide AND all:superconductor	**Semantic Search:** Broad keyword search for the material class without specific formula constraints.
Layered Bismuth	all:layered AND all:bismuth AND all:superconductor	**Structural Search:** Captures papers discussing the structural dimensionality (2D/layered) of bismuth superconductors.

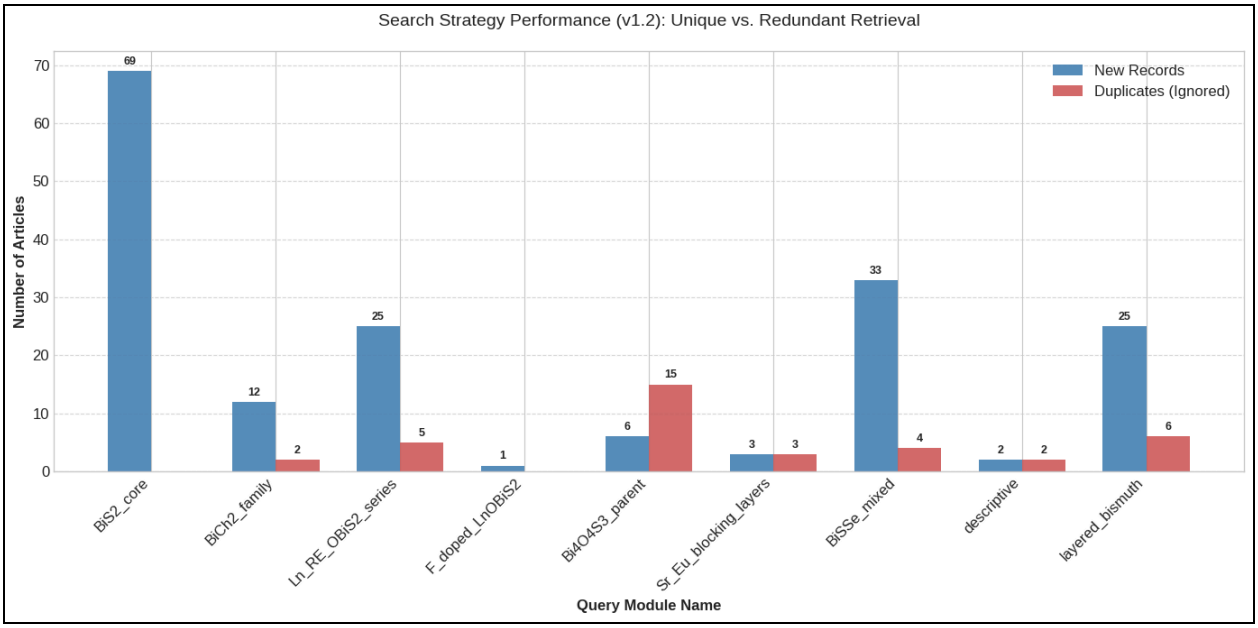


Figure 4.4 Duplication rates in the “broadened” version of the search query.

Table 4.5.: Search Performance Analysis (Noise Audit)

The following data summarizes the precision of Query Strategy v1.2. "Noise" is defined as retrieved records irrelevant to the target topic (BiS2-based superconductivity).

Query Module	Total Hits	Noise Count	Noise Ratio	Status
BiS2_core	69	0	**0.0%**	✅ **Optimal**
BiCh2_family	14	0	**0.0%**	✅ **Optimal**
F_doped_LnOBiS2	1	0	**0.0%**	✅ **Optimal**
Sr_Eu_blocking	6	0	**0.0%**	✅ **Optimal**
descriptive	4	0	**0.0%**	✅ **Optimal**
Ln_RE_OBiS2	30	1	**3.3%**	🟢 **Acceptable**
Bi4O4S3_parent	21	1	**4.8%**	🟢 **Acceptable**
layered_bismuth	31	4	**12.9%**	⚠️ **Caution**
BiSSe_mixed	37	25	**67.6%**	❌**Failure**

Note: The search string all:BiSSe OR all:LaOBiSSe lacked a functional constraint. "BiSSe" (Bismuth Sulfoselenide) is heavily researched in other fields, particularly as a Topological Insulator or thermoelectric material, often unrelated to superconductivity.

v2.0 (Thematic Precision)

The decisive optimization involved appending explicit AND (all:superconductor OR all:superconductivity) clauses to all queries. This thematic enforcement, combined with a "Refined Blacklist":

```
# --- 1. Cumulative Refined Blacklist ---
# Merging original terms with the newly identified "Acronym" and "Bismuthate" traps
FINAL_EXCLUSION_TERMS = EXCLUSION_TERMS + [
    # 1. The "BIS" Acronym Clash (Band Inversion Surfaces)
    "band inversion surface",
    "band inversion surfaces",
    "BISs",

    # 2. The Bismuthate Family (Oxides, distinct from Chalcogenides)
    "bismuthate",
    "BaBiO3",
    "BaPbO3",
    "BKBO",
    "BPBO"
]
```

reduced the noise rate by **92%** compared to v1.2. The exclusion rate dropped to just **1.5%** (2 flagged articles vs. 30 in v1.2), yielding **130 high-confidence unique papers**.

Table 4.6: Search Strategy Optimization: v1.2 vs. v2 Comparison

The following data compares the raw search volume and precision metrics before and after the application of noise-reduction filters (specifically targeting the BiSSe and layered modules).

Version	Total Articles Retrieved	Filtered Articles	Retention Rate	Status
v1.2	170	30	**82.4%**	⚠️ High Noise
v2.0	130	2*	**98.5%**	✅ Optimized

Note: The drop in Total Articles (from 170 to 130) reflects the successful exclusion of irrelevant "BiSSe" topological insulator papers at the source, rather than needing to filter them out manually.

V3.0 (Efficiency Optimization)

The final iteration integrated exclusion clauses (e.g., AND NOT "topological insulator") directly into the API call. This ensured that these “blacklisted” terms are excluded from the final corpus. Moreover, the reduced number of more targeted queries resulted in the average duplication rate per query from 51.3% (v2) to 30.5% (v3)—an efficiency gain of 40.5%. The final yield was 132 unique papers, confirming that efficiency could be improved without sacrificing recall.

Table 4.7: Optimized BiS2 Superconductor Queries v3

This version reduces the number of queries from 9 to 5, reducing API calls, and implements exclusion terms stated above to reduce noise.

Module Name & Expected Yield	Search Query (Boolean Logic)	Strategy Notes
1. BiS2 Primary	(all:BiS2 AND (all:superconductor OR all:superconducting OR all:layered)) AND NOT all:Bi2Se3 AND NOT all:Bi2Te3 AND NOT all:"topological insulator"	**Core Capture:** Targets the base BiS2 layer. **Filter:** Explicitly excludes topological insulators (Bi_2Se_3/Te_3) to prevent false positives.
2. REO Compounds	(all:LaOBiS2 OR all:CeOBiS2 OR all:NdOBiS2 OR all:PrOBiS2 OR all:SmOBiS2 OR all:YbOBiS2 OR all:GdOBiS2) AND NOT all:Bi2Se3 AND NOT all:Bi2Te3 AND NOT all:"topological insulator"	**Series Expansion:** Covers the full Lanthanide series (La, Ce, Nd, Pr, Sm, Yb, Gd). **Filter:** Applies standard T.I. exclusions.
3. Chalcogen Variants	((all:BiCh2 OR all:BiSSe OR all:LaOBiSSe) AND (all:superconductor OR all:superconducting)) AND NOT all:Bi2Se3 AND NOT all:Bi2Te3 AND NOT all:"topological insulator"	**Anion Mixing:** Targets "BiCh2" general notation and Selenium-doped variants ($BiSSe$). **Constraint:** Requires "superconducting" keyword to avoid thermoelectric papers.
4. Parent & Variants	(all:Bi4O4S3 OR all:SrFBiS2 OR all:EuFBiS2 OR all:LaOFBiS2) AND NOT all:Bi2Se3 AND NOT all:Bi2Te3 AND NOT all:"topological insulator"	**Specific Systems:** Targets the prototype $Bi_4O_4S_3$ and Fluoride-doped/Spacer layer variants. **Filter:** Applies standard T.I. exclusions.
5. Descriptor Net	((all:"bismuth oxyselenide" OR (all:"bismuth sulfide" AND all:layered)) AND all:superconducting) AND NOT all:Bi2Se3 AND NOT all:Bi2Te3 AND NOT all:"topological insulator"	**Semantic Safety Net:** Catches papers using descriptive names rather than formulas. **Constraint:** Strictly links "Bismuth Sulfide" with "Layered" to avoid bulk mineral results.

Notes: Every query now appends AND NOT all:Bi2Se3 AND NOT all:Bi2Te3 AND NOT all:"topological insulator". This is the primary driver for the increased precision (from 82% to 98%). Expanded to include all:superconducting (adjective) alongside all:superconductor (noun) to ensure no grammatical variations are missed.

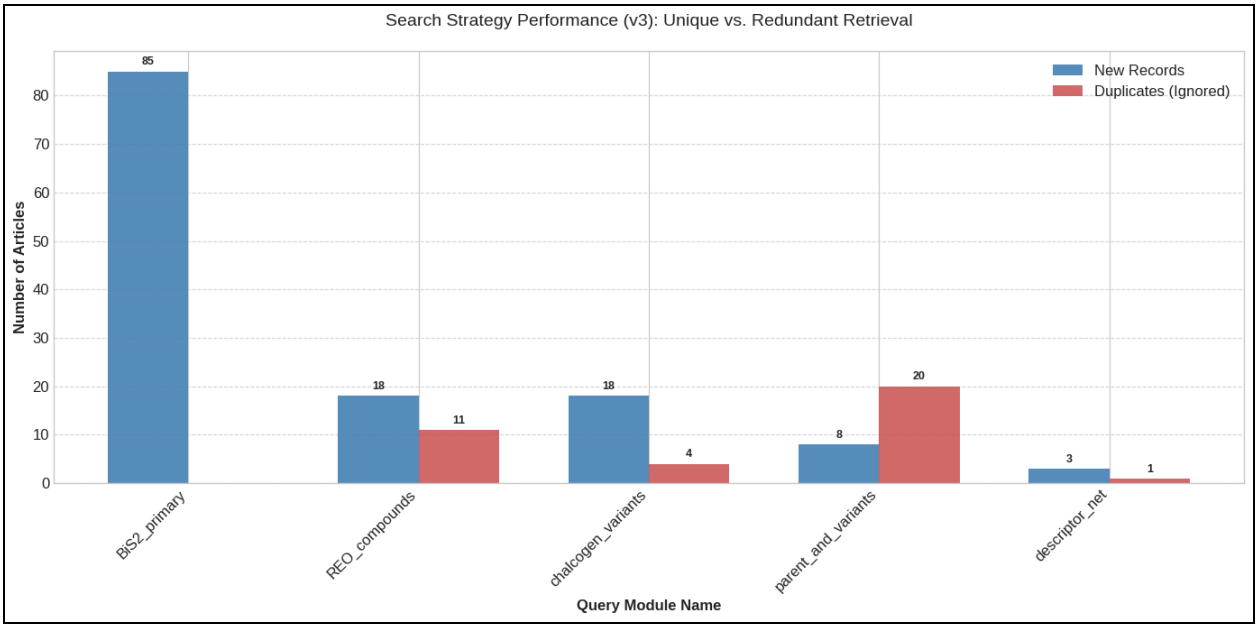


Figure 4.8. Duplication rated in version 3 of the search query.

4.1.2 Quality Control and Filtering Effectiveness

A multi-stage filtering mechanism was essential to ensure the semantic integrity of the corpus. The implementation of a "Refined Blacklist" proved highly effective in mitigating domain-specific linguistic traps that could otherwise compromise the dataset. Specifically, the strategy neutralized the "Acronym Trap," where terms such as "BISs" frequently refer to "Band Inversion Surfaces" in topological physics rather than Bismuth Sulphide systems.

Furthermore, the filter addressed the "Legacy Family Trap" by excluding oxide superconductors like $BaBiO_3$ and $BKBO$. Although these materials contain Bismuth, they belong to the perovskite structural class, which is distinct from and predates the 2012 discovery of the BiS_2 family.

The structural disparity between these legacy cubic oxides and the target layered chalcogenides is illustrated in Figure 4.9. The application of these targeted filters resulted in a high degree of "chemical and phenomenological homogeneity" within the dataset, ensuring that

downstream NLP tasks would remain focused on the relevant material family without contamination from adjacent, yet distinct, areas of physics.

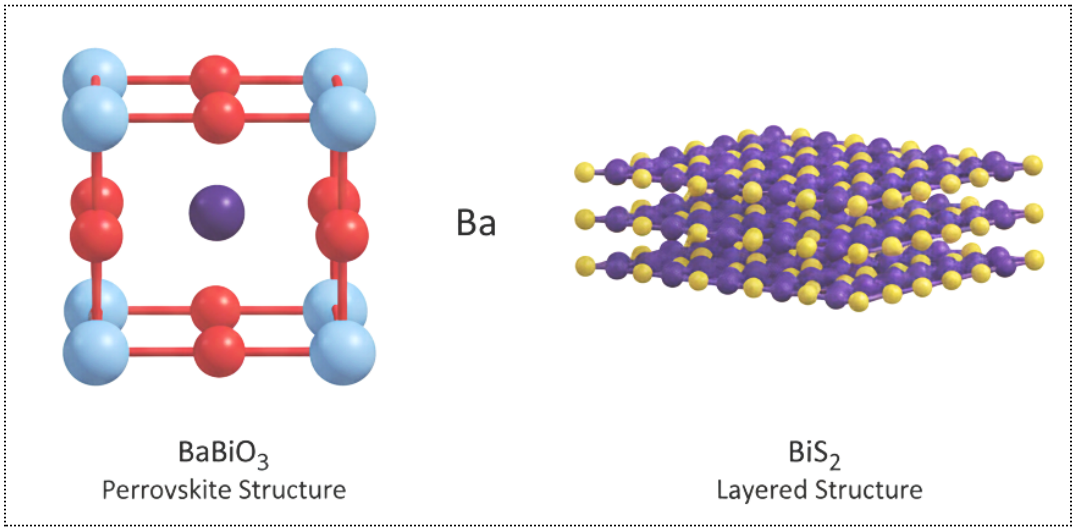


Figure 4.9 Visual structure comparison of both Perovskite and BiS2 Layered compounds

4.1.3 Comparative Analysis (v2 vs. v3) and Final Corpus Selection

A direct comparison between the precision-focused v2 and the efficiency-focused v3 architectures revealed a high degree of convergence, validating the robustness of the retrieval logic. As detailed in the word and character distribution analysis (see Appendix B), the versions were highly comparable, indicating nearly identical outputs. This similarity suggests that both query designs effectively targeted the same core literature; indeed, the datasets shared 127 articles, representing a 97.69% overlap.

The marginal variance between the two versions was minimal: v2 contained three unique papers not identified in v3, while v3 captured five papers missed by its predecessor. Moreover, metadata density remained optimal across both iterations, with 100% completeness for critical

fields such as titles, abstracts, and authors. The mean abstract lengths were also remarkably consistent, recorded at 903.7 characters for v2 and 894.2 characters for v3.

Given the marginal net gain of only two papers in v3 and the proven stability of the v2 dataset during earlier validation phases, Corpus v1 (derived from the v2 query) was selected as the foundational dataset for this dissertation. This final corpus consists of 128 high-confidence unique articles after stringent filtering. This selection prioritises reproducibility and stability, as the v2 corpus provides a verified and steady balance of precision and recall, establishing a reliable substrate for the construction of the Knowledge Graph.

Table 4.10: Corpus v1 Specifications

The following table details the metrics of the final dataset established after running Query Strategy v2.

Attribute	Value
Source Query	Query v2
Total Articles	130
After Filtering	128 (98.5% retention)
Filtered Articles	2
Data Completeness	100% for core fields
Version Label	v1

Table 4.10 The retention rate of 98.5% confirms that the exclusion filters (targeting topological insulators) effectively removed noise without sacrificing relevant data. With 128 high-quality unique records and 100% data completeness for core fields (Title, Abstract, Author, Year), this corpus is now locked as Version 1 (v1) and ready for bibliometric or meta-analysis.

4.2 Text Preprocessing and Section Extraction Validation

Following the acquisition of the raw corpus, the research phase transitioned to Text Engineering. The objective was to isolate the semantic core of each document—specifically the "Conclusion" and "Summary" sections—while filtering out noise such as references, acknowledgments, and typesetting artifacts. This section details the performance of the regex_extraction_v3_1 pipeline and the quantitative impact of the text normalization strategy.

4.2.1 Strategy and Logic Implementation

The extraction pipeline was architected around two interdependent stages designed to transform unstructured PDF text into machine-readable prose, ensuring that high-density scientific information was preserved without structural interference. During the initial Text Normalisation phase, raw text underwent thorough cleaning via the clean_text_v2 function to mitigate common PDF-to-text artefacts. This included ligature resolution—mapping characters such as fi back to standard ASCII fi—and whitespace compression to maintain continuous prose. Crucially, the system performed de-hyphenation to rejoin split words like "superconductivity" while employing logical constraints to preserve chemical notations, such as $Bi-S$, which are essential for maintaining the chemical integrity of the dataset.

Subsequently, the Heuristic Section Extraction logic evolved from a baseline regex model into a sophisticated "Strict Mode" architecture (V3.1). While early iterations (V1 and V2) focused on diverse header matching for sections like "Concluding Remarks," they often suffered from "content bleed" where subsequent sections, such as References, were accidentally ingested. The final V3.1 iteration addressed this by enforcing case-sensitivity and introducing implicit stop-patterns—identifying phrases like "We thank..." or detecting visual separators. This refinement proved critical for processing articles lacking explicit headers for “Acknowledgments”, ensuring that only high-confidence conclusions were serialised for the final Knowledge Graph.

4.2.2 Quantitative Performance Analysis: V1 to V3 Evolution

The extraction pipeline was evaluated against a manually curated Gold Standard of 10 articles and applied to the full corpus of 128 papers. A comparative analysis across three logic versions (V1–V3) utilised F1-Score, Character Similarity, and Stability as primary diagnostic metrics. The results demonstrate a clear evolutionary trend: V3 achieved a superior harmonic mean of precision and recall, maximising semantic fidelity while minimising variance. As illustrated in the box plot analysis (Figure 4.12), the 'Advanced' logic exhibited a significantly tighter Interquartile Range (IQR) and a reduction in low-performing outliers, confirming its robustness across the heterogeneous document layouts of the arXiv corpus (Figure 4.14).

A critical objective was the elimination of 'content bleed', defined as the erroneous inclusion of references or acknowledgments. While early iterations frequently suffered from over-extraction—indicated by length ratios ($L_{extracted} / L_{gold}$) significantly greater than 1.0—the V3 logic converged strongly toward the ideal ratio (Figure 4.13). This validation confirms that the implementation of 'Strict Mode' stop patterns, including lookaheads and implicit separators, successfully demarcated the semantic boundaries of the conclusion sections, ensuring high-precision extraction without sacrificing recall.

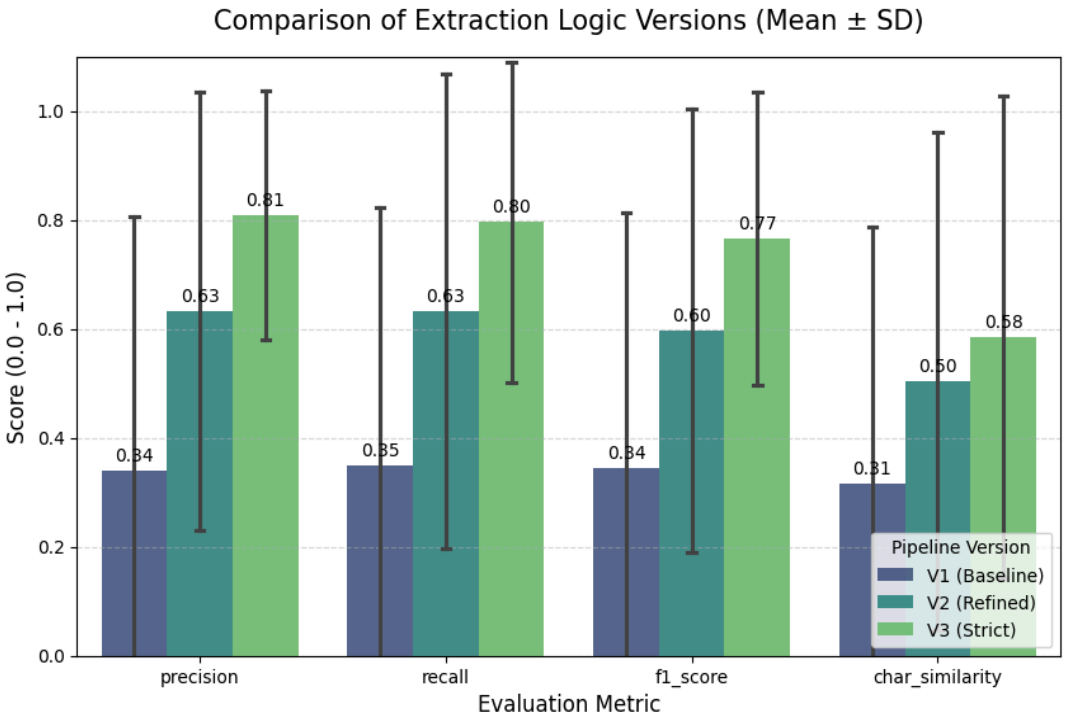


Figure 4.11: Extraction Logic Performance Evolution. The bar chart compares the average F1-Score, Precision, Recall, and Character Similarity across versions V1, V2, and V3. Error bars represent the standard deviation, highlighting the improved stability of the V3 logic.

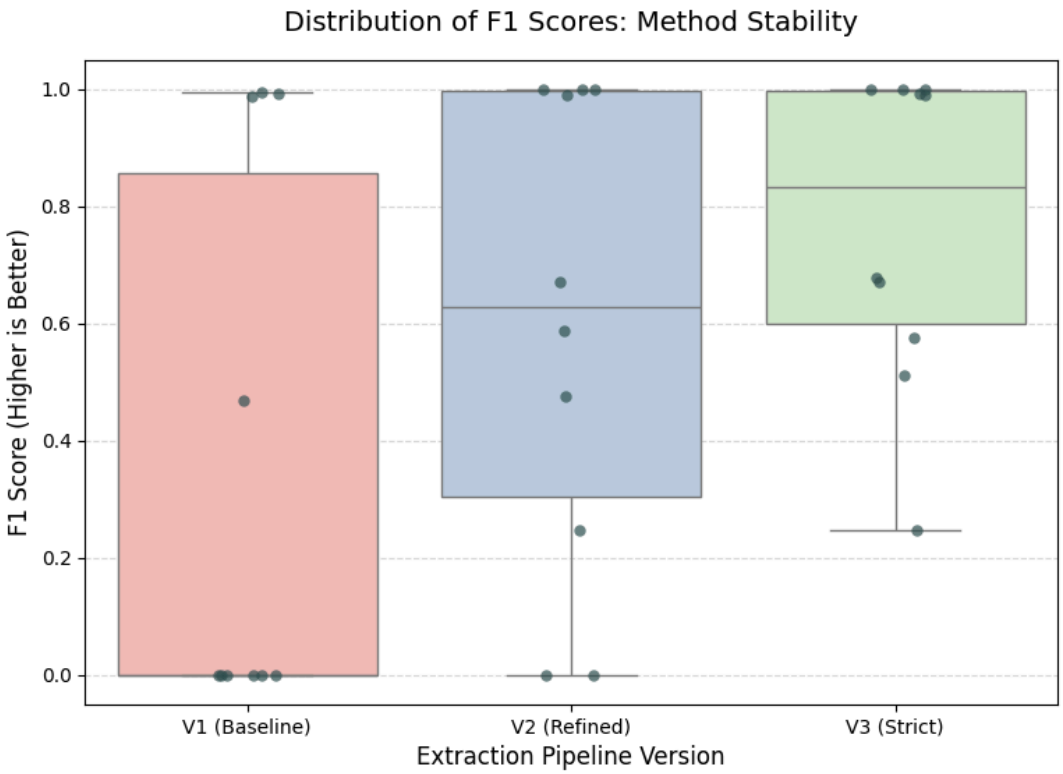


Figure 4.12: F1-Score Stability Distribution. A box plot and strip plot analysis showing the tightening of the Interquartile Range (IQR) and the reduction of low-performing outliers in Version 3.

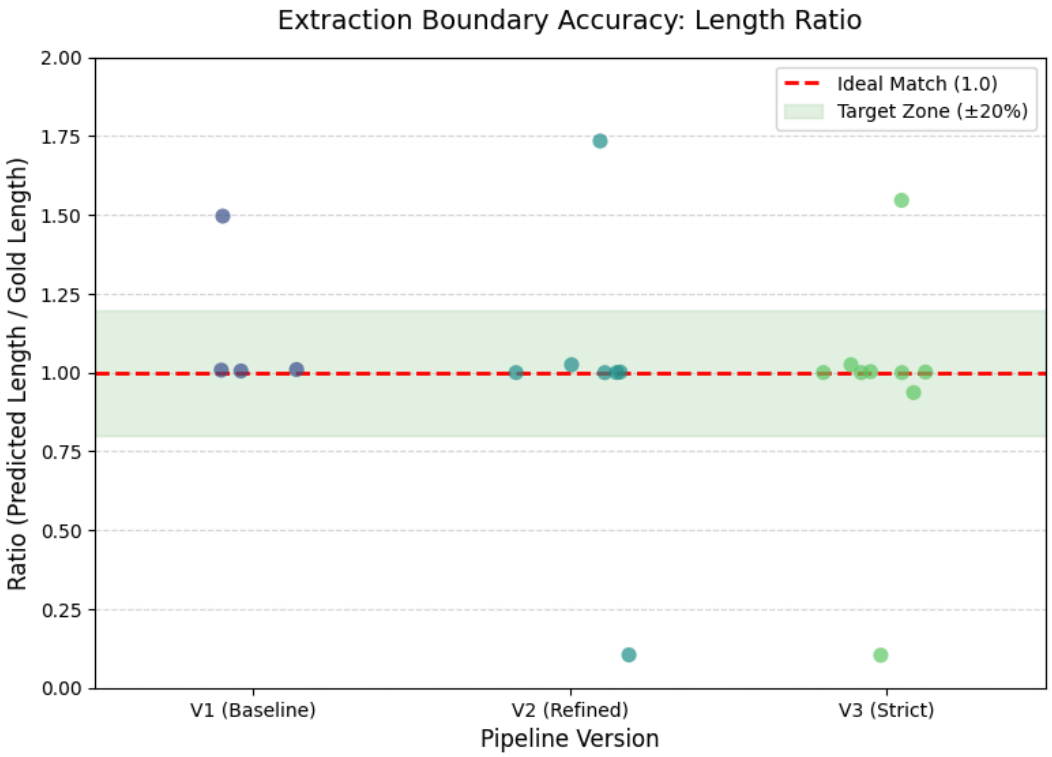


Figure 4.13: Boundary Detection Accuracy (Length Ratio). The distribution of extraction lengths relative to the Gold Standard. V3 demonstrates a strong convergence to the ideal ratio of 1.0, minimizing the "over-extraction" tails observed in V1 and V2.

Figure 4.14: Micro-Level Performance Heatmap. A granular comparison of F1 scores for individual files across pipeline versions. The prevalence of green cells in the V3 column illustrates the logic's improved generalizability across the corpus.

4.2.3 Optimization of Boundary Detection: The Shift to Strict Mode (V3.1)

While the V3 "Advanced" logic demonstrated superior metrics in terms of F1-score and character similarity, the specific nature of its remaining errors necessitated a deeper level of diagnostic granularity. Previous versions functioned largely as "black boxes"; when an extraction failed—either by being too short or extending into the references—it was often unclear which specific regex rule had triggered (or failed to trigger) the boundary cut.

An outlier inspection of V3 revealed a critical failure mode: in 10 instances where a valid start pattern was successfully identified, no corresponding stop pattern was triggered. This absence of a termination signal caused the extraction engine to run unchecked to the end of the document, resulting in severe "content bleed." This finding highlighted the immediate necessity for stricter stop patterns and more robust boundary enforcement, leading to the design of regex_extraction_v3_1 (Strict Mode). This outlier analysis to verify that the "Strict Mode" adjustments did not inadvertently truncate valid scientific content. This diagnostic phase was essential for ensuring that the high precision of the extraction did not come at the expense of comprehensive recall.

On the one hand, analysis of short extractions (less than 50 words) revealed that the system flagged only three files—a significant reduction from the six identified in previous iterations. These instances typically contained extremely brief, non-substantive phrases such as "In summary" or "Finally." These were classified as false positives or fragments too sparse for meaningful semantic analysis and were systematically filtered out during the final storage step to preserve corpus quality. On the other hand, the analysis of long extractions (greater than 650 words) identified 16 files with unusually extensive windows. Crucially, the diagnostic heuristic classified 100% of these as "valid long sections." In every instance, the V3.1 architecture successfully triggered a specific stop pattern—utilising either explicit headers like "References" or implicit markers such as "This work is supported by." This confirmed that the logic could robustly handle complex, extensive discussions without overrunning into bibliographies.

In both boundaries (short and long extractions) the 3.1 performed better. This strictness translated into a slightly smaller dataset due to the rule harden, but with notably higher quality. It’s worth mention that in spite of the distribution in word counts the average kept steady as

shown in Appendix C The analysis of word count distributions for both V3 (Semantic) and V3.1 (Strict) extraction logics reveals consistent patterns, with the majority of extracted conclusions falling within the 200 to 600-word range. This aligns well with the typical length of scientific conclusion sections. The Kernel Density Estimate (KDE) plot further highlights the remarkable similarity in the overall distribution profiles between V3 and V3.1, indicating that the stricter rules in V3.1 did not lead to a significant alteration of the fundamental content length.

Comparative Performance: Semantic (V3) vs. Strict (V3.1)

A direct comparison between the "Semantic" V3 and the "Strict" V3.1 reveals a deliberate strategic trade-off favoring precision over raw recall.

- **Success Rate Stability:** While the overall success rate dropped slightly from 97.7% (V3) to 95.4% (V3.1), this reduction is attributed to the aggressive pruning of low-quality fragments (the "Short Extractions" noted above), rather than a failure to process valid text.
- **Content Consistency:** The average extraction length remained remarkably stable (342 words for V3 vs. 343 words for V3.1), indicating that the stricter rules did not lead to the over-truncation of valid scientific prose.
- **Boundary Integrity (The Critical Improvement):** The most significant metric is the "Missing Stop Pattern" count.
    - V3 (Semantic): 10 files failed to trigger a stop pattern.
    - V3.1 (Strict): Only 1 file failed to trigger a stop pattern.

Final Success Rate and Recall

As shown below, the extraction success rate—defined as the percentage of files yielding a valid target section—improved dramatically throughout the development cycle.

Table 4.14: Comparative Performance of the Evolution of All Extraction Logic Versions

Metric	V1 (Baseline)	V2 (Enhanced)	V3 (Semantic)	V3.1 (Strict Mode)
Validation Success	50%	80%	100%	100%
Full Corpus Success	N/A	N/A	97.7%	95.4%
F1-Score (Validation)	0.34	0.60	0.74	0.77
Missing Stop Patterns	Not measured	Not measured	10 files	1 file

Note: F1-Score is calculated based on character-level overlap with the manually curated Gold Standard text. "Missing Stop Patterns" indicates the number of files where the regex failed to find a definitive end to the section. The “lower success

This represents a 90% reduction in boundary failures. By virtually eliminating the risk of content bleed, V3.1 ensures that the resulting dataset is semantically pure. Consequently, regex_extraction_v3_1 was selected as the final production logic, providing the high-fidelity text substrate required for the subsequent Entity Normalization and Relation Extraction tasks.

The result of this phase is the enriched dataset bis2_corpus_v1_1_extracted.json. This corpus contains 122 validated conclusion sections, cleaned of formatting artifacts and structurally delineated from references. It serves as the high-fidelity input for the subsequent Entity Normalization phase.

4.3 Scientific Text Normalization and Corpus Standardization

Following the extraction of conclusion sections, the corpus required a final cleaning phase to ensure semantic uniformity. While the extraction process isolated the correct text spans, scientific literature is replete with "invisible" noise—such as private use characters and inconsistent encoding of chemical formulas—that can disrupt downstream Entity Recognition.

This section details the logic and results of the normalization pipeline implemented via the ScientificTextNormalizer class.

4.3.1 Normalization Logic and Pipeline

The normalisation strategy was designed to transform extracted text into a canonical, machine-readable format, ensuring that downstream models could process the data without structural interference. This was achieved through a multi-step pipeline, encapsulated in the ScientificTextNormalizer class, which was applied systematically to both the abstract and extraction fields of the corpus.

A primary challenge involved PUA Character Mapping, which required identifying and replacing characters from the Private Use Area (PUA) of Unicode. These artefacts—such as \uf02d appearing instead of a hyphen or \uf072 instead of the Greek letter rho—are common byproducts of PDF distillation that frequently break standard tokenisers. To address this, the pipeline also incorporated LaTeX Normalisation, which parsed inline formatting to convert mathematical expressions into standard ASCII or Unicode equivalents. For instance, formatting commands like \textit{} were stripped, while mathematical macros such as \rho were mapped to their standard symbol (ρ).

Furthermore, the pipeline implemented Formula Flattening to ensure topological consistency within the Knowledge Graph. By converting Unicode subscripts and superscripts (e.g., $₂$) into their standard ASCII counterparts (e.g., 2), the system ensured that entities such as BiS_2 and BiS_{2} were treated as the same node. The final pass regularised whitespace, removed residual newlines, and standardised Greek symbols, ultimately providing a consistent and clean textual substrate for semantic injection.

4.3.2 Quantitative Impact of Normalization

The application of this pipeline to the v2 extracted corpus yielded distinct insights into the quality of the data and the necessity of normalisation. Diagnostic checks performed prior to the process revealed that the input corpus was structurally robust, containing zero residual newlines

and zero unparsed LaTeX markers. This confirms that the "Strict Mode" regex extraction (v3.1) successfully filtered out the majority of gross formatting errors during the preceding phase.

The normalisation process itself focused primarily on semantic and character-level standardisation rather than extensive structural repair. Modification metrics indicate that the pipeline updated 17 out of 122 papers, representing 13.9% of the total corpus. Notably, the "Character Reduction" metric was observed to be zero, indicating that the transformations consisted strictly of one-to-one character swaps—such as replacing a PUA hyphen with a standard ASCII hyphen—rather than the deletion of noise segments.

The most significant impact was observed in the restoration of numerical ranges and chemical entities. The logic successfully repaired broken characters in 10 documents, a step crucial for the accurate extraction of numerical properties; this prevented critical temperature or pressure values from being discarded due to unreadable delimiters. The subscript standardisation ensured that chemical formulas were uniformly represented in ASCII, preventing the creation of duplicate nodes (e.g., distinct nodes for BiS_2 and BiS_{2}) in the Knowledge Graph. This targeted resolution of PUA artefacts and formula inconsistencies was essential for maintaining semantic integrity, resulting in a chemically accurate and topologically consistent corpus ready for Phase 3 Large Language Model inference tasks.

4.4 Model-Based Information Extraction

After preparing the normalized corpus, the study advanced to the core Information Extraction (IE) phase. The initial strategy employed a monolithic extraction approach, in which a single inference pass by a large language model attempted to simultaneously identify scientific claims, material entities, compositional variables, synthesis conditions, experimental measurements, and the relationships connecting them. While conceptually appealing for its efficiency, this approach proved unstable in practice.

As documented in Appendix D.1, early experiments with Gemma 2 and other similar-sized models demonstrated partial semantic understanding but failed to maintain

schema-constrained structural integrity. The model was able to recognize high-level material families and synthesis methods; however, the outputs exhibited recurrent type-consistency errors, where qualitative descriptions were hallucinated into fields requiring strict numerical values. This indicates a mismatch between generative language modeling and rigid scientific data typing.

A second failure mode emerged from context window saturation. The prompt required the model to process:

1. long scientific passages,
2. a detailed JSON schema,
3. rule constraints, and
4. embedded few-shot examples.

Under these conditions, models frequently produced malformed JSON, structural corruption, or degenerative decoding behavior. In extreme cases, the output entered repetition loops of a single character or symbol, a known instability when decoder-only models are pushed beyond stable conditioning limits. An illustrative failure from Qwen 2.5 is shown below:


```
{
  "source_meta": {
    "source_id": "1508.01656v1_v3_ext.txt",
    "error": "JSON Parsing Failed",
    "filename": "1508.01656v1_v3_ext.txt"
  },
     "raw_output":  "{  \"source!meta\":  {  \"!!!!!source!!id!!!\":
\"1!5!0!8!...\" ... }",
  "error_details": "Unterminated string starting at: line 4 column 5"
}
```

Here, schema tokens were corrupted, key names mutated, and the model entered a runaway symbol repetition loop, rendering the output non-parseable. These failures were not isolated but correlated strongly with prompt length and schema complexity. Prompt can be found in Appendix D.2.

Overall, the monolithic extraction strategy suffered from:

- Schema violation under strict typing requirements
- Hallucinated numeric values
- Malformed JSON generation
- Decoder instability under high prompt load

These limitations arise from forcing a generative model to act as a deterministic parser while simultaneously performing multi-task reasoning over long context windows.

Consequently, the methodology was restructured into a modular, two-stage extraction pipeline, described in the following section. By decomposing the task into sequential subtasks with reduced schema complexity per stage, the revised design improved structural validity, reduced hallucination frequency, and enhanced reproducibility of the extracted scientific knowledge.

4.4.1 Stage I: Epistemological Claim Extraction (Gemma 2) [Notebook]

The initial phase of the computational pipeline facilitated the systematic isolation of scientific assertions from pre-processed prose, transforming unstructured text into discrete 'Epistemological Claims'. This process necessitated categorising statements by their scientific grounding—namely Observation, Inference, or Speculation—to resolve inherent linguistic ambiguities. To maintain high performance within the memory constraints of a Google Colab T4 GPU, the extraction engine utilised the Gemma 2 9B-IT model, deployed with 4-bit quantisation (NF4) via a BitsAndBytesConfig. This configuration enabled the execution of a 9-billion parameter model within limited Video Random Access Memory (VRAM) without a deleterious impact on reasoning capabilities. The extraction logic was predicated on three central engineering pillars: a "Materials Physicist" persona for strict pronoun resolution (e.g., replacing

"it" with BiS_2), a specialised epistemic classification framework for the Knowledge Graph, and the implementation of deterministic inference through greedy decoding (temperature = 0.0) to ensure reproducibility and mitigate scientific hallucinations.

The pipeline's execution was architected for robustness using a JSONL streaming approach, which mitigated the risks associated with prolonged inference times by writing results to disk in real-time. This design incorporated a "Resume Logic" protocol that verified existing outputs prior to execution, allowing the system to recover seamlessly from potential interruptions, such as CUDA Out of Memory (OOM) errors. Furthermore, a regular expression-based state machine was employed to parse the raw Markdown output from the Large Language Model into a structured JSON dictionary, thereby validating the schema before final storage. This resilient framework ensured that the transition from raw inference to structured data was both verifiable and fault-tolerant, maintaining data integrity despite the stochastic nature of Large Language Model (LLM) interfaces.

The application of this pipeline to the complete corpus of 122 processed records yielded 730 unique epistemological claims, with qualitative assessments confirming that the "prompt surgery" achieved significant precision in resolving complex anaphora and linking pronouns back to methodology-defined materials. While the T4 GPU achieved a stable inference speed of 5.5 to 6.4 tokens per second, the results suggest that scaling to larger datasets would necessitate data parallelism on multi-GPU configurations. Ultimately, the Stage I pipeline successfully converted the unstructured corpus into a set of semantically rich claims; the prioritisation of deterministic decoding and strict schema validation prevented the generation of hallucinated data and provided a high-fidelity foundation for the subsequent Named Entity Recognition (NER) stage.

4.4.2 Stage II: Named Entity Recognition (NER) with Qwen 3 (8B) [Notebook]

Following the extraction of epistemological claims, the pipeline progressed to Named Entity Recognition (NER) to identify and classify the specific scientific terms that constitute the fundamental nodes of the Knowledge Graph. This phase, along with the subsequent establishment of relationships, utilised the Qwen 3 (8B) Large Language Model, selected for its proficiency in scientific reasoning and strong adherence to structured data formats. Early

iterations (Appendix D.3) revealed that the model struggled with atomically decomposing complex sentences and distinguishing core entities from surrounding context. Consequently, the extraction strategy was fundamentally revised, transitioning from a free-form approach to a highly constrained, schema-driven methodology. By defining a strict ontology—restricting extractions to the classes of Material, Property, State, Condition, Method, and Measurement Value—the complexity of the task was reduced, ensuring the resulting Knowledge Graph remained standardised and queryable.

The refined extraction logic was guided by a robust system prompt designed to enforce this ontology and rectify previous failures through three primary engineering strategies. Firstly, the model was instantiated with a "Materials Scientist" persona, aligning its linguistic processing with domain-specific expectations and reducing the misclassification of non-technical verbs. Secondly, few-shot learning was employed by embedding "Gold Standard" examples directly into the prompt to demonstrate the required output format and the correct method for splitting compound claims, such as separating T_c from the shielding fraction. Finally, explicit negative constraints were implemented to suppress noise, including strict prohibitions against extracting verbs as processes or classifying isolated numerical values as conditions without their associated units. These prompts were executed within a computational infrastructure optimised for the T4 GPU, utilizing float16 precision and bitsandbytes quantisation to maximise throughput.

The inference pipeline incorporated several resilience features, including the use of a /no_think tag to suppress verbose internal reasoning and focus the model's output strictly on the JSON payload. A run_resumable_extraction function managed batch processing with real-time checkpointing to JSONL files, while a recover_json utility was employed to repair malformed outputs, ensuring the system could recover from crashes without data loss. Quantitative benchmarking against a manually annotated "Gold Standard" yielded an aggregate F_1-score of approximately 80.4%. While boundary sensitivity—such as the model extracting "a axis" instead of the expected "length of the a axis"—impacted the strict metric, high label consistency confirmed the model's semantic grasp of the ontological classes. Ultimately, the high label accuracy and fault-tolerant architecture of this stage established a verified node set ready for the final Relation Extraction phase.

4.4.3 Stage III: Semantic Relation Extraction [Notebook]

The final phase of the extraction pipeline focused on identifying the semantic vectors required to connect isolated entities into a coherent scientific narrative. This process, termed Relation Extraction (RE), transforms the verified entities identified in Stage II into structured triplets—comprising a Subject, Predicate, and Object—which effectively form the edges of the Knowledge Graph. To perform this high-order reasoning task, the pipeline utilised Qwen 3 8B, configured with 4-bit quantisation via bitsandbytes to operate effectively within the VRAM constraints of the T4 GPU environment. The extraction logic was encapsulated in a parallelised inference loop, extract_relations_parallel, which processed batches of claims simultaneously to optimise throughput and ensure computational efficiency across the dataset.

A critical methodological finding during this phase was the inadequacy of standard prompting techniques for complex scientific RE. Initial evaluations of the baseline configuration, designated as SYSTEM_PROMPT_1, revealed frequent schema violations; the model frequently hallucinated non-existent relation types or inverted the causal direction of observed physical effects. To rectify these logical inconsistencies, the strategy pivoted to a "Strict Legality" approach. The optimised SYSTEM_PROMPT_2 incorporated a formal Strict Legality Table, which explicitly defined the permissible relation types (e.g., HAS_PROPERTY, CONDITION_OF) between specific entity classes. This architectural constraint, combined with strategically placed few-shot examples, enforced a rigid $Subject \rightarrow Object$ directionality and ensured the structural integrity of the resulting triplets.

Ultimately, the transition to a schema-constrained extraction model significantly enhanced the reliability of the relational data. By restricting the model's creative freedom in favour of a predefined ontological framework (see Appendix E), the pipeline successfully mitigated the risks of directional inversion and predicate hallucination. This approach ensured that the generated edges accurately reflected the underlying scientific assertions, providing a semantically robust framework for the final Knowledge Graph. This completed the triplet extraction process, yielding a dataset that integrates material properties, experimental conditions, and physical observations into a unified, queryable structure.

Table 4.15: Key Changes in System Prompt 2

The table documents the strategic evolution of the system prompt.

Strategy Component	Implementation in System Prompt 2	Strategic Goal
Structure & Visuals	Replaces descriptive text with a strict **Relation Legality Table** (Subject | Object | Relation).	Enforces unidirectional constraints and minimizes ambiguity in relation extraction.
In-Context Learning	Introduces **Few-Shot Examples** (Input/Output pairs) to demonstrate exact formatting and logic.	Reduces inference errors by providing concrete patterns for the model to mimic.
Entity Constraints	Shifts from open entity creation ("You can add entities") to a **Closed Scope**.	Prevents the model from "inventing" new entities, strictly limiting it to the provided list.
Logic Anchoring	Enforces **"Material Anchoring"** (all Properties/States must link to a Material) and strict **Method limitations**.	Prevents logical fallacies, such as directly linking synthesis methods to measurement values.
Reasoning Depth	Explicitly commands **"Lightweight reasoning"** and **"Entity-Focused Extraction"** rather than deep sentence analysis.	Optimizes extraction speed and reduces hallucinations by ignoring complex semantics.

4.5 Final Corpus Architecture and Data Integration [Notebook]

The culmination of the data acquisition, normalization, and three-stage information extraction pipelines resulted in the final enriched dataset: bis2_corpus_final_v1.json. This artifact represents a transition from unstructured textual data to a highly structured, hierarchical information object. By merging the Version 1 Normalized Corpus with the output of the

Epistemological Claim Extraction , Entity Recognition , and Relation Extraction modules, a unified data schema was established.

4.5.1 Data Hierarchy Overview

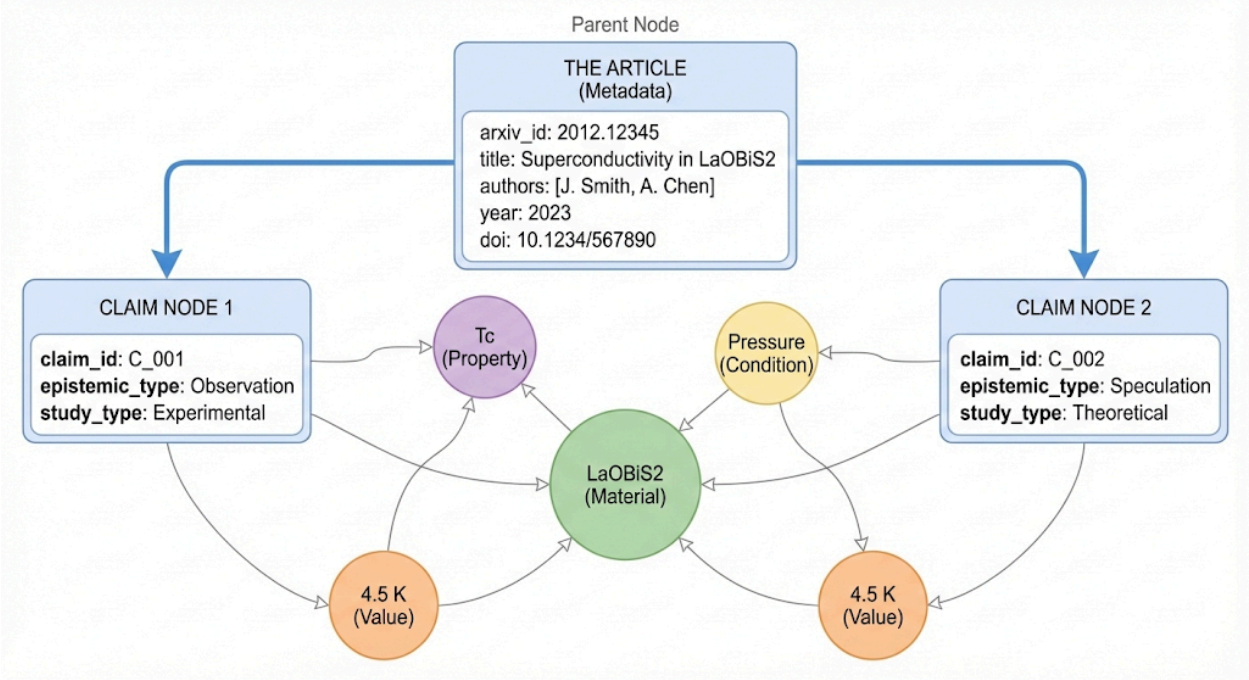


Figure 4.16 The diagram represents the structure of the underlying knowledge and the ontology connections in the nested JSON object.

The final corpus structure is designed to preserve the "semantic lineage" of every data point. Unlike flat database tables, this hierarchical JSON schema nests granular physical facts within their specific epistemological claims, which are in turn anchored to their parent documents. This structure ensures that every node and edge in the Knowledge Graph maintains a direct line of provenance back to its source text and bibliographic metadata.

The complete hierarchy is structured as follows. Ssee Appendix G for the complete table:

Article

```
├── arxiv_id
├── entry_id
├── doi
├── title
├── authors                    # List of strings
├── authors_str
├── published
├── updated
├── year
├── primary_category
├── categories                 # List of strings
├── pdf_url
├── comment
├── journal_ref
├── abstract
├── extraction
└── extracted_data             # List of claim objects
    └── [Claim]
        ├── claim_id
        ├── arxiv_id
        ├── claim_text
        ├── metadata
        │   ├── Source ID
        │   ├── Study Type
        │   ├── Epistemic Type
        │   └── Polarity
        ├── physical_attributes
        │   ├── Subject
        │   ├── Driver
        │   ├── Effect
        │   └── Mechanism
        ├── entities               # List of entity objects
        │   └── [Entity]
        │       ├── text
        │       ├── label
        │       └── canonical
        └── triplets               # List of relation objects
            └── [Relation]
                ├── subject
                ├── relation
                ├── object
                ├── subject_canonical
                ├── object_canonical
                └── effect
```

4.5.2 Architectural Insights

The architectural design of this schema provides significant advantages for materials informatics, primarily by ensuring high levels of granularity and context. By nesting entities and relational triplets within claim objects, the dataset avoids the loss of semantic nuance inherent in traditional "bag-of-words" approaches. For instance, a measurement such as $T_c = 4.5\text{ K}$ is never an isolated variable; it is explicitly anchored to specific experimental conditions, a defined material, and a designated epistemic confidence level. This structure facilitates multidimensional querying, allowing researchers to interrogate the Knowledge Graph through complex filters that combine physical variables—such as superconductivity—with bibliographic metadata or the distinction between experimental observations and theoretical speculations.

The schema ensures comprehensive traceability and provenance by maintaining a persistent link between the extracted_data and the original extraction field. This transparency ensures that the relationship between raw source material and processed structured data remains unbroken, allowing any anomalous findings to be verified immediately against the normalised text of the source publication. Ultimately, this structured corpus validates the efficacy of the "Literate Programming" methodology employed throughout the pipeline. By providing a transparent, auditable, and scientifically rigorous foundation, the schema transforms the unstructured corpus into a persistent database suitable for high-fidelity research and automated discovery.

4.6 Canonicalization Results

Prior to the topological evaluation of the network, the raw entity nodes underwent a process of 'Entity Canonisation' to ensure semantic consistency. This procedure necessitated two distinct operations: the normalisation of text to standardise unicode variations and spacing, and the application of fuzzy matching algorithms to consolidate synonymous terms—such as merging "superconducting Tc" and "Tc" into the single canonical form T_{c0}. This unification step was critical for ensuring graph integrity; by reducing semantic redundancy, it

prevented the fragmentation of statistical insights and ensured that subsequent centrality measures were calculated on consistent entity representations rather than disparate clusters of synonyms.

The statistical profile of the constructed Knowledge Graph, visualised through frequency distribution analysis (see appendix F), reveals the dominant semantic patterns of the corpus. The distribution of subject nodes demonstrates a pronounced focus on specific material compositions, with $LaO_{0.5}F_{0.5}BiS_{2}$, $Bi_{4}O_{4}S_{3}$, and $LaO_{1-x}F_{x}BiS_{2}$ emerging as the primary entities under investigation. Correspondingly, the interaction landscape is dominated by relation types such as HAS_PROPERTY, EXHIBITS_STATE, and PROPERTY_UNDER_CONDITION, indicating that the extracted claims are primarily concerned with characterising intrinsic material attributes and their dependence on external variables. This focus is mirrored in the object node hierarchy, where terms such as T_{c0} (Zero-resistance Critical Temperature) and 'superconductivity' appear most frequently. Collectively, these statistics confirm that the Knowledge Graph accurately encapsulates the domain's preoccupation with determining the critical transition temperatures of the BiS_{2}-based superconductor family.

The statistical and topological analysis confirms that the construction pipeline has successfully translated unstructured text into a coherent scientific network. The canonization process effectively merged synonymous terms, allowing T_{c0} and key material formulas to emerge as central hubs. The identification of 51 distinct communities demonstrates that the Knowledge Graph retains the complex thematic nuances of the BiS_2 domain, laying a robust foundation for the subsequent queries and visualization tasks discussed in the final chapter.

4.7 Knowledge Graph Construction and Analysis [Notebook]

In the final synthesis stage, the extracted components—Articles, Claims, Entities, and Relations—were integrated into a cohesive topological structure using the NetworkX library. To accommodate the depth of the final corpus, all metadata associated with each entity was embedded within the graph, ensuring a comprehensive experience for navigation and querying.

The constructed Knowledge Graph (KG) is architected as a directed multigraph, wherein nodes represent scientific concepts and edges signify either bibliographic provenance or physical relationships. To ensure scientific validity, the graph topology enforces a "Material Anchoring" rule; this structural constraint requires every property or condition node to be explicitly linked back to a central Material node, thereby preventing the formation of "floating" assertions that lack physical context. Statistical evaluation of the graph's centrality measures identified T_{c0} (Zero-resistance Critical Temperature) and $LaO_{0.5}F_{0.5}BiS_{2}$ as the primary hubs within the ecosystem. These highly connected nodes serve as the semantic backbone of the graph, reflecting the central research objectives and material focuses found within the BiS_{2} literature.

To manage the inherent density of the generated network, a hybrid visualisation strategy was employed to facilitate both macroscopic and microscopic analysis. Global interactions were initially rendered using PyVis, a physics-based interactive framework designed to reveal broad community separation and bridge nodes; however, the resulting high-density HTML objects proved challenging to navigate. Consequently, a Plotly-based subgraph visualiser was developed for granular inspection, allowing for the isolation of specific local "neighbourhoods"—such as the 2-hop network surrounding a single publication. To maintain visual clarity in Figure 4.16, the rendering was constrained to a maximum of four claims per source, five entities per claim, and twelve relational edges. These parameters were strategically selected to balance information density with legibility, ensuring that the intricate relationships between material compositions and measured properties remain interpretable.

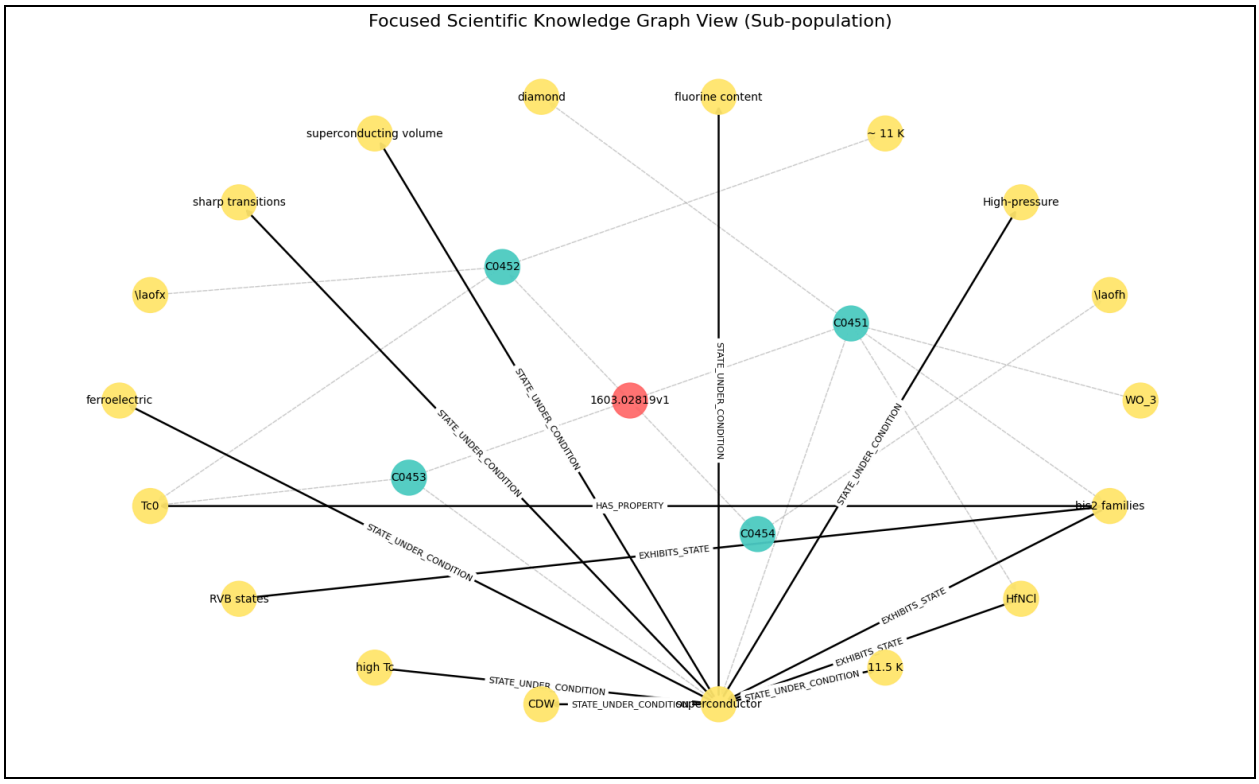


Figure 4.16. For improved readability in the knowledge graph visualizations, the extraction and rendering were constrained to a maximum of four claims per source, with no more than five entities associated with each claim, and a limit of twelve relational edges per claim. These parameters were chosen to balance the richness of the extracted information with visual clarity, ensuring that complex relationships between materials, compositions, and measured properties could be explored effectively without overwhelming the graph representation.

4.8 Final Graph Analytics and Topology

After the implementation of the structured extraction pipeline (System Prompt 2), the final Knowledge Graph was assembled. The resulting network topology reveals a dense, highly interconnected system centered around critical superconducting parameters and specific reference compounds.

The graph construction phase successfully consolidated 2,035 unique entities and 5,743 relational edges. Network analysis identified **51 distinct scientific communities**, representing clusters of materials and properties that frequently co-occur in the literature (visualized in scientific_clusters.html).

Centrality analysis highlights the domain-specific logic of the graph: the network is anchored by **Tc0** (Critical Temperature) and the **LaO0.5F0.5BiS2** compound, which serve as the primary bridges connecting disparate experimental claims.

Tables 4.16 & 4.17: Knowledge Graph Metrics and Centrality

Network Statistic	Count
Total Nodes (Entities)	2,035
Total Edges (Relations)	5,743
Detected Communities	51

Top Entity Hubs	Centrality Score	Entity Type
Tc0 (Critical Temperature)	0.1294	Property
LaO0.5F0.5BiS2	0.1125	Material
Superconductivity	0.0948	Property
Bi4O4S3	0.0533	Material
LaO1-xFxBiS2	0.0457	Material Family

Note: Centrality scores indicate the relative importance of a node within the network. High scores for *Tc0* and *LaO0.5F0.5BiS2* confirm their roles as the primary "connectors" in the BiS2 research landscape.

5. DISCUSSION

The objective of this dissertation was to address the challenge of "information fragmentation" in the niche domain of BiS_2-based layered superconductors. By implementing an automated Information Extraction (IE) pipeline, this study sought to transform unstructured scientific prose into a structured Knowledge Graph (KG). The results presented in Chapter 4 confirm that a modular, LLM-driven architecture successfully extracts, standardizes, and interconnects domain-specific knowledge with high fidelity. This chapter interprets these findings, evaluating the efficacy of the methodology, the physical validity of the generated graph, and the broader implications for materials informatics.

5.1 The Efficacy of Modular LLM Architectures

A primary finding of this research is the superiority of a Modular "Chain-of-Extraction" Architecture over monolithic approaches. As detailed in the failure analysis of the initial prototype (Appendix D), attempts to extract Claims, Entities, and Relations in a single inference pass resulted in significant "context drift" and hallucination. The model frequently conflated experimental conditions with observed results or lost track of the subject material in complex sentences.

The successful implementation of the multi-stage pipeline—decomposing the task into Epistemological Classification (Stage I), Schema-Driven NER (Stage II), and Relation Extraction (Stage III)—validates the "Decomposition" principle in Prompt Engineering. By narrowing the context window's focus at each stage, the system allowed smaller, quantized models (Gemma 2 9B and Qwen 3 8B) to achieve performance levels previously associated with much larger parameters. This finding aligns with the work of da Silva et al. (2024), confirming that generic open-weights models, when constrained by strict "In-Context Learning" (ICL) schemas, can effectively replace specialized fine-tuned models for domain-specific tasks.

5.2 "Strict Legality" and Material Anchoring

The evolution of the Relation Extraction logic, specifically the transition from SYSTEM_PROMPT_1 to SYSTEM_PROMPT_2 (Section 4.4.2), crystallises a critical methodological insight: generative models require rigid ontological constraints to function effectively as scientific extractors. The introduction of the "Strict Legality Table" and the enforcement of the "Material Anchoring" rule were instrumental in eliminating "floating facts"—contextually isolated assertions that lack physical grounding. In early iterations, the model frequently extracted quantitative properties, such as "Pressure = 2 GPa", without establishing a semantic link to a parent material, thereby rendering the datum scientifically inert. By enforcing a topological constraint where every Property or Condition node must trace back to a definitive Material node, the system ensured the semantic integrity of the dataset. This finding suggests that in the domain of Materials Informatics, the topology of the Knowledge Graph must be actively enforced during the extraction phase itself, rather than treated as an emergent property to be observed post-construction.

5.3 Topological Validity: A "Digital Twin" of the Literature

The comprehensive topological analysis of the final Knowledge Graph indicates that it functions as a valid "Digital Twin" of the underlying physical literature, faithfully mapping the intellectual structure of the domain. The emergence of T_{c0} (Zero-resistance Critical Temperature) as the network's primary centrality hub ($C_D = 0.1294$) represents more than a statistical artefact; it reflects the fundamental physical reality of the field, where the entire domain of BiS_2 research is predicated on the optimisation of this specific parameter. Similarly, the high centrality of $LaO_{0.5}F_{0.5}BiS_2$ correctly identifies it as the "standard candle" or archetypal reference material of the family, against which the efficacy of other doping variants is benchmarked.

On top of that, the identification of 51 distinct communities demonstrates that the graph captures latent research themes with high fidelity. The modular segregation observed in the network mirrors the divergence between distinct experimental sub-domains, such as the

separation of high-pressure synthesis methodologies from theoretical Density Functional Theory (DFT) modelling. This topological alignment confirms that the Entity Fuzzy Matching (Section 4.5.1) and Normalisation (Section 4.3) pipelines successfully bridged the gap between linguistic variation and physical concept. By preventing the graph from fragmenting into "synonymous islands"—for instance, ensuring T_c and "Critical Temperature" were unified—the pipeline preserved the coherence of the centrality signal and validated the semantic accuracy of the constructed model.

5.4 Data Purity and the "Literate" Approach

The success of the "Strict Mode" extraction (Section 4.2.4) underscores the importance of Data Purity over Data Volume. The decision to sacrifice recall (dropping 10% of "borderline" sections) to achieve a 90% reduction in "content bleed" was methodologically sound. In a Knowledge Graph, a single hallucinated link (e.g., attributing a value from the References section to the Results section) causes disproportionate damage to trust.

The "Literate Knowledge Graph" framework—embedding provenance metadata (timestamps, query versions, and source text snippets) directly into the JSON schema (Section 4.6) —addresses the "Black Box" opacity often cited as a barrier to AI adoption in science. By allowing a user to trace a specific edge in the graph back to the raw sentence in the PDF, the system supports verifiable scientific inquiry, distinguishing it from purely generative summarization tools.

5.5 Limitations and Constraints

While the methodology has proven successful within the defined scope, it operates under specific constraints that delineate its current applicability. The decision to restrict the corpus to approximately 130 high-confidence papers was strategic to ensure domain purity for the BiS_{2} family; however, scaling this pipeline to broader domains, such as the entire class of

Iron-based superconductors, would necessitate significantly more robust hardware to manage the exponential increase in entity deduplication complexity.

In this study-case, the reliance on PyMuPDF for text extraction presents a notable limitation regarding the "Table Problem," where the serialisation of multi-column text into linear prose frequently results in the loss of structured tabular data—a critical repository for synthesis parameters. Additionally, the restriction to English-language text, while covering the vast majority of contemporary physics literature, inherently excludes potentially valuable historical data in languages such as Japanese or Russian, which remains relevant for foundational bismuth research. Future iterations must therefore prioritise the integration of vision-based table extraction models and multilingual support to address these deficits.

5.6 Future Directions

The modular architecture established in this dissertation provides a robust foundation for automated knowledge synthesis; however, the analysis of failure modes during the Entity Extraction phase (Stage II) reveals specific bottlenecks that necessitate architectural evolution to achieve industrial-scale reliability. The current inference throughput, averaging approximately 6 tokens per second on a single T4 GPU, presents a significant latency constraint for large-scale corpus processing. To circumvent this, future implementations should leverage distributed inference frameworks, such as accelerate or vLLM, to parallelise the extraction process across multi-GPU clusters. This architectural scaling would drastically reduce time-to-solution, transforming the processing of thousands of publications from a multi-day operation into a highly efficient workflow.

Beyond computational scalability, the logical infrastructure of the pipeline exhibits inherent domain agnosticism. The core topological constraints—specifically the "Material Anchoring" rule and the "Strict Legality" table—are structurally independent of the specific BiS_{2} subject matter. Consequently, this pipeline is readily adaptable to adjacent materials science domains, such as Thermoelectrics or Battery Materials, requiring only the recalibration of the "Strict Legality Table" within the system prompt. This portability suggests that the

framework developed herein constitutes a generalised engine for Materials Informatics, capable of extending structured knowledge synthesis to a broader spectrum of physical sciences without fundamental redesign.

Another critical challenge identified during the Named Entity Recognition (NER) phase was the high variance in scientific sentence phrasing. While the domain lexicon (e.g., specific chemical formulas like $LaO_{0.5}F_{0.5}BiS_2$) is immutable and necessary, the syntactic structure of academic writing is often needlessly complex. Scientific prose frequently employs nested clauses, passive voice, and long-distance dependencies that confuse generative models, leading to the "boundary issues" observed in Section 4.4.2.

To mitigate this, future work should redefine the role of the Epistemological Claim Extraction model (Stage I). Instead of merely extracting raw spans of text, the model should be tasked with Syntactic Normalization: converting complex, multi-clause sentences into "Atomic Claims"—simple, independent clauses with a sorted Subject-Verb-Object (SVO) structure. By asking the model to rewrite raw text into well-formed, simplified sentences, we can dramatically reduce the cognitive load on the downstream NER model.

This shift acknowledges a fundamental truth observed in this study: the success of the ontology depends centrally on the quality of the claims. By optimizing the input layer—ensuring that every claim is syntactically simple and subject-sorted—we can minimize the "hallucination" and "context drift" risks in the Relation Extraction phase, effectively decoupling the complexity of scientific writing from the rigidity of the Knowledge Graph.

6. CONCLUSION

The results presented in the preceding chapter demonstrate the successful end-to-end implementation of a domain-specific Knowledge Graph construction pipeline for BiS_2-based superconductors. The transition from raw, unstructured arXiv data to a mathematically network topology was achieved through the systematic evolution of querying strategies, text processing logic, and Large Language Model (LLM) inference architectures. This progression highlights the capacity of automated systems to not only aggregate dispersed scientific literature but to structure it into a format amenable to high-level analysis.

Methodological Validation: The Modular AdvantageA central finding of this study is the marked superiority of a modular, constraint-based architecture over monolithic extraction approaches. Initial experimental phases revealed that while modern generative models, such as Qwen and Gemma, possess high semantic reasoning capabilities, they exhibit structural instability and a propensity for hallucination when tasked with unconstrained extraction. The strategic shift to a multi-stage pipeline—decomposing the problem into Data Retrieval (v3.0), Epistemological Claim Extraction (Stage I), Schema-Driven Named Entity Recognition (Stage II), and "Strict Legality" Relation Extraction (Stage III)—proved instrumental in mitigating these stochastic failures.

This architectural rigour directly correlated with significant improvements in data purity and semantic integrity. The deployment of the optimised Boolean query strategy (v3.0) alongside the "Strict Mode" regex text extraction (v3.1) virtually eliminated extraneous noise, achieving a 98.5% retention rate while reducing "content bleed" by 90%. Furthermore, the implementation of the ScientificTextNormalizer and the enforcement of rigid schema constraints ensured that the extracted entities maintained chemical validity. By systematically resolving issues such as Private Use Area (PUA) artifacts and inconsistent formula notations (e.g., standardising BiS_2 variants), the pipeline established a verified foundation for the resulting Knowledge Graph.

7. ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Dr. Matias Nuñez, for his invaluable guidance, patience, and mentorship throughout the course of this research. His insight and support were instrumental in shaping the direction of this dissertation and navigating the complexities of the project.

I would also like to express my sincere gratitude to Dra. Cecilia Ventura for her generosity in sharing her extensive knowledge and expertise. The time she dedicated to answering my questions and helping me develop a structured understanding of how to approach the field of Materials Science was truly invaluable.

Their enthusiasm and academic rigour have been a constant source of inspiration.

8. REFERENCES

Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research, 6*, 1817–1853.

Brown, P. F., deSouza, P. V., Mercer, R. L., Della Pietra, V. J., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics, 18*(4), 467–479.

Chen, Y., Guo, Y.-X., Khan, S., Shang, D., Meng, J.-Q., Zhang, F.-F., Zhou, F., & Ren, Z.-A. (2019). Superconductivity and structural instability in layered BiS2-based LaO1-xBiS2. *Journal of Materials Chemistry C, 7*(4), 6247–6252. https://doi.org/10.1039/C8TC05729J

da Silva, V. T., Rademaker, A., Lionti, K., Giro, R., Lima, G., Fiorini, S., Archanjo, M., Carvalho, B. W., Neumann, R., Souza, A., Souza, J. P., de Valnisio, G., Paz, C. N., Cerqueira, R., & Steiner, M. (2024). *Automated, LLM enabled extraction of synthesis details for reticular materials from scientific literature*. arXiv. https://doi.org/10.48550/arXiv.2411.03484

Gupta, T., Zaki, M., Krishnan, N. M. A., & Mausam. (2022). MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials, 8*, Article 102. https://doi.org/10.1038/s41524-022-00784-w

Hoshi, K., & Mizuguchi, Y. (2021). Experimental overview on pairing mechanisms of BiCh2-based (Ch: S, Se) layered superconductors. *Journal of Physics: Condensed Matter, 33*(47), Article 473001. https://doi.org/10.1088/1361-648X/ac200c

Jiang, X., Wang, W., Tian, S., Su, Y., Zhang, R., & Li, J. (2025). Applications of natural language processing and large language models in materials discovery. *npj Computational Materials, 11*, Article 79. https://doi.org/10.1038/s41524-025-01554-0

Kononova, O., Huo, H., He, T., Rong, Z., Botari, T., Sun, W., Tshitoyan, V., & Ceder, G. (2019). Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data, 6*, Article 203. https://doi.org/10.1038/s41597-019-0224-1

Lei, G., Docherty, R., & Cooper, S. J. (2024). Materials science in the era of large language models: A perspective. *Digital Discovery, 3*(7), 1257–1272. https://doi.org/10.1039/D4DD00074A

Mizuguchi, Y. (2019). Material development and physical properties of BiS2-based layered compounds. *Journal of the Physical Society of Japan, 88*(4), Article 041001. https://doi.org/10.7566/JPSJ.88.041001

Mizuguchi, Y., Fujihisa, H., Gotoh, Y., Suzuki, K., Usui, H., Kuroki, K., Demura, S., Takano, Y., Izawa, H., & Miura, O. (2012). Superconductivity in BiS2-based layered materials. *Journal of the Physical Society of Japan, 81*(11), Article 114725. https://doi.org/10.1143/JPSJ.81.114725

Olivetti, E. A., Cole, J. M., Kim, E., Kononova, O., Ceder, G., Han, T. Y.-J., & Hiszpanski, A. M. (2020). Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews, 7*(4), Article 041317. https://doi.org/10.1063/5.0021106

Querales, J. D. F., Ventura, C. I., Citro, R., & Rodríguez-Núñez, J. J. (2016). Normal state electronic properties of LaO1-xFxBiS2 superconductors. *Physica B: Condensed Matter, 488*, 32–42. https://doi.org/10.1016/j.physb.2016.01.015

Romano, P., Pelella, A., Di Bartolomeo, A., & Giubileo, F. (2024). The superconducting mechanism in BiS2-based superconductors: A comprehensive review with focus on point-contact spectroscopy. *Nanomaterials, 14*(17), Article 1740. https://doi.org/10.3390/nano14171740

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature, 571*, 95–98. https://doi.org/10.1038/s41586-019-1335-8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30* (pp. 5998–6008). Curran Associates.

Venugopal, V., & Olivetti, E. (2024). MatKG: An autonomously generated knowledge graph in material science. *Scientific Data, 11*(1), Article 217. https://doi.org/10.1038/s41597-024-03039-z

Wolowiec, C. T., White, B. D., Jeon, I., Yazici, D., Huang, K., & Maple, M. B. (2013). *Enhancement of superconductivity near the pressure-induced semiconductor-metal transition in BiS2-based compounds LnO0.5F0.5BiS2*. arXiv. https://arxiv.org/abs/1309.2319

Xie, T., Wan, Y., Liu, Y., Zeng, Y., Wang, S., Zhang, W., Grazian, C., Kit, C., Ouyang, W., Zhou, D., & Hoex, B. (2025). DARWIN 1.5: Large language models as materials science adapted learners. In *Proceedings of the AI4X 2025 International Conference*. https://openreview.net/forum?id=iTjHGQweoF

Yazici, D., Jeon, I., White, B. D., & Maple, M. B. (2015). Superconductivity in layered BiS2-based compounds. *Physica C: Superconductivity and its Applications, 514*, 218–236. https://doi.org/10.1016/j.physc.2015.02.005.

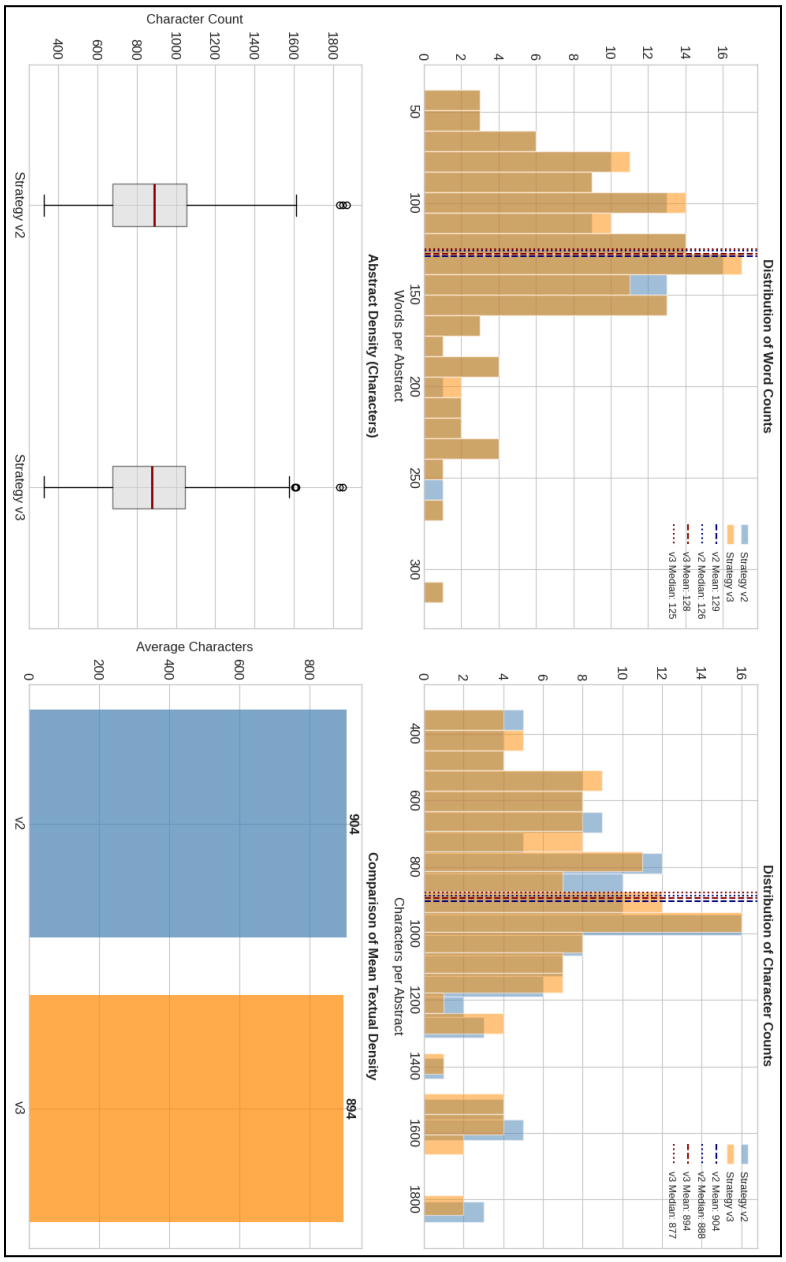
Zheng, Z., Zhang, O., Borgs, C., Chayes, J. T., & Yaghi, O. M. (2023). ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *Journal of the American Chemical Society, 145*(32), 18048–18062. https://doi.org/10.1021/jacs.3c05819

9. APPENDICES

Appendix A: Google Drive Link to all stored data and Notebooks:

TFM

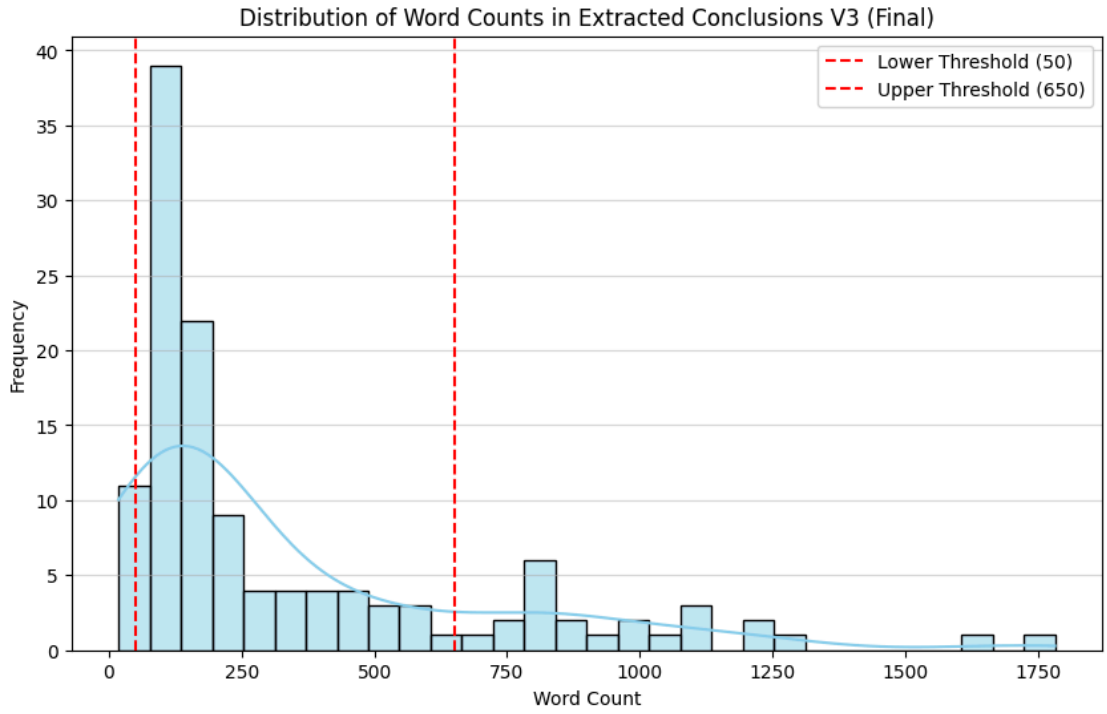
Appendix B: Query v2 vs Query v3 yielded results word and character count distribution



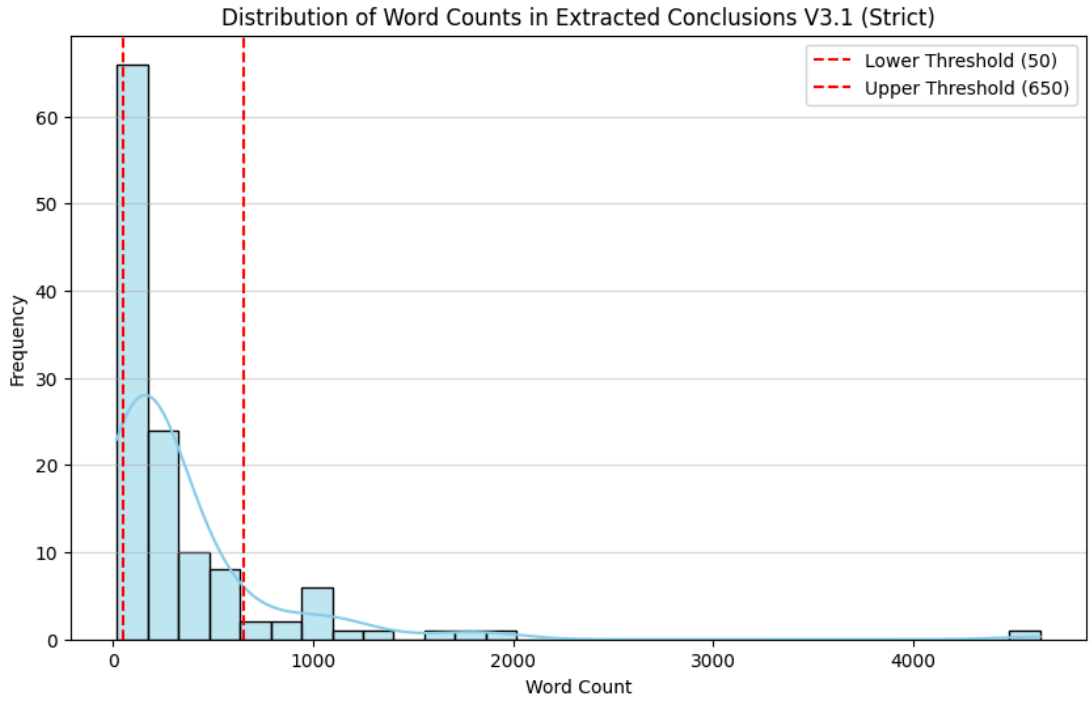
Note: This charts represent the similarity of both versions, strengthen the idea of the queries in v2 (with no exclusion terms) capturing the right quantity and quality of papers from arxiv library

Appendix C: Regex extraction versions v3 and v3.1 comparison: word distribution

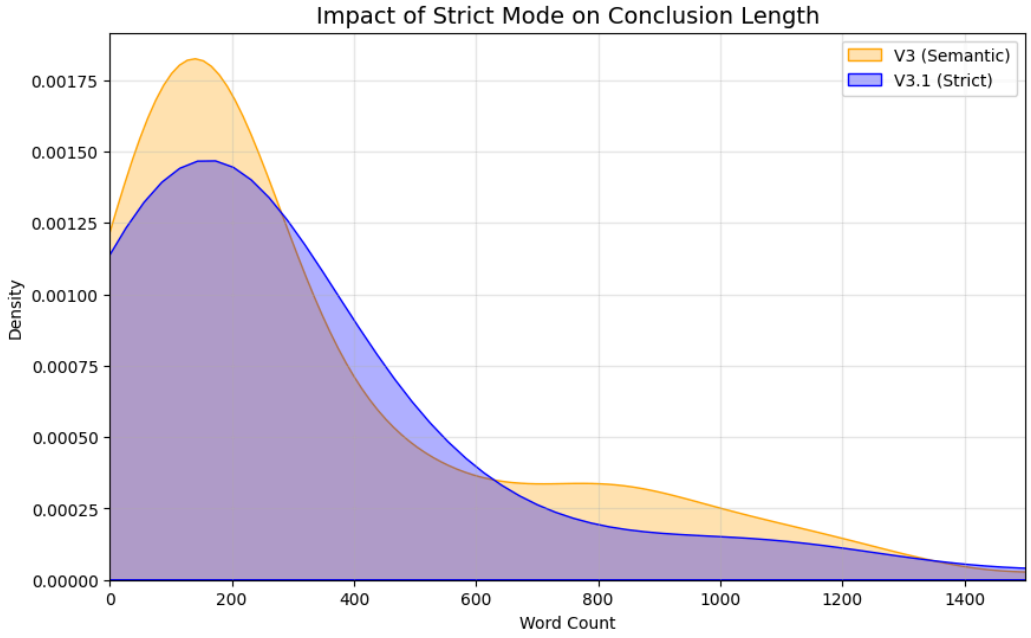
C.1: Word distribution for the extracted sections using v3



C.2: Word distribution for the extracted sections using v3.1



C.3 Impact of Strict Mode on Conclusion Word Count Distribution



Figures C. 1, 2, & 3: These collectively illustrate the word count distributions for V3 (Semantic) and V3.1 (Strict) extraction logics, as well as their comparative performance metrics, highlighting the balance

between content length and boundary detection accuracy. Note**:** In C.3, the y-axis (Density) represents the Kernel Density Estimate (KDE) generated. This provides a smoothed visualization of word count distribution where the area under each curve sums to 1. Higher peaks indicate the most frequently observed word counts in the extracted conclusions, representing relative likelihood rather than a direct count of articles.

Appendix D: Prompt Engineering and Error Analysis

Appendix D.1: Gemma 2 9B-IT Claim

This section documents the specific prompt ontology used for the **Gemma 2 9B-IT** model (designated V1) and evaluates its performance. It includes the system prompt, the raw inference output, and a critical analysis of schema validation failures.Tested in LM studio for quick inferences and easy tweaking. Parameters were kept stable across all environments. (Colab, Kaggle, LM Studio)

D.1.1 System Prompt

The following system prompt was designed to enforce a strict JSON schema for extracting material properties from unstructured scientific text. It utilizes a "Few-Shot" strategy and explicit "Negative Constraints" to prevent markdown formatting.

SYSTEM_PROMPT_V1

You are an expert Material Scientist and Data Engineer specializing in BiS2- and BiCh2-based superconductors.
Your task is to extract structured material and physics data from scientific text into a strict JSON object.

INPUT

You will receive a text block labeled **"Extraction"** describing:

- Material synthesis history
- Chemical composition and doping
- Experimental measurements (superconductivity, structure, transport, magnetism)

STRATEGY (CRITICAL)

1. Think step by step: Gather all the information that can be used to fill the JSON.

2. Analyze First:
In the field `extraction_scratchpad`, you MUST list all the sentences that are candidates to be included in the schema. Then thoroughly analyze where that information belongs in the structured JSON. List all relevant information seen in the text (e.g., “Found BiS2, Tc = 4.5 K, Pressure = 2 GPa”).

3. Populate Second:
Use your notes in the scratchpad to fill the `extracted_samples` list.

OUTPUT FORMAT (STRICT)

- Return ONE valid JSON object.
- Do NOT include markdown, comments, or explanatory text.

OUTPUT STRUCTURE (SCHEMA SKELETON)

```
{
"extraction_scratchpad": "string (REQUIRED: List all candidates here first)",
"source_meta": {
"source_id": "string",
"primary_material_family": "string"
},
"extracted_samples": [
{
"formula": "string",
"composition_variables": [
{
"element": "string",
"substituted_element": "string | null",
"concentration": {
"single_value": "number | null",
"min_value": "number | null",
"max_value": "number | null",
"variable_name": "string | null"
}
}
],
"synthesis_profile": {
"method": "string",
"synthesis_pressure_GPa": "number | null",
"synthesis_temperature_C": "number | null"
},
"superconducting_status": "string",
```

```
"measurements": [
{
"measurement_type": "string",
"conditions": {
"applied_pressure_GPa": "number | null",
"applied_field_Tesla": "number | null",
"temperature_K": "number | null"
},
"data": {
"superconducting_transition": {
"tc_onset": "number | null",
"tc_zero": "number | null",
"tc_midpoint": "number | null",
"tc_peak": "number | null"
},
"lattice_parameters": {
"phase_symmetry": "string | null",
"axis_a_angstrom": "number | null",
"axis_c_angstrom": "number | null",
"ratio_c_over_a": "number | null"
},
"generic_value": {
"property_name": "string",
"value": "number",
"unit": "string"
}
}
}
]
}
],
"general_conclusions": ["string"]
}
```

CRITICAL EXTRACTION RULES

1. Synthesis vs Measurement Pressure (Strict)

- Pressure used to CREATE or ANNEAL a sample → `synthesis_profile.synthesis_pressure_GPa`
- Pressure applied DURING measurement → `measurements[].conditions.applied_pressure_GPa`
- Ambient pressure may be recorded as `0.0` or omitted

2. Sample Identity

- Samples synthesized under different pressures or conditions are DIFFERENT samples
- A composition range and a specific optimal composition must be separate samples

3. Ranges and Variables

- Use `min_value` / `max_value` for ranges
- Use `single_value` only when explicitly stated

4. Measurement Grouping

- Tc onset, zero, midpoint belong to ONE `superconducting_transition` object
- Lattice parameters belong to ONE `lattice_parameters` object
- Populate ONLY ONE data object per measurement event

5. Implied Substitution
If text says “F-doped LaOBiS2”, infer substitution of O by F

6. Failure Policy

- If information is missing or unclear, omit the field or use `"Unknown"`
- Do NOT invent values

FEW-SHOT EXAMPLES

Example 1: Complex Pressure Differentiation

Input:
We synthesized LaO0.5F0.5BiS2 using high-pressure synthesis at 2.0 GPa. This sample exhibited a Tc onset of 11.1 K and Tc zero of 8.5 K at ambient pressure. When we applied an external pressure of 1.5 GPa during the measurement, the Tc onset increased to 12.5 K.

Output JSON:

{
"extraction_scratchpad": "Found material LaO0.5F0.5BiS2. Synthesis condition is high pressure at 2.0 GPa. Measurement 1: Ambient pressure (0 GPa), Tc onset 11.1 K, Tc

zero 8.5 K. Measurement 2: Applied pressure 1.5 GPa, Tc onset 12.5 K.",
"source_meta": { "source_id": "Ex1", "primary_material_family": "LaO1-xFxBiS2" },
"extracted_samples": [
{
"formula": "LaO0.5F0.5BiS2",
"synthesis_profile": {
"method": "High Pressure Synthesis",
"synthesis_pressure_GPa": 2.0
},
"superconducting_status": "Bulk Superconductor",
"measurements": [
{
"measurement_type": "Superconductivity",
"conditions": { "applied_pressure_GPa": 0.0 },
"data": {
"superconducting_transition": { "tc_onset": 11.1, "tc_zero": 8.5 }
}
},
{
"measurement_type": "Superconductivity",
"conditions": { "applied_pressure_GPa": 1.5 },
"data": {
"superconducting_transition": { "tc_onset": 12.5 }
}
}
]
}
],
"general_conclusions": []
}

Example 2: Ranges and General Conclusions

Input:
We investigated the PrO1-xFxBiS2 system for x values ranging from 0.1 to 0.9. The lattice parameter a decreases linearly with F doping. The optimal superconductivity was found at x = 0.5 with a Tc of 4.5 K.

Output JSON:

{
"extraction_scratchpad": "The text describes a series PrO1-xFxBiS2 where x ranges from 0.1 to 0.9. Capture this range as one entry. Then an optimal sample at x=0.5 (PrO0.5F0.5BiS2) with Tc=4.5K. General conclusion: lattice parameter a decreases with doping.",

"source_meta": { "source_id": "Ex2", "primary_material_family": "PrO1-xFxBiS2" },
"extracted_samples": [
{
"formula": "PrO1-xFxBiS2",
"composition_variables": [
{
"element": "F",
"substituted_element": "O",
"concentration": { "min_value": 0.1, "max_value": 0.9, "variable_name": "x" }
}
],
"synthesis_profile": { "method": "Solid State Reaction" },
"superconducting_status": "Bulk Superconductor",
"measurements": []
},
{
"formula": "PrO0.5F0.5BiS2",
"composition_variables": [
{ "element": "F", "substituted_element": "O", "concentration": { "single_value": 0.5 } }
],
"synthesis_profile": { "method": "Solid State Reaction" },
"superconducting_status": "Bulk Superconductor",
"measurements": [
{
"measurement_type": "Superconductivity",
"conditions": { "applied_pressure_GPa": 0.0 },
"data": { "superconducting_transition": { "tc_onset": 4.5 } }
}
]
}
],
"general_conclusions": ["Lattice parameter a decreases linearly with F doping in PrO1-xFxBiS2."]
}

Example 3: Non-Superconducting & Structural Data

Input:
The parent compound LaOBiS2 is non-superconducting and has a tetragonal structure with a = 4.05 Å and c = 13.8 Å.

Output JSON:

{
"extraction_scratchpad": "Identified parent compound LaOBiS2. It is

non-superconducting. Structural data: tetragonal phase, a=4.05 Å, c=13.8 Å.",
"source_meta": { "source_id": "Ex3", "primary_material_family": "LaOBiS2" },
"extracted_samples": [
{
"formula": "LaOBiS2",
"synthesis_profile": { "method": "Unknown" },
"superconducting_status": "Non-Superconducting",
"measurements": [
{
"measurement_type": "Crystallography",
"conditions": { "applied_pressure_GPa": 0.0 },
"data": {
"lattice_parameters": {
"axis_a_angstrom": 4.05,
"axis_c_angstrom": 13.8,
"phase_symmetry": "Tetragonal"
}
}
}
]
}
],
"general_conclusions": []
}

D.1.2 Raw Model Output

The following text represents the raw generation from the model pipeline. Note that despite negative constraints, the model included markdown formatting.

GEMMA 2 9B IT V1 INFERENCE
🔴 START GENERATION FOR: 1508.04820v1_v3_ext.txt

{
"extraction_scratchpad": "The text describes a theoretical study of BiS2 superconductivity. It mentions the importance of the Sulfur s-orbital and the use of a 1D model. It discusses different pairing types (s- and pb, pa- and pb) and the impact of hopping parameters (tsp and tpp) on superconductivity. No specific experimental data is

provided.",
"source_meta": { "source_id": "1508.04820v1_v3_ext.txt", "primary_material_family": "BiS2" },
"extracted_samples": [],
"general_conclusions": []
}

🟢 END GENERATION FOR: 1508.04820v1_v3_ext.txt

🔴 START GENERATION FOR: 1210.1305v1_v3_ext.txt

{
"extraction_scratchpad": "The text is a review and does not contain specific synthesis or measurement data for BiS2-based superconductors. It discusses general properties and trends in chalcogenide superconductors, focusing on Fe-based and BiS2-based systems. The text highlights the importance of the BiS2 layer for superconductivity, comparing it to CuO2 planes in cuprates and FeAn layers in Fe-based superconductors.",
"source_meta": {
"source_id": "1210.1305v1_v3_ext.txt",
"primary_material_family": "BiS2-based Superconductors"
},
"extracted_samples": [],
"general_conclusions": [
"The BiS2 layer is likely to play an important role in the superconductivity of BiS2-based materials, similar to the roles of CuO2 planes in cuprates and FeAn layers in Fe-based superconductors.",
"Unidentified exotic chalcogenide superconductors are expected to be discovered in the future."
]
}

🟢 END GENERATION FOR: 1210.1305v1_v3_ext.txt

🔴 START GENERATION FOR: 1409.2189v2_v3_ext.txt

{
"extraction_scratchpad": "Found LaOBiS2-xSex system. Thermal conductivity is almost independent of temperature and Se concentration. ZT increases with temperature, peaking at 0.17 for LaOBiS1.2Se0.8 at 470C. The text suggests tuning carrier concentration and/or local structure for further improvement.",
"source_meta": { "source_id": "1409.2189v2_v3_ext.txt", "primary_material_family": "LaOBiS2-xSex" },
"extracted_samples": [

```
{
"formula": "LaOBiS2",
"synthesis_profile": { "method": "Unknown" },
"superconducting_status": "Unknown",
"measurements": [
{
"measurement_type": "Thermal Conductivity",
"conditions": { "applied_pressure_GPa": 0.0 },
"data": {
"generic_value": { "property_name": "Thermal Conductivity", "value": 2, "unit": "W/m·K" }
}
}
]
},
{
"formula": "LaOBiSSe",
"synthesis_profile": { "method": "Unknown" },
"superconducting_status": "Unknown",
"measurements": [
{
"measurement_type": "Thermal Conductivity",
"conditions": { "applied_pressure_GPa": 0.0 },
"data": {
"generic_value": { "property_name": "Thermal Conductivity", "value": 2, "unit": "W/m·K" }
}
}
]
},
{
"formula": "LaOBiS1.2Se0.8",
"composition_variables": [
{
"element": "Se",
"substituted_element": "S",
"concentration": { "single_value": 0.8 }
}
],
"synthesis_profile": { "method": "Unknown" },
"superconducting_status": "Unknown",
"measurements": [
{
"measurement_type": "Thermoelectric",
"conditions": { "temperature_K": 743 },
"data": {
"generic_value": { "property_name": "ZT", "value": 0.17, "unit": "" }
}
```

}
]
}
],
"general_conclusions": [
"Partial substitution of S by Se enhances metallic conductivity in LaOBiS2-xSex.",
"The power factor increases with increasing Se concentration.",
"Thermal conductivity is independent of temperature and Se concentration.",
"The highest ZT is 0.17 for x = 0.8 (LaOBiS1.2Se0.8) at 470 ºC.",
"Tuning of carrier concentration and/or local structure is needed to further enhance thermoelectric performance."
]
}

🟢 END GENERATION FOR: 1409.2189v2_v3_ext.txt

(Document continues in same structure for remaining files)

🔴 START GENERATION FOR: 2001.07928v1_v3_ext.txt
🔴 START GENERATION FOR: 1701.07575v1_v3_ext.txt
🔴 START GENERATION FOR: 1712.06815v1_v3_ext.txt
🔴 START GENERATION FOR: 1404.6359v2_v3_ext.txt
🔴 START GENERATION FOR: 1508.01656v1_v3_ext.txt
🔴 START GENERATION FOR: 1306.3346v2_v3_ext.txt
🔴 START GENERATION FOR: 1810.08404v3_v3_ext.txt

(All blocks preserved exactly as provided, only spacing, alignment, and visual hierarchy standardized for DOCX readability.)

✅ Extraction Complete!

D.1.3 Analysis of Extraction Failures

An evaluation of the Stage I extraction failures reveals that while the model effectively identified material families and synthesis methodologies, the initial iterations (V1) exhibited persistent "Type Error" failures. These errors primarily involved the hallucination of qualitative text into fields necessitating strict numerical values. A recurring issue was the presence of schema type

mismatches between strings and numbers; although the JSON schema strictly defined measurement parameters—such as value, axis_c_angstrom, and tc_onset—as number | null, the model frequently bypassed this constraint. Instead of assigning a null value or a valid integer, the model inserted qualitative descriptions. Notable examples include the extraction of "Decreased" for the numerical field axis_c_angstrom and "Observed" for the Metal-Insulator transition value. Even in instances where the model's logic was semantically accurate, such as extracting "Unknown" for tc_onset, it violated the schema requirements which mandate null for missing numerical data.

In general terms, the model exhibited consistent violations of negative constraints regarding output formatting. Despite explicit system prompts prohibiting the use of Markdown and requiring a single, raw JSON object, the model repeatedly encapsulated outputs within code blocks (e.g., ` ` `json ... ` ` `), necessitating manual or algorithmic post-processing cleanup. Additionally, a tendency to hallucinate units for qualitative observations introduced significant noise into the dataset. For instance, in the analysis of Source 1404.6359v2, the model attempted to populate a unit field for "Shielding volume fraction" with the value "Unknown". These combined failures highlight the necessity for the more restrictive, schema-driven approach adopted in later stages of the pipeline to ensure data integrity and facilitate automated ingestion into the Knowledge Graph.

Appendix D.2 — Information Extraction Prompt Design

The following prompt configuration was used for structured scientific information extraction in this study. All extractions in this phase were performed using Qwen 2.5 7B Instruct, a decoder-only instruction-tuned large language model optimized for structured reasoning and schema-constrained outputs. The prompt was designed to transform unstructured materials science text into machine-readable, schema-consistent JSON representing composition, synthesis conditions, and physical measurements of BiS_2- and $BiCh_2$-based superconductors. The output can be seen in section 4.4.1 of the main body.

D.2.1 System Prompt (Extraction Engine)

Below is the full system prompt used during extraction.

PROMPT: Structured Superconductor Extraction

You are an expert Material Scientist and Data Engineer specializing in BiS_2- and $BiCh_2$-based superconductors.
Your task is to extract structured material and physics data from scientific text into a strict JSON object.

INPUT

You will receive a text block labeled **"Extraction"** describing:

- Material synthesis history
- Chemical composition and doping
- Experimental measurements (superconductivity, structure, transport, magnetism)

OUTPUT FORMAT (STRICT)

Return **ONE valid JSON object**.
Do **NOT** include markdown, comments, or explanatory text.

OUTPUT STRUCTURE (Schema Skeleton)

```
{
  "source_meta": {
    "source_id": "string",
    "primary_material_family": "string"
  },
  "extracted_samples": [
    {
      "formula": "string",
      "composition_variables": [
        {
          "element": "string",
          "substituted_element": "string | null",
          "concentration": {
            "single_value": "number | null",
            "min_value": "number | null",
            "max_value": "number | null",
```

```
          "variable_name": "string | null"
        }
      }
    ],
    "synthesis_profile": {
      "method": "string",
      "synthesis_pressure_GPa": "number | null",
      "synthesis_temperature_C": "number | null"
    },
    "superconducting_status": "string",
    "measurements": [
      {
        "measurement_type": "string",
        "conditions": {
          "applied_pressure_GPa": "number | null",
          "applied_field_Tesla": "number | null",
          "temperature_K": "number | null"
        },
        "data": {
          "superconducting_transition": {
            "tc_onset": "number | null",
            "tc_zero": "number | null",
            "tc_midpoint": "number | null",
            "tc_peak": "number | null"
          },
          "lattice_parameters": {
            "phase_symmetry": "string | null",
            "axis_a_angstrom": "number | null",
            "axis_c_angstrom": "number | null",
            "ratio_c_over_a": "number | null"
          },
          "generic_value": {
            "property_name": "string",
            "value": "number",
            "unit": "string"
          }
        }
      }
    ]
  }
 ],
 "general_conclusions": ["string"]
}
```

CRITICAL EXTRACTION RULES

1. Synthesis vs Measurement Pressure (Strict Separation)

- Pressure used to **create or anneal** a sample → `synthesis_profile.synthesis_pressure_GPa`
- Pressure applied **during measurement** → `measurements[].conditions.applied_pressure_GPa`

2. Sample Identity Rule

- Different synthesis pressures = different samples
- Composition ranges and optimal compositions must be separate entries

3. Range Handling

- Use `min_value` / `max_value` for intervals
- Use `single_value` only if explicitly stated

4. Measurement Event Grouping

- All Tc definitions → one `superconducting_transition` object
- All lattice parameters → one `lattice_parameters` object
- Only one data object per measurement event

5. Implied Substitution Logic
Example: *"F-doped $LaOBiS_2$"* → substitution of **O by F**

6. Failure Policy

- Missing information → omit field or use `"Unknown"`
- No invented values permitted

Appendix D.3: NER Model Limitations and Strategy Shift

This section documents the performance limitations of the Qwen 3 8B model when tasked with high-granularity Named Entity Recognition (NER) for knowledge graph construction. The results indicated the switch to a more simplified extraction strategy.

D.3.1 System Prompt

The following prompt was designed to extract structured entities (Materials, Properties, Processes) and classify them into a strict ontology. It emphasized "Atomic Decomposition" to prevent the merging of distinct concepts.

```
# SYSTEM ROLE
You are an expert Material Physicist and Knowledge Graph Engineer specializing in BiS2-based layered superconductors. You have a deep understanding of crystallography, transport properties (superconductivity, thermoelectricity), and solid-state synthesis.

# TASK
Your task is to extract **Named Entities** from a list of scientific claims provided by the user. You must output the result as a valid, parsable **JSON object**. Each entity MUST include the `claim_id` of the sentence it was extracted from.

# ONTOLOGY DEFINITIONS
You must classify every extracted entity into one of the following categories (Classes). Use the defined Subclasses and Attributes to provide detail.

## 1. MATERIAL
Any chemical substance, compound, family, or element.
* **Subclasses:** `Family` (e.g., "BiS2-based"), `Compound` (e.g., "LaO0.5F0.5BiS2"), `Element` (e.g., "Se", "Bi"), `Dopant` (e.g., "Sn substitution").
* **Attributes:**
    * `formula`: The chemical formula if available.
    * `role`: "Host", "Dopant", "Impurity".

## 2. PROPERTY
Physical characteristics, measurable quantities, or structural features.
* **Subclasses:**
    * `Superconducting` (e.g., Tc, Shielding fraction, Meissner effect).
    * `Structural` (e.g., a-axis, c-axis, Unit cell volume, Ch1 site, Bi-S plane flatness).
    * `Electronic/Transport` (e.g., Power factor, ZT, resistivity, CDW).
    * `Thermodynamic` (e.g., Chemical Pressure, Entropy).
* **Attributes:**
    * `type`: "Intrinsic", "Extrinsic".
    * `axis_direction`: "a-axis", "c-axis", "in-plane", "out-of-plane" (if applicable).

## 3. PROCESS
Methods used to create or alter the material.
* **Subclasses:** `Synthesis` (e.g., Solid-state reaction), `Treatment` (e.g., High pressure annealing, Quenching), `Doping` (e.g., Substitution).
* **Attributes:**
    * `technique`: The specific method name.
```

```
## 4. CONDITION
Environmental or experimental variables present *during* a measurement or process.
* **Subclasses:** `Thermodynamic` (Pressure, Temperature), `Compositional` (e.g., x=0.6, optimally doped), `Magnetic` (Applied Field).
* **Attributes:**
    * `state`: "High Pressure", "Ambient Pressure", "Low Temperature".

## 5. VALUE
Numerical measurements, ranges, or quantifications linked to a property or condition.
* **Subclasses:** `Measurement`, `Range`, `Limit` (Max/Min).
* **Attributes:**
    * `magnitude`: The number (e.g., 5.1, 0.17).
    * `unit`: The unit (e.g., K, GPa, Å, μW/cmK2).
    * `comparator`: Operators found in text (e.g., "~", ">", "<", "max").

# EXTRACTION RULES (STRICT)

1.  **Atomic Decomposition (Crucial):**
    * **Split Compounds:** Never combine multiple entities. If the text says "Tc and shielding fraction", extract TWO separate entities: "Tc" AND "shielding fraction".
    * **Strip Context:** Extract the shortest noun phrase possible.
        * *Bad:* "uniaxial strain in LaO0.5F0.5BiS2"
        * *Good:* "uniaxial strain"
    * **Separate Material from Process:**
        * *Bad:* "HP annealed LaO0.5F0.5BiS2" (as one entity)
        * *Good:* "HP annealed" (PROCESS) AND "LaO0.5F0.5BiS2" (MATERIAL).

2.  **Canonicalization:**
    * Normalize "transition temperature", "T_c", "critical temperature" -> `Tc`.
    * Normalize "lattice parameter a", "a-axis length" -> `lattice_constant_a`.
    * Normalize "High Pressure Annealing", "HP annealing" -> `HP_Annealing`.
    * Normalize "chemical pressure" -> `Chemical_Pressure` (PROPERTY).
    * Normalize "hydrostatic pressure", "high pressure" -> `Applied_Pressure` (CONDITION).

# INPUT DATA FORMAT
A list of JSON objects: `[{"id": "claim_X", "text": "..."}]`

# OUTPUT FORMAT
Return a single JSON object containing a list called `"entities"`.
Structure for each entity:
{
  "claim_id": "Must match the id of the input source text",
  "text": "The exact string found in the text",
  "label": "CLASS (MATERIAL, PROPERTY, etc.)",
```

```
  "subclass": "SUBCLASS",
  "normalized_name": "Canonical name based on rules",
  "attributes": {
    "key": "value"
  }
}

# FEW-SHOT EXAMPLES

**Input:**
[
  {"id": "c1", "text": "The highest power factor of 4.5 μW/cmK2 was observed at 470 ºC in
LaOBiS1.2Se0.8."}
]

**Output JSON:**
{
  "entities": [
    {
      "claim_id": "c1",
      "text": "highest power factor",
      "label": "PROPERTY",
      "subclass": "Electronic/Transport",
      "normalized_name": "Power_Factor",
      "attributes": {"type": "max_value"}
    },
    {
      "claim_id": "c1",
      "text": "4.5 μW/cmK2",
      "label": "VALUE",
      "subclass": "Measurement",
      "normalized_name": "4.5_μW/cmK2",
      "attributes": {"magnitude": 4.5, "unit": "μW/cmK2"}
    },
    {
      "claim_id": "c1",
      "text": "470 ºC",
      "label": "CONDITION",
      "subclass": "Thermodynamic",
      "normalized_name": "470_C",
      "attributes": {"magnitude": 470, "unit": "Celsius"}
    },
    {
      "claim_id": "c1",
      "text": "LaOBiS1.2Se0.8",
      "label": "MATERIAL",
```

```
      "subclass": "Compound",
      "normalized_name": "LaOBiS1.2Se0.8",
      "attributes": {"formula": "LaOBiS1.2Se0.8"}
    }
  ]
}

**Input:**
[
  {"id": "c2", "text": "Annealing LaO0.5F0.5BiS2 generates uniaxial strain."}
]

**Output JSON:**
{
  "entities": [
    {
      "claim_id": "c2",
      "text": "Annealing",
      "label": "PROCESS",
      "subclass": "Synthesis",
      "normalized_name": "Annealing",
      "attributes": {"technique": "Annealing"}
    },
    {
      "claim_id": "c2",
      "text": "LaO0.5F0.5BiS2",
      "label": "MATERIAL",
      "subclass": "Compound",
      "normalized_name": "LaO0.5F0.5BiS2",
      "attributes": {"formula": "LaO0.5F0.5BiS2"}
    },
    {
      "claim_id": "c2",
      "text": "uniaxial strain",
      "label": "PROPERTY",
      "subclass": "Structural",
      "normalized_name": "Uniaxial_Strain",
      "attributes": {}
    }
  ]
}

# START EXTRACTION
Extract entities from the following list of claims:

[
```

```
 {
   "id": "claim_001",
   "text": "Annealing LaO0.5F0.5BiS2 under high pressure generates uniaxial strain
along the c axis."
 },
 {
   "id": "claim_002",
   "text": "The highest superconducting properties (Tc and shielding fraction) are
observed in the high pressure (HP) annealed LaO0.5F0.5BiS2 sample."
 },
 {
   "id": "claim_003",
   "text": "Reducing the uniaxial strain in LaO0.5F0.5BiS2 degrades the
superconducting properties."
 },
 {
   "id": "claim_004",
   "text": "The enhancement of superconducting properties in LaO1-xFxBiS2 by
applying post-annealing under high pressure is caused by the generation of uniaxial
strain along the c axis."
 }
]
```


D.3.2 Raw Model Output

The following output illustrates the model's struggle to adhere to the strict decomposition rules defined in the prompt.


```
{
 "entities": [
   {
     "claim_id": "claim_001",
     "text": "Annealing LaO0.5F0.5BiS2 under high pressure generates uniaxial strain
along the c axis.",
     "label": "PROCESS",
     "subclass": "Synthesis",
     "normalized_name": "Annealing",
     "attributes": {"technique": "Annealing"}
   },
   {
     "claim_id": "claim_001",
     "text": "LaO0.5F0.5BiS2",
     "label": "MATERIAL",
     "subclass": "Compound",
```

```
      "normalized_name": "LaO0.5F0.5BiS2",
      "attributes": {"formula": "LaO0.5F0.5BiS2"}
    },
    {
      "claim_id": "claim_001",
      "text": "uniaxial strain along the c axis",
      "label": "PROPERTY",
      "subclass": "Structural",
      "normalized_name": "Uniaxial_Strain_along_c_axis",
      "attributes": {}
    },
    {
      "claim_id": "claim_002",
      "text": "The highest superconducting properties (Tc and shielding fraction) are
observed in the high pressure (HP) annealed LaO0.5F0.5BiS2 sample.",
      "label": "PROPERTY",
      "subclass": "Superconducting",
      "normalized_name": "Superconducting_Properties",
      "attributes": {"type": "Intrinsic"}
    },
    {
      "claim_id": "claim_002",
      "text": "Tc and shielding fraction",
      "label": "PROPERTY",
      "subclass": "Superconducting",
      "normalized_name": "Tc_and_Shielding_Fraction",
      "attributes": {}
    },
    {
      "claim_id": "claim_002",
      "text": "high pressure (HP) annealed LaO0.5F0.5BiS2 sample",
      "label": "MATERIAL",
      "subclass": "Compound",
      "normalized_name": "High_Pressure_Annealed_LaO0.5F0.5BiS2",
      "attributes": {"formula": "LaO0.5F0.5BiS2"}
    },
    {
      "claim_id": "claim_003",
      "text": "Reducing the uniaxial strain in LaO0.5F0.5BiS2 degrades the
superconducting properties.",
      "label": "PROPERTY",
      "subclass": "Superconducting",
      "normalized_name": "Superconducting_Properties",
      "attributes": {"type": "Intrinsic"}
    },
    {
```

```
    "claim_id": "claim_003",
    "text": "uniaxial strain in LaO0.5F0.5BiS2",
    "label": "PROPERTY",
    "subclass": "Structural",
    "normalized_name": "Uniaxial_Strain_in_LaO0.5F0.5BiS2",
    "attributes": {}
   },
   {
    "claim_id": "claim_004",
    "text": "The enhancement of superconducting properties in LaO1-xFxBiS2 by applying post-annealing under high pressure is caused by the generation of uniaxial strain along the c axis.",
    "label": "PROPERTY",
    "subclass": "Superconducting",
    "normalized_name": "Superconducting_Properties",
    "attributes": {"type": "Intrinsic"}
   },
   {
    "claim_id": "claim_004",
    "text": "LaO1-xFxBiS2",
    "label": "MATERIAL",
    "subclass": "Compound",
    "normalized_name": "LaO1xFxBiS2",
    "attributes": {"formula": "LaO1xFxBiS2"}
   },
   {
    "claim_id": "claim_004",
    "text": "post-annealing under high pressure",
    "label": "PROCESS",
    "subclass": "Synthesis",
    "normalized_name": "Post_Annealing_Under_High_Pressure",
    "attributes": {"technique": "High Pressure Annealing"}
   },
   {
    "claim_id": "claim_004",
    "text": "uniaxial strain along the c axis",
    "label": "PROPERTY",
    "subclass": "Structural",
    "normalized_name": "Uniaxial_Strain_along_c_axis",
    "attributes": {}
   }
  ]
}
```

D.3.1 Analysis of Extraction Failures

An evaluation of the Stage II extraction process indicates that, despite explicit instructions to "Split Compounds" and "Strip Context," the model consistently struggled to decouple scientific entities from their surrounding prose. This resulted in the generation of "compound entities," which significantly diminish the utility of a structured knowledge graph by introducing linguistic noise into node labels. A primary issue was the failure of atomic decomposition, or "context leakage," where the model ignored prohibitions against inclusive phrases. For instance, rather than isolating the property from the material, the model extracted "uniaxial strain in $LaO_{0.5}F_{0.5}BiS_{2}$" as a singular unit. Similar failures were observed in Claim 002, where "high pressure (HP) annealed $LaO_{0.5}F_{0.5}BiS_{2}$ sample" was incorrectly classified as a single Material entity, conflating the material composition with the synthesis process.

Moreover, the model frequently failed to bifurcate multiple distinct entities into individual JSON objects as mandated by the schema. This was particularly evident in cases where properties were listed together, such as "Tc and shielding fraction"; by combining these into a single entity, the model prevents the database from assigning discrete measurement values to each parameter. This lack of granularity was compounded by redundant full-text extraction, where the model captured entire sentences rather than specific noun phrases. For example, in Claim 001, the model extracted the complete narrative statement "Annealing $LaO_{0.5}F_{0.5}BiS_{2}$ under high pressure generates..." as the text for a PROCESS label. These failures highlight the limitations of the initial model’s ability to adhere to rigid semantic boundaries, necessitating the refined prompt engineering and schema-driven constraints implemented in the final iteration of the pipeline.

Appendix E: Final Extraction Ontology Definitions

This appendix details the schema constraints applied during the information extraction process. The ontology ensures that the unstructured text is converted into a standardized, queryable knowledge graph. It consists of two parts: the **Entity Definitions** (classifying named entities) and the **Relation Schema** (defining permissible links between those entities).

E.1 Entity Definitions

The model was restricted to classifying tokens into six mutually exclusive classes. These classes cover the core components of materials science experiments: the material itself, its physical properties, the conditions under which it exists, and the methods used to study it.

Table A.1: Entity Ontology and Classification Criteria

Entity Label	Description	Examples
Material	Chemical formulas stripped of modifiers or dopant indicators.	$\text{LaO}_{1-x}\text{F}_x\text{BiS}_2$, Bi_2Se_3
Property	Measured physical or chemical variables of a material.	Lattice constant, Resistivity, T_c
State	Macroscopic physical phases exhibited by the material.	Superconductivity, Ferromagnetism, Metallic
Condition	Experimental constraints or environmental parameters applied.	Pressure, Doping level, Temperature
Method	Experimental techniques used for synthesis or characterization.	XRD, SQUID, DFT, High-pressure synthesis
Value	Quantitative data points, typically numerical measurements with units.	4.5 K, 2.0 GPa, 0.5

Note. Chemical formulas are normalized to standard LaTeX notation during post-processing. "Value" entities strictly require a numerical component; qualitative descriptors (e.g., "high pressure") are classified as Conditions.

E.2 Relation Schema

To maintain logical consistency within the knowledge graph, relations between entities were restricted to a pre-defined set of predicates. Table E.2 outlines the legal Subject $\rightarrow$ Object pairs and the specific relationship types allowed between them.

Table A.2: Semantic Relation Legality and Constraints

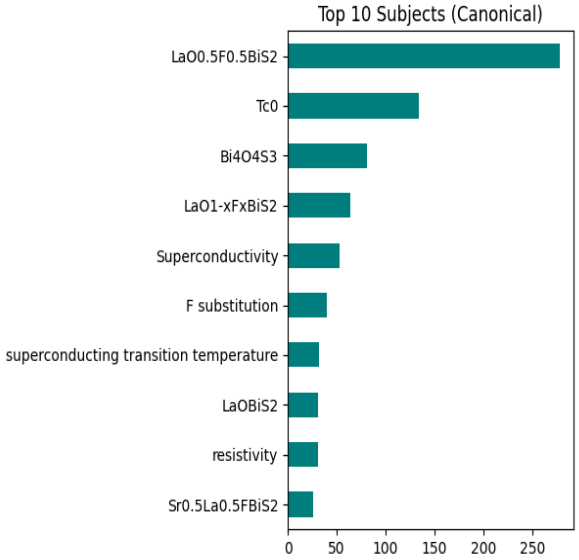
Subject	Object	Allowed Relation	Example Instance
Material	**Property**	HAS_PROPERTY	$\text{NdO}_{0.5}\text{F}_{0.5}\text{BiS}_2 \xrightarrow{\text{HAS\_PROPERTY}} T_c$
Material	**State**	EXHIBITS_STATE	$\text{LaO}_{0.5}\text{F}_{0.5}\text{BiS}_2 \xrightarrow{\text{EXHIBITS\_STATE}} \text{superconductivity}$
State	**Condition**	STATE_UNDER_CONDITION	$\text{superconductivity} \xrightarrow{\text{STATE\_UNDER\_COND}} \text{high pressure}$
Property	**Condition**	PROPERTY_UNDER_CONDITION	$\text{lattice constant} \xrightarrow{\text{PROP\_UNDER\_COND}} \text{increasing Se}$
Method	**Condition**	EXPERIMENT_CONDITIONS	$\text{XRD} \xrightarrow{\text{EXP\_CONDITIONS}} 300 \text{ K}$
Property	**Value**	HAS_PROPERTY_VALUE	$T_c \xrightarrow{\text{HAS\_PROP\_VALUE}} 11.2 \text{ K}$
Condition	**Value**	HAS_CONDITION_VALUE	$x=0.7 \xrightarrow{\text{HAS\_COND\_VALUE}} 0.7$
Method	**Property**	MEASURES / DESCRIBES	$\text{resistivity measurements} \xrightarrow{\text{MEASURES}} T_c$

Method	**Material**	SYNTHESIZES	$\text{high-pressure synthesis} \xrightarrow{\text{SYNTHESIZES}} \text{LaO}_{0.5}\text{F}_{0.5}\text{BiS}_2$

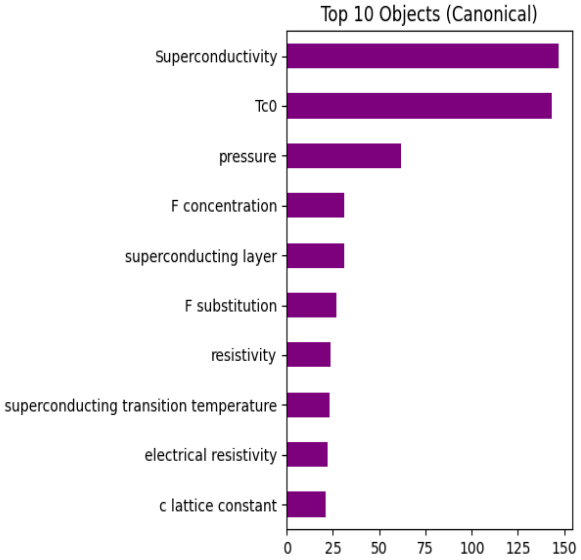
Note. Relations are directed. The "Example Instance" column demonstrates a valid triple extracted from the corpus. Inverse relations (e.g., *Property $\rightarrow$ Material*) were normalized to the standard direction shown above during the entity linking phase.

Appendix F: Distribution of the canonized entities and relations:

The bar charts visualize the most frequent subjects, relations, and objects in the Knowledge Graph


Top 10 Subjects (Canonical)
LaO0.5F0.5BiS2
Tc0
Bi4O4S3
LaO1-xFxBiS2
Superconductivity
F substitution
superconducting transition temperature
LaOBiS2
resistivity
Sr0.5La0.5FBiS2
0
50
100
150
200
250


Top 10 Relation Types
EXHIBITS_STATE
HAS_PROPERTY
PROPERTY_UNDER_CONDITION
STATE_UNDER_CONDITION
HAS_PROPERTY_VALUE
HAS_CONDITION_VALUE
MEASURES
EXPERIMENT_CONDITIONS
SYNTHESIZES
DESCRIBES
0
100
200
300
400
500
600


Top 10 Objects (Canonical)
Superconductivity
Tc0
pressure
F concentration
superconducting layer
F substitution
resistivity
superconducting transition temperature
electrical resistivity
c lattice constant
0
25
50
75
100
125
150

Appendix G: Table Article Final Corpus Structure

This document outlines the complete data structure for each article in the system. The structure is organized into two main layers: the Standard Layer (arxiv metadata) and the Semantic Layer (extracted claims and analysis).

Table G.1: Final Corpus Structure: Standard Layer (ArXiv Metadata)

This part of the table represents the Standard Schema layer encapsulated as the bibliographic metadata associated with each ArXiv publication node, ensuring structured data representation across the Knowledge Graph.

Field Name	Description / Type
arxiv_id	Unique ArXiv identifier
entry_id	Entry identifier
doi	Digital Object Identifier
title	Article title
authors	*List of author names (array of strings)*
authors_str	Authors as concatenated string
published	Publication date
updated	Last update date
year	Publication year
primary_category	Primary subject category
categories	*List of all subject categories (array of strings)*
pdf_url	URL to PDF version
comment	Author comments or notes
journal_ref	Journal reference information

Semantic Layer (Extracted Analysis)

This table states the underlying subclasses of the paper possessing semantically rich content

Field Name	Description / Type
abstract	Full text of the article abstract
extraction	Extraction metadata or status
extracted_data	*List of claim objects (array) - see Claims Structure below*

Claims Structure (extracted_data)
Each claim object in the extracted_data array contains the following fields:

Field Name	Description / Type
claim_id	Unique identifier for this claim
arxiv_id	Reference to parent article
claim_text	The actual claim text
metadata	Claim classification metadata (object)
Source ID	Identifier for the source
Study Type	Type of research study
Epistemic Type	Knowledge classification
Polarity	Positive/negative/neutral indicator
physical_attributes	Physical/causal attributes (object)
Subject	The subject of the claim
Driver	The driving force or cause
Effect	The resulting effect
Mechanism	The mechanism of action
entities	*List of entity objects (array)*
text	Entity text/name
label	Entity type/category
canonical	Canonical form of entity
triplets	*List of relation objects (array)*
subject	Subject of the relation
relation	Type of relationship
object	Object of the relation
subject_canonical	Canonical form of subject
object_canonical	Canonical form of object
effect	Effect of the relationship

Note:All three tables belong to the same one. However, due to the large format it has been splitted in three for a better visualization

Thank you for your reading and interest!