# LifeTabFusion: A Confidence-Guided Table Understanding via Hybrid Integration of KGs, ML, and LLMs

Vishvapalsinhji Parmar[1], Alsayed Algergawy[1]

[1]Chair of Data and Knowledge Engineering, University of Passau, Passau, Germany

**Abstract**

With the rapidly increasing volume of tabular data, there is a growing presence of semantic and structural heterogeneities, which presents significant challenges for effective table understanding in domains such as life sciences and biomedicine. Despite their structured appearance, tables often lack explicit semantics, making tasks like data integration, search, and knowledge graph construction challenging. In this paper, we introduce LifeTabFusion, a modular and forward-looking framework designed to support robust and scalable table understanding. The framework integrates three core components we have previously developed and evaluated individually on benchmark datasets: (i) domain-sensitive preprocessing for anomaly handling and normalization, (ii) lightweight machine learning models for schema-specific annotation, and (iii) scalable knowledge graph annotation via API-based lookups. Each module has demonstrated effectiveness across tasks such as Cell Entity Annotation (CEA), Column Type Annotation (CTA), and Column Property Annotation (CPA) using datasets from the SemTab challenge. Building on these foundations, LifeTabFusion proposes a hybrid architecture that selectively incorporates Large Language Models (LLMs) for contextual disambiguation and semantic enrichment. The final annotations are derived through a confidence-based fusion strategy that leverages the strengths of each component while minimizing individual weaknesses.

**Keywords**

Semantic Table Understanding, Cell Entity Annotation, Column Type Annotation, Column Property Annotation, Knowledge Graph Matching, Tabular Data Annotation

## 1. Introduction

Tables are among the most widely used formats for representing structured information, with applications ranging from scientific publications and spreadsheets to open government data and biomedical records. Their use dates back millennia, with one of the earliest known examples being a Sumerian clay tablet from the ancient city of Shuruppag (ca. 2600 BCE), organized in a tabular format to record data[1]. This early instance underscores the enduring importance of tabular representation in human knowledge preservation .

In the digital age, the use of tabular data has grown exponentially. By the end of 2028, the global volume of data is projected to reach approximately 394 zettabytes[1], a significant portion of which is expected to be structured in tabular form. Despite their readability and efficiency for human users, tables are often semantically ambiguous and lack the contextual information required for machines to process them reliably. Variability in domain-specific content, structure, language, and formatting further complicates their automated interpretation.

Table understanding aims to address this gap. It refers to the automatic interpretation of a table's structure, content, and semantics. This includes tasks like detecting tables in documents, identifying their functional elements (e.g., headers, stubs), and performing semantic interpretation, which involves linking cells to entities, classifying column types, and identifying relationships between columns [2]. These tasks form the backbone of a growing research field aiming to transform raw tabular data into semantically rich knowledge representations. To advance research in this area, the SemTab challenge[2]

*Posters & Demos Track, co-located with SEMANTiCS'25: International Conference on Semantic Systems, September 3–5, 2025, Vienna, Austria*

vishvapalsinhji.parmar@uni-passau.de (V. Parmar); alsayed.algergawy@uni-passau.de (A. Algergawy)

0000-0002-4370-2729 (V. Parmar); 0000-0002-8550-4720 (A. Algergawy)



[1]https://www.statista.com/statistics/871513/worldwide-data-created/

[2]https://www.cs.ox.ac.uk/isg/challenges/sem-tab/

become a central benchmark for table-to-knowledge graph matching systems. It provides benchmark datasets across domains such as food, biomedicine, and biodiversity, and invites systems to annotate them using structured knowledge graphs like DBpedia, Schema.org, and Wikidata.

Considering datasets from SemTab Challenge, we have recently developed systems addressing table understanding through minimalist ML-based annotation targeting DBpedia and Schema.org, domain-sensitive preprocessing pipelines for enhanced accuracy, and efficient Wikidata-driven annotation utilizing API-based lookup and caching [3, 4, 5]. This paper consolidates insights from our previous work into a forward-looking framework that strategically integrates preprocessing, knowledge graph APIs, and emerging LLMs, employing LLMs selectively rather than end-to-end, to achieve robust and scalable table understanding.

## 2. Table Understanding : An Overview

Table understanding refers to the automatic interpretation of the structure and semantics of tabular data. It enables machines to reason about the entities, types, and relationships described in tables, which are otherwise ambiguous and context-dependent. Semantic table interpretation is a subfield that focuses specifically on enriching tables with semantic annotations derived from structured knowledge graphs (KGs), such as Wikidata or DBpedia.

### 2.1. Core Tasks in Semantic Table Interpretation

The three fundamental tasks in semantic table interpretation are CEA, CTA, and CPA. Each task plays a distinct role in transforming a plain table into a semantically meaningful representation. To illustrate these tasks, consider the biomedical table shown in Table 1, which lists drugs, their biological target proteins, and approval years.

**Table 1**
Example biomedical table for illustrating semantic table understanding tasks.

| Drug Name | Target Protein | Approval Year |
|---|---|---|
| Aspirin | COX-1 | 1899 |
| Metformin | AMPK | 1994 |
| Imatinib | BCR-ABL | 2001 |

CEA links individual cell values to entities in a knowledge graph, enhancing semantic depth; for example, the value “Aspirin” in the Drug Name column of Table 1 can be linked to the Wikidata[3] entity wd:Q18216, and “COX-1” to wd:Q410251, representing cyclooxygenase-1. CTA assigns semantic types to columns, such as identifying Drug Name as a subclass of Pharmaceutical Drug (wd:Q12140), Target Protein as Protein family (wd:Q417841), and Approval Year as calendar year (wd:Q3186692). CPA discovers relationships between columns using properties from a knowledge graph; for instance, Drug Name and Approval Year may be linked by publication date (wdt:P577), while Drug Name and Target Protein may use a domain-specific property like has target.

### 2.2. Techniques used for Semantic Table Interpretation

The Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) has been running annually since 2019, promoting standardized evaluation of systems with tasks such as CEA, CTA, and CPA. Systems submitted to SemTab use a wide range of strategies, Rule-based or heuristic systems, JenTab used handcrafted rules to align columns and cells with entities and properties based on label matching, type constraints, and schema-based heuristics[6]. Another system, MantisTable uses heuristics and string similarity, column-type detection, and concept linking to interpret tables by

[3] https://www.wikidata.org/wiki/Wikidata:Main_Page

using resources like DBpedia and Wikidata [7]. Also there are some ML-based systems which shows appropriate results. For instance, a system called TURL uses structure-aware Transformer encoder tailored for tabular data [8]. These systems often rely on token frequency, column uniqueness, or embedding-based similarity. Knowledge Graph (KG)-driven system, SemTEX leveraged structured lookups using DBpedia or Wikidata APIs to directly retrieve and rank candidate annotations using gradient boosting [9].

## 3. Role of LLMs in Semantic Table Annotation

To address the core challenges of data noise, limited interpretability, and scalability in semantic table interpretation, our research contributes a modular pipeline combining domain-aware preprocessing, minimalist ML models, and scalable KG-based entity linking. We also reflect on recent advances in large language models LLMs, which offer promising capabilities for hybrid frameworks. This section first summarizes our contributions, followed by a brief discussion of complementary LLM-based efforts.

### 3.1. Modular and Scalable Table Annotation

To address the challenges of noisy data, scalability, and interpretability in semantic table annotation, we developed a modular pipeline comprising three components: (i) a ML-based structure annotation module, (ii) a scalable knowledge graph-driven lookup system, and (iii) a domain-agnostic preprocessing pipeline. Our first contribution, DREIFLUSS, introduced in SemTab 2023 Round 2 [3], is a minimalist logistic regression model tailored for CTA and CPA tasks. It employs count vectorized features extracted from tabular content and uses stratified sampling to handle label imbalance when working with Schema.org and DBpedia. Despite limited training data, DREIFLUSS demonstrated competitive performance, particularly on the CPA (DBpedia) task. This work shows the efficacy of simplistic approach obtaining competitive results with better sampling techniques. Building on this foundation, we developed a scalable CEA system for SemTab 2024 that uses live Wikidata API calls to perform cell-to-entity matching. To ensure throughput and robustness, we implemented multithreaded querying via Python's `ThreadPoolExecutor`, paired with caching mechanisms and a custom rate limiter to prevent API throttling. This system efficiently handles large tables across life science domains such as biodiversity and biomedicine[4]. Most recently, we introduced a domain-aware preprocessing pipeline [5], which performs anomaly detection (e.g., missing values, special symbols), normalization (e.g., multilingual variant alignment, abbreviation expansion), and rule-based refinement prior to annotation. This step alone yielded significant performance boosts for CEA across noisy datasets. A comprehensive summary of F1 scores and performance improvements achieved by our system across different tasks and datasets from the SemTab challenge is presented in Table 2. For the reproducibility of our work, we have made all our systems available on GitHub such as DREIFLUSS[4] as well as for Wikidata-driven annotation[5] and for preprocessing[6]. Together, these contributions reflect our commitment to building interpretable, efficient, and domain-agnostic solutions for semantic table understanding. Each component is independently deployable and complements the others, setting the stage for more integrated hybrid frameworks.

### 3.2. LLM-based Annotation and Motivation for Hybrid Integration

Recent advances demonstrate LLMs' potential for semantic table interpretation, with systems like CitySTI achieving effective cell-level entity disambiguation through LLM-based ranking and cleaning [10], while GPT-3 prompting reaches over 92% F1 across CEA, CTA, and topic detection in zero-shot settings [11], and LLM-driven CPA methods show fine-tuned GPT-3.5 outperforming traditional ML systems in column relationship identification [12]. Unlike traditional approaches requiring feature engineering or API integration, LLMs leverage contextual signals from table structure, headers, and

[4]https://github.com/vishvapalsinh/cta-cpa-schemaorg-dbpedia
[5]https://github.com/vishvapalsinh/CEACTA24
[6]https://github.com/DKEPassau/PreprocessMatch

**Table 2**
Performance of our modular annotation components across tasks and datasets.

| Dataset | Task (TragetKG) | F1 Score | Improvement | Component |
|---|---|---|---|---|
| SOTAB (2023) | CTA (Schema.org) | 0.7795 | N/A | ML-based |
| SOTAB (2023) | CTA (DBpedia) | 0.7694 | N/A | ML-based |
| SOTAB (2023) | CPA (Schema.org) | 0.7801 | N/A | ML-based |
| SOTAB (2023) | CPA (DBpedia) | 0.8312 | N/A | ML-based |
| tBiomed-Large (2024) | CEA (Wikidata) | 0.9250 | N/A | API-based annotation |
| tBiodiv-Large (2024) | CEA (Wikidata) | 0.9320 | N/A | API-based annotation |
| tBiodiv-Large (2024) | CTA (Wikidata) | 0.6150 | N/A | API-based annotation |
| Wikidata Tables (2024) | CEA (Wikidata) | 0.845 | +24.1% | Preprocessing |
| Biodiversity Tables (2024) | CEA (Wikidata) | 0.696 | +31.1% | Preprocessing |
| tFood Tables (2024) | CEA (Wikidata) | 0.858 | +4.5% | Preprocessing |

surrounding text, offering enhanced capabilities for ambiguous, incomplete, or multilingual data. Building on our specialized module results, we envision hybrid frameworks that fuse structured methods' interpretability mentioned in the section above with LLMs' contextual reasoning power, as outlined in our proposed modular architecture.

## 4. Future Framework: A Modular Hybrid Architecture

The proposed architecture, which we refer to as LifeTabFusion, is a modular and hybrid system designed to address the complex challenges of semantic table understanding in real-world domains. The framework integrates four main components developed through earlier work including domain-sensitive preprocessing, parallelized semantic annotation via API calls, lightweight ML models, and LLMs for disambiguation and interpretation.

### 4.1. Pipeline of Framework

The pipeline, as illustrated in Figure 1, begins with the ingestion of raw input tables in various formats, including CSV, Excel, or tables extracted from PDFs. These input tables are initially processed through a comprehensive preprocessing module that performs data standardization, anomaly detection, and domain-aware cleaning, preprocessing steps that have been proven to significantly enhance downstream annotation accuracy.

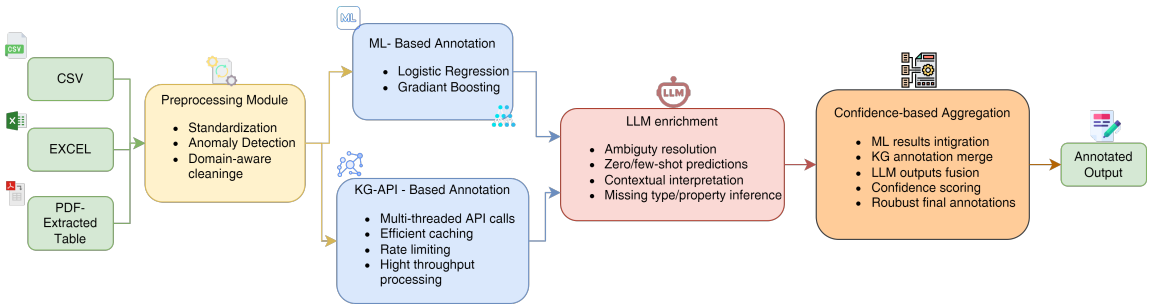


**Figure 1:** Proposed pipeline for LifeTabFusion

Following preprocessing, the cleaned table data is directed through two parallel processing paths to maximize efficiency and coverage. The first path employs a ML-based system that handles annotation tasks using lightweight algorithms, specifically logistic regression and gradient boosting models trained on knowledge graph features extracted from Schema.org and DBpedia. The second path implements parallel knowledge graph-based entity annotation, where individual cell values are processed through

multi-threaded API calls to external knowledge graphs such as Wikidata. This approach incorporates intelligent caching mechanisms and rate-limiting strategies to ensure high-throughput annotation of large tables while maintaining annotation accuracy. In the subsequent enrichment phase, Large Language Models, such as GPT-4, are employed to validate and enhance the results from both parallel paths. The LLM layer serves multiple critical functions, resolving ambiguities in entity disambiguation, performing zero-shot and few-shot predictions for missing type or property annotations, and interpreting contextual information from neighboring cells, column headers, and table captions. This contextual understanding enables more accurate and comprehensive annotations. The outputs from all three components; ML models, KG annotations, and LLM enrichments which systematically integrated through a fusion module that employs a confidence-based selection strategy[13]. This approach ensures robust final annotations by leveraging the strengths of each component while mitigating individual weaknesses. The annotated output can be exported in desired format. Despite seeming a promising approach, the proposed framework is still in its early stages and using LLM will be costly. To reduce the cost we need to explore the possibility of using open-source LLMs or fine-tuning smaller models on domain-specific data, which might be not as effective as paid services.

## Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

## References


[1] M. Campbell-Kelly, E. Robson, M. Croarken, R. Flood, The history of mathematical tables : from sumer to spreadsheets, 2003.

[2] A. O. Shigarov, Table understanding: Problem overview, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 13 (2022).

[3] V. R. Parmar, A. Algergawy, Dreifluss: A minimalist approach for table matching, in: SemTab@ISWC, 2023.

[4] V. R. Parmar, A. Algergawy, Wikidata-driven cea and cta for life sciences table matching extending dreifluss, in: SemTab@ISWC, 2024.

[5] V. R. Parmar, A. Hadder, A. Algergawy, On the role of preprocessing on matching tables to knowledge graphs, in: EKAW-PDWT, 2024.

[6] N. Abdelmageed, S. Schindler, Jentab meets semtab 2021's new challenges, in: SemTab@ISWC, 2021.

[7] M. Cremaschi, R. Avogadro, D. Chieregato, Mantistable: an automatic approach for the semantic table interpretation, in: SemTab@ISWC, 2019.

[8] X. Deng, H. Sun, A. Lees, Y. Wu, C. Yu, Turl: Table understanding through representation learning, 2020.

[9] E. G. Henriksen, A. M. Khorsid, E. Nielsen, A. M. Stück, A. S. Sørensen, O. Pelgrin, Semtex: A hybrid approach for semantic table interpretation, in: SemTab@ISWC, 2023.

[10] D. L. T. Yue, E. Jiménez-Ruiz, Citysti 2024 system: Tabular data to kg matching using llms, in: SemTab@ISWC, 2024.

[11] J. P. Bikim, C. Atezong, A. Jiomekong, A. Oelen, G. Rabby, J. D'Souza, S. Auer, Leveraging gpt models for semantic table annotation, 2024.

[12] K. Korini, C. Bizer, Column property annotation using large language models, in: European Semantic Web Conference, Springer, 2024, pp. 61–70.

[13] P. Betz, S. Lüdtke, C. Meilicke, H. Stuckenschmidt, Rule confidence aggregation for knowledge graph completion, in: International Joint Conference on Rules and Reasoning, Springer, 2024, pp. 32–49.