

Toward Cost-Efficient Metadata Curation in Agricultural Genomics with LLMs

Yi Liu¹, Matthew L Spangler², Jennifer Clarke^{3,4}, Hongfeng Yu^{1,5}

¹ *Holland Computing Center, University of Nebraska-Lincoln, Lincoln, NE, USA*

² *Department of Animal Science, University of Nebraska-Lincoln, Lincoln, NE, USA*

³ *Department of Food Science and Technology, University of Nebraska-Lincoln, Lincoln, NE, USA*

⁴ *Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE, USA*

⁵ *School of Computing, University of Nebraska-Lincoln, Lincoln, NE, USA*

Abstract—The rapid growth of DNA sequencing and genotyping technologies has produced an unprecedented volume of genomics data in agricultural research and industry settings. While the FAIR (Findable, Accessible, Interoperable, and Reusable) principles have guided the development of public repositories such as USDA’s Ag Data Commons and the National Center for Biotechnology Information (NCBI), the curation of lab-specific data into database-aligned formats remains labor-intensive, error-prone, and difficult to scale. The online, cloud-based public large language models (LLMs) raise concerns over data confidentiality, while locally deployable LLMs often lack the capacity to handle complex curation tasks. We present a locally deployable AI framework that automates metadata curation in agricultural genomics by integrating a transformer-based schema mapping model with an LLM-driven instruction interpreter. The system balances data confidentiality and reasoning capability, achieving over 80% mapping accuracy while supporting conversational, user-guided refinement. This demonstrates a scalable pathway for secure and efficient metadata standardization aligned with FAIR principles.

Index Terms—data curation, metadata, agricultural data, FAIR data principles, large language models

I. INTRODUCTION

The advancement in DNA sequencing and genotyping technologies has created an unprecedented volume of genomics data in agricultural production and research activities. However, because these datasets are collected under diverse experimental protocols and naming conventions, their metadata often lack standardization, thereby hindering data sharing, integration, and long-term reuse. This fragmentation demands mechanisms that can automatically harmonize heterogeneous metadata structures across research groups while preserving contextual accuracy and domain semantics.

To address the challenge and harness this wealth of data, the Findable, Accessible, Interoperable, and Reusable (FAIR) principle [1] has been introduced for modern scientific data management and stewardship, enabling actionability by both humans and machines. In alignment with this principle, public databases such as USDA’s Ag Data Commons [2] have been established to support dataset discovery, access, and reuse for

USDA-funded agricultural research. Similarly, the National Center for Biotechnology Information (NCBI) [3], originally established to serve biomedical and molecular biology disciplines, has expanded its support and tools to host genomics data from agriculturally relevant species.

Despite these advancements, FAIR-compliant repositories rely on standardized metadata formats, which can differ significantly from the diverse structures found in laboratory data collections. This mismatch introduces a major bottleneck in making lab data interoperable and reusable. Translating lab-specific data into FAIR-compliant formats often requires significant domain expertise, repetitive manual mapping, and continual adaptation to evolving schemas, making it a persistent challenge in the data curation process. Researchers often spend considerable time converting metadata tables by hand or writing ad hoc scripts to match repository templates. These manual workflows are difficult to scale, prone to errors, and unsustainable given the increasing pace and diversity of data generation in agricultural genomics.

The recent advancement in artificial intelligence (AI) can support the construction of an automated curation process that transforms group- or lab-specific data collections into formats compatible with FAIR-enabled repositories (e.g., NCBI). Specifically, large language models (LLMs) can interpret natural language instructions and translate them into computational actions. This conversational interaction allows users without computational expertise to construct, adjust, and validate curation plans through natural language dialogue, reducing the technical burden of data preparation. The finalized plan can be then used to automatically generate scripts that execute the metadata conversion.

However, there are two main challenges when applying LLMs. First, the use of public, cloud-based LLMs raises concerns about data security and privacy [4]–[6]. Due to effects such as memorization, user-provided data could inadvertently be exposed. This creates vulnerabilities that may leak information and compromise confidentiality. Second, the use of private LLMs could be limited by the availability of computational resources. High-parameter models deliver stronger reasoning performance but demand excessive computational resources, making them impractical for individual labs to host. On the other hand, the budget-friendly, smaller LLM models often

This work was supported by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) under Grant No. 2024-77039-43724 ODF: National Agricultural Produces Data Cooperative: Expansion through Strategic Development.

may not be able to generate complete curation plans due to a lack of multi-step reasoning ability required for accurate metadata curation and schema conversion. We aim to balance these extremes by leveraging lightweight, locally deployable LLMs in combination with specialized schema mapping models, thereby maintaining data confidentiality while still achieving the level of reasoning required for accurate metadata conversion.

In this work, we propose a locally deployable, interactive framework that integrates schema mapping models with LLMs to automate metadata conversion while incorporating feedback from users, such as data curation experts and domain-specific researchers. The proposed framework is designed to address the above two challenges. We integrate a lightweight LLM with a transformer-based schema mapping approach. Thus, it can be deployed locally to preserve data confidentiality while reducing the computational resource requirements. Specifically, the framework combines two major AI components : (1) a schema mapping model that predicts initial conversion plans between lab-specific tables and the target data format, and (2) an LLM-based natural language instruction interpreter that translates user feedback into actionable adjustments to the conversion plan. By combining a web interface, these components provide an interactive, conversation-based data curation framework.

The contributions of this work are threefold:

- We introduce a framework that integrates a transformer-based schema mapping model with an LLM-based natural language instruction interpreter, enabling both automated prediction of conversion plans and researcher-guided refinement.
- We demonstrate that this integration allows the use of lightweight, locally deployable LLMs, which preserve data confidentiality and reduce computational demands while maintaining the reasoning capability required for metadata curation.
- We design and prototype a web-based interactive interface that connects these components, providing researchers with a conversation-driven environment for curating lab-specific data into repository-ready formats.

Together, these contributions demonstrate a practical pathway toward AI-assisted, scalable, secure, and user-centered data format curation in agricultural genomics.

II. BACKGROUND

Since the foundational breakthrough of the transformer architecture [7], self-attention mechanisms enable efficient modeling of long-range dependencies without recurrence or convolution in natural language processing. Based on this architecture, BERT (Bidirectional Encoder Representations from Transformers) [8] adopts the encoder-only structure of the transformer architecture to learn contextual representations of texts. Thus, BERT is effective for language understanding tasks such as text classification, named entity recognition, and question answering. LLMs also extend the transformer architecture. They, on the other hand, are trained to predict

the next token in a sequence for text generation, reasoning, and few-shot learning. LLMs provide generative capabilities at a higher computational cost, such as content creation and zero-shot task performance.

Recent advancements in LLMs (e.g., GPT, Llama, and Claude) show their capacities to interpret natural language instructions and perform reasoning, fostering opportunities to reduce technical barriers in research activities [9]. LLMs have been investigated to assist in research tasks such as code generation, data analysis, and workflow design. For example, LLMs are explored for workflow planning and conversational assistance from natural language descriptions.

A. LLMs for Generating Scientific Workflows from Natural Language

One emerging application is utilizing LLMs to help researchers specify and build their scientific workflows via conversational human-machine interactions with natural language. Conventionally, workflow systems require users to learn syntax and certain programming skills to build tools (e.g., PhyloFlow [10], Common Workflow Language [11], and Toil [12]). LLMs potentially have the capabilities to translate natural language descriptions into machine-executable workflows. For example, Xu et al. [13] proposed a LLM-based automated workflow model generation tool, named LLM4Workflow. Their work leverages the abilities of LLMs in contextual learning to translate user-provided textual workflow descriptions into context-aware and executable workflow models.

In addition, LLMs are able to improve the scientific workflow thanks to their capacity for contextual understanding. Lin et al. [14] introduce a LLM-driven framework for chemical experimentation. Their work can interpret natural language instructions to generate executable code. The code is validated using a modular action library. Then, by integrating real-time environmental data, the robotic arm performs the experimentation code. Their approach shifts chemists' focus from routine procedures to creative, hypothesis-driven research.

Further, human-machine interactive systems have been investigated to enhance the capabilities of LLMs, inspiring researchers to design more effective and efficient scientific workflows. Craig [15] introduces a human-LLM note-taking system. This system embeds users as participants, allowing the user to inspect, revise, and adapt the system's processes and solutions. It promotes creative problem-solving by combining human expertise with AI reasoning in a shared, transparent workspace for effective and efficient scientific workflows.

Overall, the recent work shows proof-of-concept success in using LLMs to translate natural language instructions into pipeline steps. However, LLM-generated workflows do need verification and refinement since LLMs can be prone to errors or omissions because of hallucinations [16].

B. LLM Applications for Conversational AI

Another fold of LLM applications has developed conversational assistants to support researchers across data-intensive and academic tasks [17], [18]. LLM-based conversational

AI improves efficiency in fields that rely heavily on natural language communication. For example, Zedeh and Sattler [19] propose platforms to generate personalized narratives for vulnerable patients (e.g., the elderly, palliative care recipients, and multilingual communities) in community-based health surveys using conversational AI. Their platform utilizes LLM analysis of survey data to enhance patient engagement and deliver customized guidance. Rahman and Islam [20] explore improving knowledge management using ChatGPT (a chatbot based on LLMs) in university libraries. Their results show that ChatGPT enhances information retrieval and user engagement via an intuitive conversational interface.

Moreover, conversational AI could allow researchers to focus on more impactful analytical and creative work. For example, Chen, Cao, and Madden [21] introduce a generic LLM-centric framework, named *Lingua Manga*, to assist in creating data curation solutions using cloud-based services such as ChatGPT. The framework embeds LLMs to interpret user needs and generate executable code for data imputation pipelines and database connectivity (e.g., SQL-based systems). The conversational interface allows users to construct data curation workflows through natural language interactions, while simultaneously exposing programming interfaces that support direct intervention by technical users. Therefore, *Lingua Manga* provides flexible interaction for both technical and non-technical users.

In summary, conversational AI empowered by LLMs shows strong potential to bridge technical barriers related to programming, workflow creation, and execution. However, the cloud-based LLM raises data security concerns, such as ethical considerations in agricultural genomics [22].

III. FRAMEWORK DESIGN

Motivated by the recent advancement in LLMs and conversational AI, we are inspired to propose an interactive framework to automate metadata conversion. The design integrates a schema matching module, an LLM module, and a conversion module with a web interface to assist users in creating, inspecting, and executing the metadata conversion plan. The framework is locally deployable to ensure data privacy. In addition, we modularize the framework to allow the user to choose the model sizes based on their computational resource limitations.

A. Overview

The framework design is illustrated in Fig. 1. Based on a lab-specific dataset and a target dataset sample provided by users, the framework first attempts to generate an initial metadata conversion plan using the schema mapping module. The generated plan is displayed on the web interface for users to inspect. A user can provide corrections or additional instructions through chat on the web interface. The LLM module, then, parses the user comments to adjust instructions to modify the conversion plan. The framework allows users to iteratively modify the conversion plan through chat until the desired plan is achieved. Once the conversion is ready, the

conversion module will assist users in preparing the conversion script. The user can choose to download the conversion script and/or directly convert their data table via the web interface.

B. Schema Mapping Module

The schema mapping module is a machine learning strategy using the transformer model to predict metadata conversion plans based on given source table and target table samples. The module uses a zero-shot retrieval strategy that does not need labeled training data. Specifically, the module computes three similarity signals between each source metadata item and the candidate target metadata item.

Three considered signals are:

- **Lexical similarity** is computed using Levenshtein distance between column headers from the source (s) and target (t) tables: $lex(s, t)$. This captures character-level correspondence, accounting for insertions, deletions, and substitutions. This is effective for identifying minor variations in spelling or typographical inconsistencies.
- **Semantic similarity** is derived from contextual embeddings generated by transformer-based encoders (e.g., BERT). For each pair of column headers, embeddings are computed and compared using cosine similarity:

$$\cos(e_s, e_t) = \frac{e_s \cdot e_t}{\|e_s\| \|e_t\|}. \quad (1)$$

where e_s is the embedding from the source table, e_t is the embedding from the target table. This provides a measure of conceptual relatedness that is robust to synonyms and paraphrases beyond surface-level matching.

- **Structural similarity** indicates the data type of sample cell values within each column (e.g., numeric, date, categorical string, or percentage). This signal is a binary indicator that returns 1 if the dominant data type of the source (s) and target (t) matches (e.g., both numeric, date, or categorical) and 0 otherwise.

$$\text{struct}(s, t) = \begin{cases} 0, & \text{if type } s \neq t, \\ 1, & \text{if type } s = t. \end{cases} \quad (2)$$

This signal helps identify structural consistency, particularly in cases where headers are ambiguous or under-specified.

Based on the signals, the module computes a composite score. Each signal is normalized and aggregated into a composite score defined as:

$$\text{score}(s, t) = w_1 \cdot lex(s, t) + w_2 \cdot \cos(e_s, e_t) + w_3 \cdot \text{struct}(s, t) \quad (3)$$

where s and t stands for each source-to-target pair, e_s and e_t are their embeddings, and w_1, w_2, w_3 are tunable weights. Our preliminary grid search shows a relatively balanced performance when semantic and structural signals are favored (i.e., $w_1 = 0.2, w_2 = 0.4, w_3 = 0.3$).

Further, the module ranks source-to-target pairs for each source item. The module outputs the list of the highest-ranking

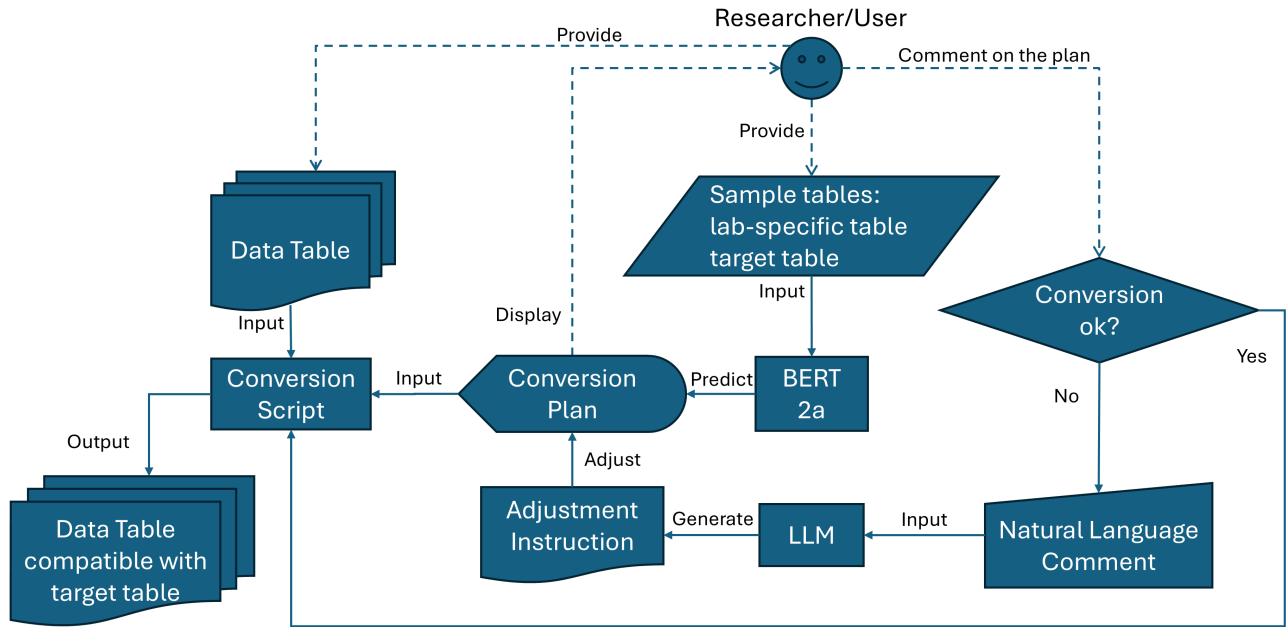


Fig. 1. Overview of the framework.

correspondence for each source item, with its composite score as the candidate mapping plan.

In addition, users can select the transformer-based encoders, ranging from compact BERT variants to larger contextual models, in the semantic similarity based on their available computational resources. This allows the flexibility to balance efficiency and semantic signal precision.

C. LLM Module

The LLM module is a natural-language instruction interpreter implemented using a local LLM service based on Ollama [23]. This module translates user comments into structured, machine-readable modifications to the conversion plan. Fig. 2 presents the pipeline for translating user comments into structured action, which are validated and applied to update the conversion plan. This module enables iterative human refinement through conversational interaction.

When a user reviews the predicted mappings, they can provide descriptive comments such as “map source item A to target B” or “remove the mapping of source item C.” The LLM first parses natural-language input into a JSON-formatted action object. Each action object encodes a specific modification (e.g., `map_source_to_target`, `unmap`, or `rename_source`) with the associated target and source attributes.

To ensure reliability and prevent unintended modifications, the generated JSON object is validated against a predefined list of allowed action formats. Only actions that match the format in the list are allowed to be executed by the module. This is a safeguard to mitigate the risk of illegal modifications, which can arise from the generative variability of LLM outputs. If an invalid modification is identified, the system will inform the user to resubmit the comment in the chat (e.g., the second

TMA chat record shown in Fig. 3. It preserves the stability of the conversion plan while still leveraging the flexibility of natural-language input.

Once validated, the parsed JSON-formatted action object is executed to adjust the conversion plan. The revised plan is updated and displayed to the user through the web interface, allowing continuous inspection and iterative refinement.

This module allows users to choose a range of LLMs from Ollama’s library based on their available computational resources. Similar to the flexible option in model selection of the semantic similarity in the schema mapping module, this allows users to balance efficiency and performance.

D. Conversion Module

The conversion module is responsible for transforming the validated conversion plan into an executable metadata conversion script that performs the actual format integration between the source and target tables. With the user-confirmed conversion plan displayed on the web interface, this module provides two options for the user. The user may download the generated script (e.g., Python script) for local execution within their own data pipelines, or they can execute the conversion directly within the web interface. This module completes the end-to-end workflow from schema prediction to metadata conversion. By coupling automated script generation with secure local execution, the conversion module ensures that metadata conversion can be generated efficiently, reproducibly, and without compromising data confidentiality.

E. Web Interface

The web interface is the interactive front end of the framework, connecting the user with the underlying modules. It is designed to provide a transparent and conversational workflow

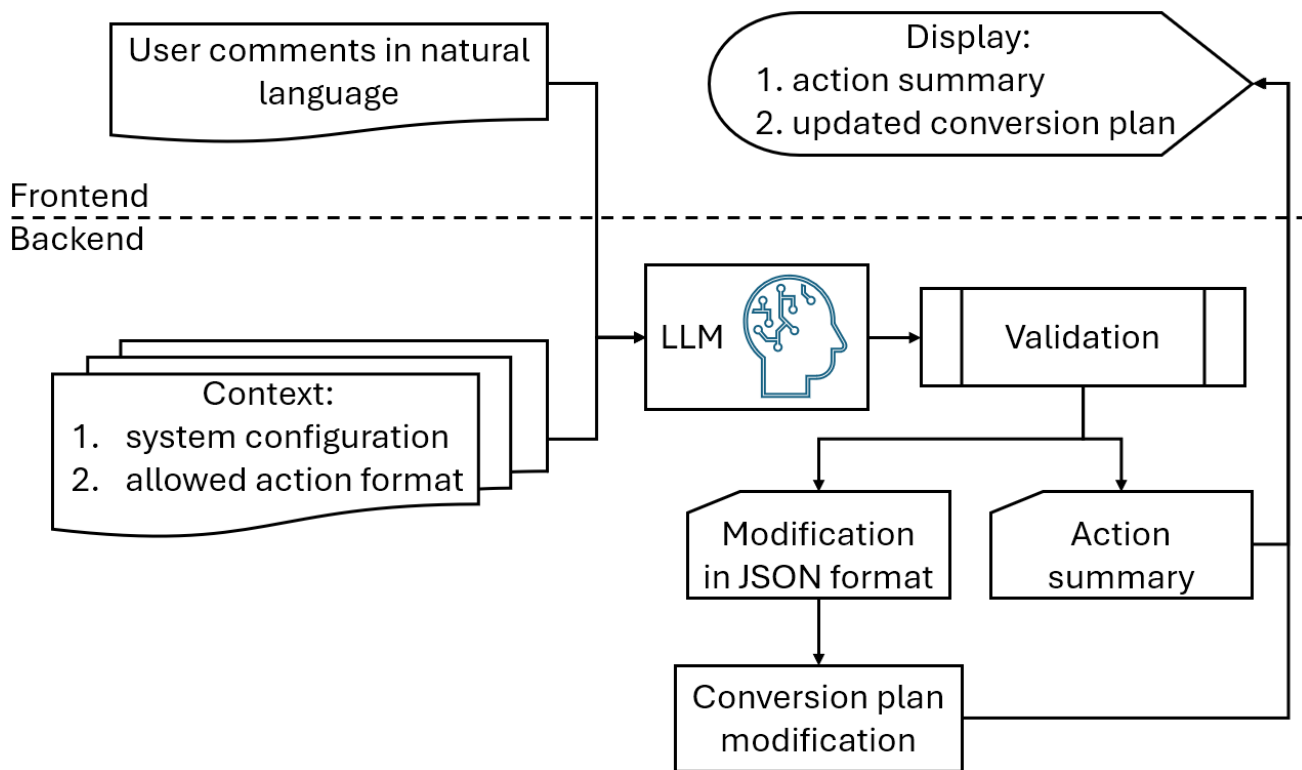


Fig. 2. The pipeline for translating user comments into structured action.

for the user to visualize, validate, and refine the conversion plan in real-time. Fig. 3 illustrates the layout of the interface.

The interface is divided into four primary functional panels: Chat History, Chat Conversation, Current Mapping Plan, and Conversion Controls.

1) **Chat History Panel** (left column)

This panel maintains records of previous conversion tasks. Each entry record shows the project name (e.g., SRA metadata shotgun and Metagenome Env 1.0). This feature supports users in revisiting prior work and reloading the corresponding conversion plans and conversion scripts.

2) **Chat Conversation Panel** (center column)

The conversation panel implements a chat-based interaction between the user and the LLM module (Table Mapping Assistant or TMA). The user provides natural-language comments in the lower input box, such as “map Library Preparation_alt to library_selection” or “drop the mapping of Library Preparation”. The LLM module will respond to the user comment. Any action that is interpreted as invalid will trigger error messages that guide the user to reformat their comments. This panel provides the interface for users to iteratively revise the conversion plan without programming knowledge.

3) **Current Mapping Plan Panel** (right column)

The mapping panel displays the current state of the conversion plan. It lists all detected source and target column pairs along with their composite similarity scores.

This panel provides real-time visualization of the current conversion plan. This panel also provides the starting point for users. To begin with a new project, users can type the project name and upload sample source and target CSV files to trigger the schema mapping module to generate the initial conversion plan. This panel will display the plan once the generation is finished, and enable the chat conversation panel for users to inspect the plan.

4) **Conversion Controls Panel** (lower left column)

After the mapping plan is finalized, users can access the Conversion Module through the two buttons here: (i) Download Script, which generates a Python-based conversion script for local execution; and (ii) Convert a given CSV file, which allows direct conversion of uploaded files through the web interface. Both options are based on the current conversion plan. Note that data confidentiality is maintained if the web interface is deployed locally for direct file conversion.

This web interface design integrates all framework functionalities shown in Fig. 1 into an intuitive and effortless workflow. The interface focuses on transparency, reproducibility, and accessibility, enabling researchers without programming experience to build metadata conversion to convert from a lab-specified data table to a FAIR-compliant data table (e.g., NCBI) via a guided, interactive process.

Table Mapping Assistant

Logout

Chat History

- SRa metadata shotgun
- Demo
- Metagenome Env 1.0 Shotgun

Chat Conversation

John: Drop the mapping of "Library Preparation". Then, mapping "Library Preparation" to "library_ID".

TMA: "operation": unmap, "detail": {"source": "Library Preparation"},
"operation": map_source_to_target, "detail": {"source": "Library Preparation", "target": "library_ID"}
Oct. 7, 2025, 6:45 p.m.

John: map "Library Preparation_alt" to "library_selection".

TMA: There was an error processing your instruction. Please try again.
Expecting value: line 4 column 5 (char 132)
Oct. 7, 2025, 6:48 p.m.

John: map source column "Library Preparation_alt" to target column "library_selection".

TMA: "operation": map_source_to_target, "detail": {"source": "Library Preparation_alt", "target": "library_selection"}
Oct. 7, 2025, 6:51 p.m.

John: "fil" maps to the target column "filename".

TMA: "operation": map_source_to_target, "detail": {"source": "fil", "target": "filename"}
Oct. 7, 2025, 6:56 p.m.

Type your message here... (Press Enter to send, Alt+Enter for new line)

Send

Current Mapping Plan

Mapping Title

Enter a title for the mapping plan

Upload Sample Source Table CSV (include column titles and a few data rows).

Choose File No file chosen

Upload Sample Target Table CSV (include column titles and a few data rows).

Choose File No file chosen

Upload and Process

source column	target column	composite score
librarySource	library_source	0.847387
fil	filename	0.000000
fastaFile	fasta_file	0.767483
des_des	design_description	0.578176
ass	assembly	0.599662
Title	title	1.000000
Library Preparation_alt	library_selection	0.758382
Library Preparation	library_ID	0.000000
Instrument Model	instrument_model	0.942717
Filename2	filename2	1.000000
	bioproject_accession	0.000000
	host	0.000000
	description	0.000000
	filename3	0.000000

Fig. 3. Layout of the web interface.

IV. FRAMEWORK DEMONSTRATION AND ASSESSMENT

A. Experimental Environment

We implemented our framework on a local PC and tested deployment on a high-performance computing cluster (HCC [24]) to ensure that the proposed framework could be deployed and benefit from the flexibility of choosing different models in the schema mapping and LLM modules in both low and high computational resource environments.

We also assessed the initial conversion plan prediction generated by the schema mapping module. The experiments used an NCBI-compliant sample table with 32 columns as the target sample table. To evaluate the schema mapping module under conditions that simulate real-world heterogeneity, we generated three variant sets of the table header names:

- an abbreviation-based set,
- a camel-case set, and
- a proper-case/rephrased set.

From these variants, we constructed two types of evaluation data. Table I demonstrates examples of generated variants in the three sets.

- 15 sub-tables were created by sampling mixed header styles from the variant sets to represent heterogeneous mixtures.
- 3 full variant tables were created by assigning each of the 32 columns to the same transformation strategy.

The 15 sub-tables were used to evaluate the performance of the initial plan prediction generated by the schema mapping module. The three full variant tables assess sensitivity to different variant generation methods.

B. Workflow Demonstration

The conversational interactions captured in the screenshot (Fig. 3) highlight how the user can iteratively revise the predicted conversion plan. In this section, we describe a typical workflow when using the web interface:

- 1) **Starting a new project:** The user starts by typing the project name and uploading sample tables in the "Current Mapping Plan" panel. The user first types the name of the project in "Mapping Title". Then, the user uploads a small sample of their lab-specific (source) table and a sample of the target table. The "Upload and Process" button triggers the schema mapping module, which predicts candidate mappings and displays them in the mapping plan table.
- 2) **Review the conversion plan:** The table lists each predicted source–target column pair alongside its composite score. High scores indicate strong confidence in the match, while low scores suggest uncertain matches. This display allows users to quickly identify potential errors or ambiguous mappings. Our observations suggest that:

$$\text{score}(s, t) = \begin{cases} (0.8, 1] & \text{if } s \text{ matches } t, \\ (0.6, 0.8] & \text{if } s \text{ ambiguously matches } t, \\ [0, 0.6] & \text{if } s \text{ likely not match } t. \end{cases} \quad (4)$$

- 3) **Revise the plan:** Using the chat box, the user can type natural-language comments to instruct the LLM module to modify the plan. The LLM interprets these comments, converts them into a structured JSON action, and validates them against a list of allowable operations. Valid actions are executed on the current plan, and the updated mapping table is updated and displayed in real-time. Any unrecognized actions trigger an error message, prompting the user to reformat their request.
- 4) **Generate conversion scripts or convert data:** Once satisfied with the conversion plan, the user can choose to download a Python conversion script using the "Download Script" button, or directly convert their full table via the "Convert CSV" upload box and the "Convert" button. The script consists of the code to execute the conversion and the instructions to use the script on the command line.

C. Conversion Plan Prediction Assessment

To assess the effectiveness of the initial prediction of the schema mapping module, we compared its performance on the 15 sub-tables and the three full variant tables. For each sub-table, we calculated the proportion of predicted mappings that correctly matched the target table. Across the 15 sub-tables (Fig. 4, accuracy ranged from 70.00% to 90.00%, with an average of 81.40% and a standard deviation of 0.0714. We observed that the sub-tables where lexical or semantic cues were preserved achieved higher accuracy, while abbreviation-heavy variants exhibited a relatively lower performance. These results demonstrate that the framework maintains consistent performance across heterogeneous header mixes but degrades when headers are shortened to the point of lexical ambiguity.

The full variant tables assess the sensitivity of the schema mapping module to different variant generation methods. Fig. 5 shows the accuracy distribution across the results of the three full variant tables, with an average of 85.42% and a standard deviation of 0.1301. Abbreviation-based variants achieved only 75.00% accuracy, camel-case variants reached 100.00% accuracy, and proper-case/rephrased variants achieved 81.25%. This verifies our observation in the 15 sub-tables assessment. The reduction in performance for abbreviations is likely due to excessive lexical shortening, which causes significant loss in both lexical and semantic signals. The camel-case variants retain lexical cues and therefore map easily, while rephrased headers map correctly but yield lower composite scores because of reduced lexical overlap despite semantic correspondence.

The assessment findings show that the capability of the schema mapping approach with the zero-shot strategy could predict an effective initial conversion plan. The average accuracy was above 80% across heterogeneous sub-tables. This allows the user to spend less time on iterative revision of the plan. In other words, the efficiency is enhanced. But the approach remains challenged by highly abbreviated variants

TABLE I
EXAMPLES OF GENERATED VARIANTS.

Original Name	Variant Set 1	Variant Set 2	Variant Set 3
bioproject_accession	bio_acc	bioprojectAccession	Bioproject Accession
isolation_source	iso_sou	isolationSource	Isolation Source
samp_mat_process	sam_mat_pro	sampMatProcess	Samp Mat Process
source_material_id	sou_mat_id	sourceMaterialId	Source Material Identifier
library_source	lib_sou	librarySource	Library Preparation
platform	pla	platform	Sequencing Platform

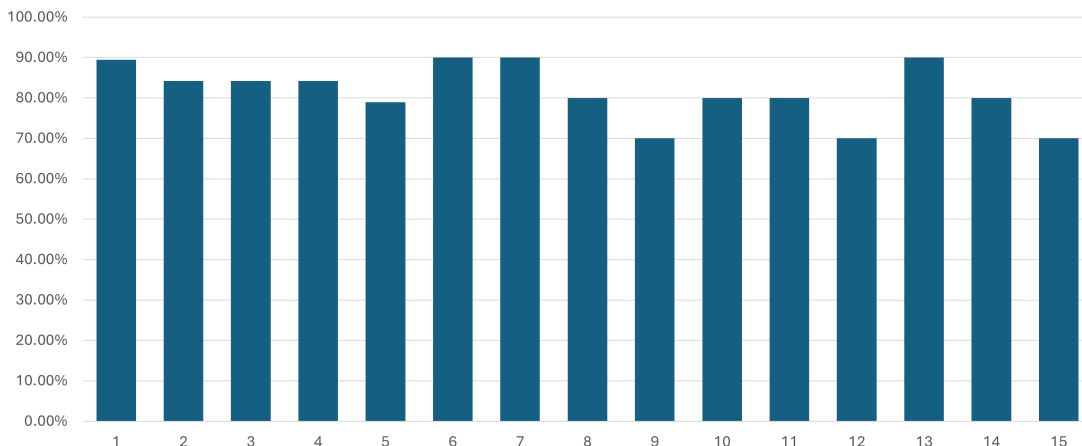


Fig. 4. The distribution of accuracy across the 15 sub-tables.

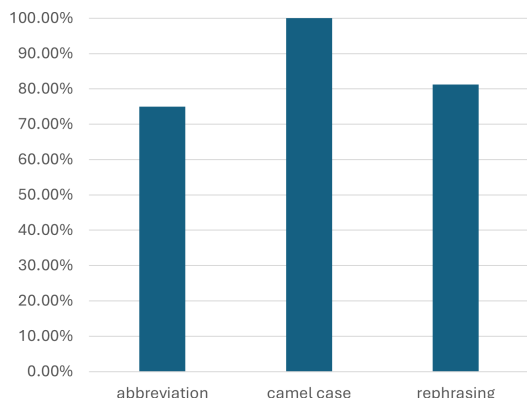


Fig. 5. The distribution of accuracy across the three full variant tables.

when the shortening starts to show loss in lexical and semantic cues.

V. CONCLUSION AND FUTURE WORK

This work presented an interactive framework that automates metadata conversion for agricultural genomics. The demonstration and assessment present that the proposed framework can generate initial metadata conversion plans with an average accuracy exceeding 80% across heterogeneous sub-tables. Then, using the conversational interface and LLM, the proposed framework allows the user to iteratively revise the plan with natural language comments. The demonstration showed that users could inspect predicted mappings, correct

ambiguities, and export executable conversion scripts. Since the framework can be deployed into a local environment or a controlled environment, the conversion does not disclose data to external services. These capabilities address two central challenges in agricultural genomics curation: reducing the labor required to align lab-specific data formats with FAIR-compliant standards and preserving data privacy through local deployment.

The results also showed limitations.

First, the assessment relied on a single standardized table expanded via variant generation due to lack of testing data. This assessment enabled controlled manipulation of header diversity and provided a reproducible environment for measuring the behavior of the schema mapping module. But the assessment may be biased from the diversity of real-world submissions. Incorporating real-world testing cases could offer an opportunity to comprehensively evaluate the performance under a wider range of practical conditions and to observe behavior in a context where agricultural genomics submissions consist of additional variability (e.g., lab-specific field conventions, heterogeneous levels of metadata completeness, and multi-table organizational structures). Second, the investigation focused on prototyping the integration of schema matching with advanced machine learning models. We characterized the behavior of the composite similarity strategy to understand how different signals and their relative weights influence performance. Future work could incorporate comparative baselines across established schema-matching approaches to situate our method within the broader landscape

of automated metadata alignment tools. Such comparisons would help identify scenarios where different approaches provide complementary strengths. Third, the composite similarity weights were set to emphasize semantic and structural cues over lexical overlap based on preliminary experiments. These weights directly influence mapping outcomes and may not generalize across databases or domains. An adaptive strategy could iteratively adjust weights within the human-in-the-loop framework by implicitly incorporating feedback from the user. Over time, this process may converge toward user-specific weighting strategies.

Overall, the proposed framework represents a practical step toward scalable, privacy-preserving metadata curation in agricultural genomics. By enabling local deployment and modular selection of model sizes, it accommodates flexible computational budgets while maintaining data confidentiality.

Future research will focus on four objectives.

- 1) **Broaden evaluation.** Future studies need to test real-world scenarios from multiple repositories, such as the European Nucleotide Archive, the Bovine Genome Database, and the USDA Ag Data Commons, to better capture real-world evaluation.
- 2) **Enhance similarity retrieval.** The three signals provide a useful baseline, but they may miss capturing deeper semantic relationships. Our future work will investigate retrieval models that jointly exploit lexical, semantic, and structural features, such as graph neural networks or pretrained models fine-tuned on schema matching tasks. Incorporating domain ontologies or controlled vocabularies could also improve alignment in specialized fields.
- 3) **Adaptive composite weighting.** The predefined selection of composite score weights limits generalization across datasets. One direction is to learn weights dynamically from user input during interactive sessions. Techniques such as Bayesian optimization or reinforcement learning could adjust the balance between similarity signals based on the success of previous mappings. This would allow the system to personalize its scoring strategy for different datasets and user preferences.
- 4) **Integration with downstream analyses.** The logs of mapping actions and corrections can provide valuable feedback. Our future work will investigate implementing mechanisms to collect usage statistics, identify recurring mismatches or frequent adjustments in column alignments. These analytics could generate reports for the user, highlighting common formatting issues and suggesting practices that facilitate automated conversion. This closed-loop feedback system would enable the framework to evolve alongside user needs and promote best practices in metadata curation.

ACKNOWLEDGMENT

This work was supported by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) under Grant No. 2024-77039-43724 ODF:

National Agricultural Produces Data Cooperative: Expansion through Strategic Development. Any views, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect those of USDA or NIFA. This work used computational resources of the Holland Computing Center at the University of Nebraska-Lincoln.

REFERENCES

- [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hoofit, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160018, Mar. 2016.
- [2] U. N. A. Library and K. S. Division, "Ag Data Commons research repository," <https://agdatacommons.nal.usda.gov/>, 2025.
- [3] "National Center for Biotechnology Information," <https://www.ncbi.nlm.nih.gov/>.
- [4] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting Training Data from Large Language Models," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2633–2650.
- [5] J. Lee, T. Le, J. Chen, and D. Lee, "Do Language Models Plagiarize?" in *Proceedings of the ACM Web Conference 2023*, ser. WWW '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 3637–3647.
- [6] A. Satvathy, S. Verberne, and F. Turkmen, "Undesirable Memorization in Large Language Models: A Survey," Mar. 2025.
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
- [9] Y. Zhang, S. A. Khan, A. Mahmud, H. Yang, A. Lavin, M. Levin, J. Frey, J. Dunmon, J. Evans, A. Bundy, S. Dzeroski, J. Tegner, and H. Zenil, "Exploring the role of large language models in the scientific method: From hypothesis to discovery," *npj Artificial Intelligence*, vol. 1, no. 1, p. 14, Aug. 2025.
- [10] J. Alvarez-Jarreta, G. de Miguel Casado, and E. Mayordomo, "PhyloFlow: A fully customizable and automatic workflow for phylogenetic reconstruction," in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2014, pp. 1–7.
- [11] M. R. Crusoe, S. Abeln, A. Iosup, P. Amstutz, J. Chilton, N. Tijanić, H. Ménager, S. Soiland-Reyes, B. Gavrilović, C. Goble, and T. C. Community, "Methods included: Standardizing computational reuse and portability with the Common Workflow Language," *Commun. ACM*, vol. 65, no. 6, pp. 54–63, May 2022.
- [12] J. Vivian, A. A. Rao, F. A. Nothhaft, C. Ketchum, J. Armstrong, A. Novak, J. Pfeil, J. Narkizian, A. D. Deran, A. Musselman-Brown, H. Schmidt, P. Amstutz, B. Craft, M. Goldman, K. Rosenbloom, M. Cline, B. O'Connor, M. Hanna, C. Birger, W. J. Kent, D. A. Patterson, A. D. Joseph, J. Zhu, S. Zaranek, G. Getz, D. Haussler, and B. Paten, "Toil enables reproducible, open source, big biomedical data analyses," *Nature Biotechnology*, vol. 35, no. 4, pp. 314–316, Apr. 2017.
- [13] J. Xu, W. Du, X. Liu, and X. Li, "LLM4Workflow: An LLM-based Automated Workflow Model Generation Tool," in *2024 39th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, Oct. 2024, pp. 2394–2398.

- [14] Y. Lin, Z. Wang, L. Zhang, C. Zhang, and X. Hei, "ChemAI: Empowering Robots to Automate Chemical Experiments with Large Language Models," in *2025 11th International Conference on Computing and Artificial Intelligence (ICCAI)*, Mar. 2025, pp. 369–373.
- [15] D. B. Craig, "Advancing Scientific Workflows: A Human-LLM Note-Taking System with Case-Based Reasoning," in *2025 IEEE Conference on Artificial Intelligence (CAI)*, May 2025, pp. 302–305.
- [16] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 42:1–42:55, Jan. 2025.
- [17] P. Kulkarni, A. Mahabaleshwarkar, M. Kulkarni, N. Sirsikar, and K. Gadgil, "Conversational AI: An Overview of Methodologies, Applications & Future Scope," in *2019 5th International Conference On Computing, Communication, Control And Automation (ICCCUBEA)*, Sep. 2019, pp. 1–7.
- [18] J. Gao, M. Galley, and L. Li, "Neural Approaches to Conversational AI," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ser. SIGIR '18. New York, NY, USA: Association for Computing Machinery, Jun. 2018, pp. 1371–1374.
- [19] P. M. Zadeh and D. Sattler, "Improving Accessibility and Readability of Survey Reports in Digital Health Platforms using Conversational AI," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2023, pp. 4987–4989.
- [20] M. H. Rahman and M. N. Islam, "The Impact of ChatGPT for Enhancing Knowledge Management in University Libraries," *Journal of Web Librarianship*, vol. 18, no. 4, pp. 177–196, Oct. 2024.
- [21] Z. Chen, L. Cao, and S. Madden, "Lingua Manga: A Generic Large Language Model Centric System for Data Curation," Sep. 2023.
- [22] D. Coles, L. J. Frewer, and E. Goddard, "Ethical Issues and Potential Stakeholder Priorities Associated with the Application of Genomic Technologies Applied to Animal Production Systems," *Journal of Agricultural and Environmental Ethics*, vol. 28, no. 2, pp. 231–253, Apr. 2015.
- [23] Ollama, "Ollama's documentation," <https://docs.ollama.com>, 2025.
- [24] University of Nebraska, "Holland Computing Center | Nebraska," <https://hcc.unl.edu/>, 2025.