PharmaSUG 2026 - Paper AI-141

Accelerating CDISC SEND Conversion with AI: From Raw Preclinical Data to Regulatory-Ready Datasets

Nattawit Pewngam, Chotika Chatgasem, Ravis Technology;
Titipat Achakulvisut, Department of Biomedical Engineering, Faculty of Engineering, Mahidol University

ABSTRACT

The Standard for Exchange of Nonclinical Data (SEND), developed by the Clinical Data Interchange Standards Consortium (CDISC), defines the standardized structure and format for submitting nonclinical study data to regulatory authorities. Converting extensive unstructured study materials, often consisting of reports, tables, and scanned documents, into SEND-compliant data sets remains a manual, error-prone, and time-consuming process that relies on repetitive data entry. This inefficiency reduces consistency, traceability, and overall regulatory readiness. Here, we introduce **CDISC-SEND Conversion platform,** an automated framework that integrates large language models (LLMs) with retrieval-augmented generation (RAG), to streamline and standardize this data transformation. Our platform rapidly normalizes and maps unstructured study content into SEND structures defined in the SEND Implementation Guide (SENDIG v3.1.1). Controlled terminology and sponsor metadata are retrieved dynamically to produce traceable, auditable, and standards-compliant mappings that demonstrate conformance and regulatory alignment. An expert review stage enables human validation and ensures accuracy before final data set approval. Results show that the workflow reduces preparation time from several weeks to less than a day while improving data consistency, and strengthening the key quality dimensions of completeness, structure, conformance, and format. Although originally developed for nonclinical SEND, the same architecture extends to clinical Study Data Tabulation Model (SDTM), providing a scalable and regulatory-aligned framework for AI-driven data standardization.

INTRODUCTION

The Clinical Data Interchange Standards Consortium (CDISC) establishes global standards for regulatory data submission across both clinical and nonclinical domains. Within this framework, the Standard for Exchange of Nonclinical Data (SEND) specifies the structure, format, and controlled terminology required for submitting nonclinical study data to regulatory agencies, thereby supporting traceability and consistent review across toxicology and pharmacology studies (CDISC 2021). Persistent challenges center on four areas including completeness, structure, conformance, and format (Zorn 2023). Completeness requires that each domain includes all Required and Expected variables and that Required fields contain valid values. Structure demands alignment with SEND Implementation Guide domain definitions, including correct records, keys, relationships, and value-level metadata. Conformance involves adherence to variable attributes, controlled terminology (National Cancer Institute (NCI) 2025), and business rules. Format addresses machine readability and consistent representations such as ISO date and time patterns and standardized units. Data quality and fitness issues often arise when narrative source materials are transcribed into spreadsheets, resulting in inconsistent identifiers, untraceable mappings, repetitive entries, and data sets that cannot reproduce reported results. Incorrect use of controlled terminology further undermines consistency. These risks indicate the need for comprehensive conformance workflows, consistent terminology control, and early validation to generate efficient, submission-ready data sets (Tibbs-Slone and Walker 2021).

To address these challenges, the CDISC-SEND conversion platform provides an automated and transparent solution for converting raw nonclinical data into standardized SEND outputs. The platform applies large language models (LLMs) to interpret and structure study content while referencing controlled terminology and SEND Implementation Guide (v3.1.1) standards. The resulting structured JSON enables early validation using tools such as Pinnacle 21 (Gupta, A. 2020), and an integrated review interface allows subject-matter experts to verify and finalize mappings while maintaining compliance with CDISC requirements.

Although large language models (LLMs) are probabilistic by nature and can exhibit emergent capabilities at scale (Wei et al. 2022), recent advances have demonstrated their strong capability to interpret domain-specific scientific content and produce structured outputs with high fidelity. Recent

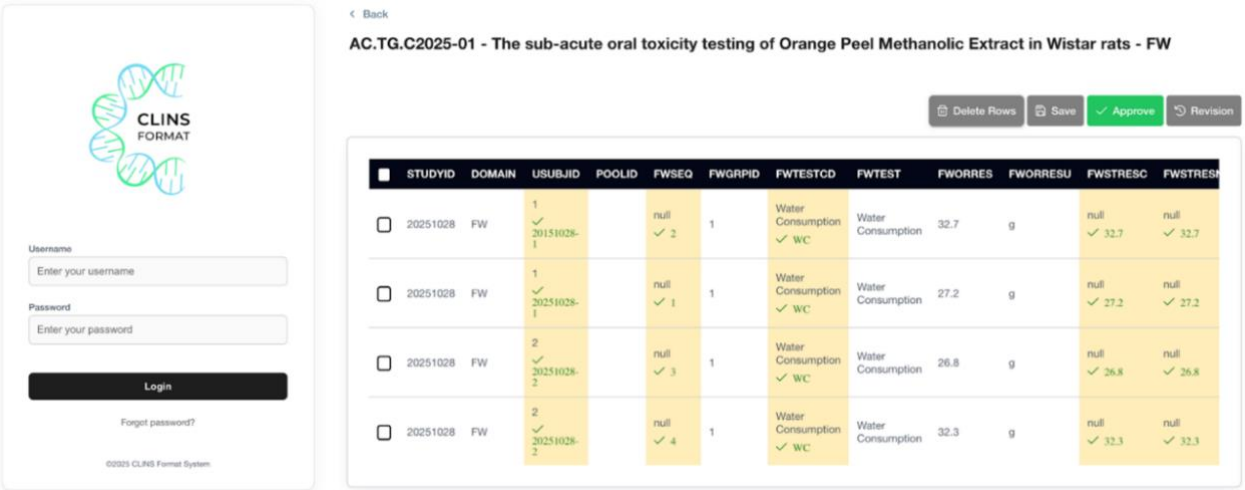
studies have reported promising performance of LLMs in biomedical data-extraction and schema-guided information-retrieval tasks, particularly when applied within structured, constraint-driven prompt environments (Brinkmann et al. 2024; Dagdelen et al. 2024). Their emergent reasoning abilities allow them to capture relationships, infer variable contexts, and match terminology across complex data sets that are difficult to standardize manually. These models can combine retrieval-based grounding and expert oversight to deliver consistent, auditable, and regulation-ready outputs while substantially reducing turnaround time (Lewis et al. 2020). Our platform integrates contextual engineering, which emphasizes structured reasoning environments for large language models (Mei et al. 2025; Dai et al. 2025), with retrieval-augmented generation (RAG), enabling LLMs to operate within a guided reasoning sequence bounded by schema and validation constraints aligned with SENDIG v3.1.1. This controlled design reinforces contextual reliability and ensures verifiable structured outputs. We can apply these guided standards with validation layers to improve the consistency of the LLM structured output.

Through this automated framework, the CDISC-SEND conversion platform improves consistency, reduces manual effort, and strengthens traceability throughout the SEND preparation process. Although initially developed for nonclinical SEND, the same architecture can be adapted for clinical SDTM, establishing a unified, AI-enabled approach to regulatory data standardization. The following section outlines the platform's overall design and processing architecture, describing how the system integrates automation, contextual reasoning, and validation to generate SEND-compliant data sets efficiently.

PLATFORM OVERVIEW

The CDISC-SEND conversion platform automates the transformation of nonclinical study data into SEND-formatted data sets, improving consistency, completeness, structure, conformance, and format in alignment with CDISC standards (Display 1). The architecture follows a sequence of Input, Processing, and Output (Figure 1). In the Input layer, users upload study reports, tables, and laboratory results through a secure web interface. In the Processing layer, large language models with RAG interpret study content and align extracted data with domain structures defined in the SEND Implementation Guide version 3.1.1. Two cooperating modules enable this workflow. The Master module manages project configuration, authentication, job orchestration, metadata and lineage persistence across the micro-frontend and microservices tiers, and packaging of outputs for JSON export. The SEND module performs standards-grounded domain and variable mapping, completes value-level metadata, harmonizes units and formats, and prepares structured outputs.

The Output layer focuses on expert review and export. Generated data appear as structured JSON that reviewers can confirm, revise, or reject to meet SENDIG and sponsor expectations. The SEND module provides controlled terminology and value recommendations with supporting rationale, while the Master module packages approved outputs and coordinates export for downstream conformance checks and submission assembly. Approved data sets are ready for validation with tools such as Pinnacle 21 Community before submission. Together, the Master and SEND modules, implemented as modular microservices, support scalability, transparency, and long-term maintainability and remain adaptable as CDISC standards evolve.



Display 1. Platform Interface

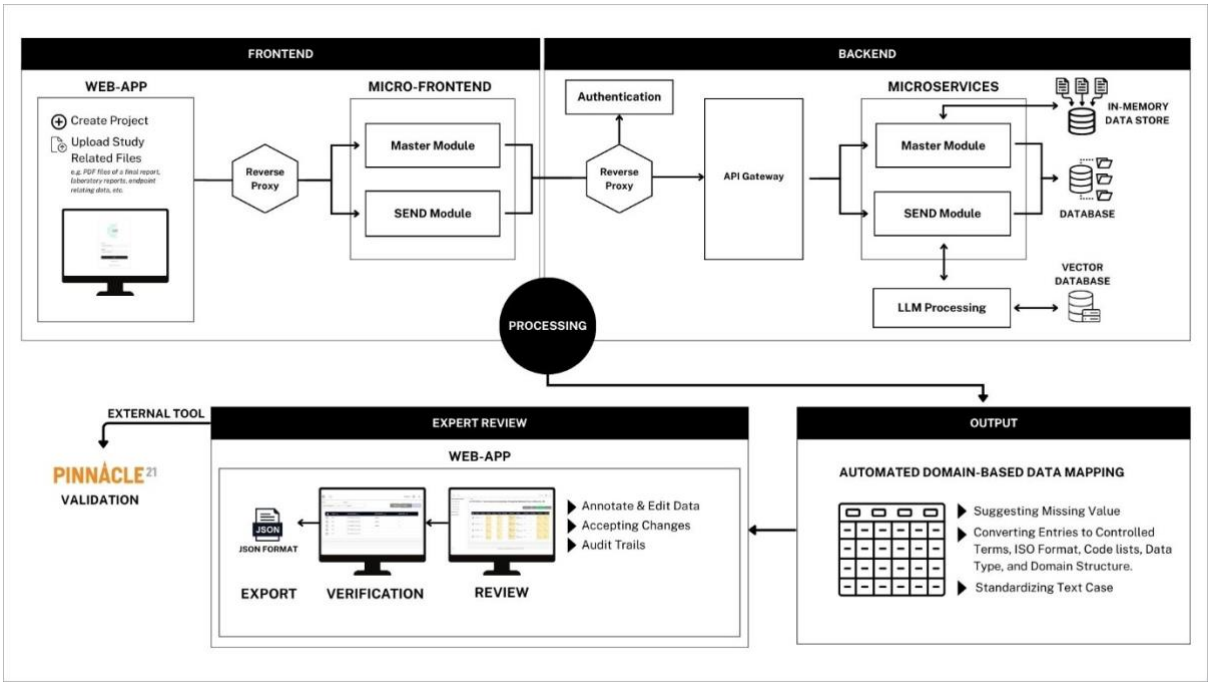


Figure 1. Platform Overview

LLM-RAG PROCESSING FRAMEWORK

The CDISC-SEND conversion platform employs a controlled LLM–RAG framework grounded in contextual engineering to convert unstructured nonclinical study content into SEND-compliant data sets (Figure 2). Study reports and tables are first normalized and segmented into traceable data units, after which large language models interpret the extracted content according to domain schemas embedded from SENDIG v3.1.1. Controlled terminology is dynamically retrieved through RAG to validate variables and ensure alignment with CDISC standards.

The pipeline operates through a governed multi-stage sequence of reasoning and validation layers. Each inference step is context-aware, constrained by schema definitions, and reinforced through retrieval from authoritative sources, ensuring that automated decisions remain explainable and auditable. All intermediate results are captured as structured JSON records preserving both raw and standardized representations, enabling transparent traceability without revealing proprietary mechanisms.

To mitigate the probabilistic nature of large language models, contextual engineering and deterministic normalization precede every inference step, constraining the input space and minimizing ambiguity. RAG restricts reasoning to verified knowledge bases such as CDISC-controlled terminology and sponsor-approved metadata, reducing the likelihood of hallucination or schema drift. Confidence thresholds and provenance tracking ensure that each mapped value is grounded in explicit evidence, reinforcing trust and auditability.

Each data set produced by the pipeline undergoes both automated and expert validation. Structural integrity and controlled-term compliance are verified programmatically, while subject-matter specialists review semantic accuracy through an interactive quality-assurance interface. This hybrid governance model ensures that LLMs function not as autonomous generators but as contextually engineered reasoning agents within a reproducible, inspectable data transformation system that combines the adaptability of AI with the rigor required for regulatory compliance and submission readiness.

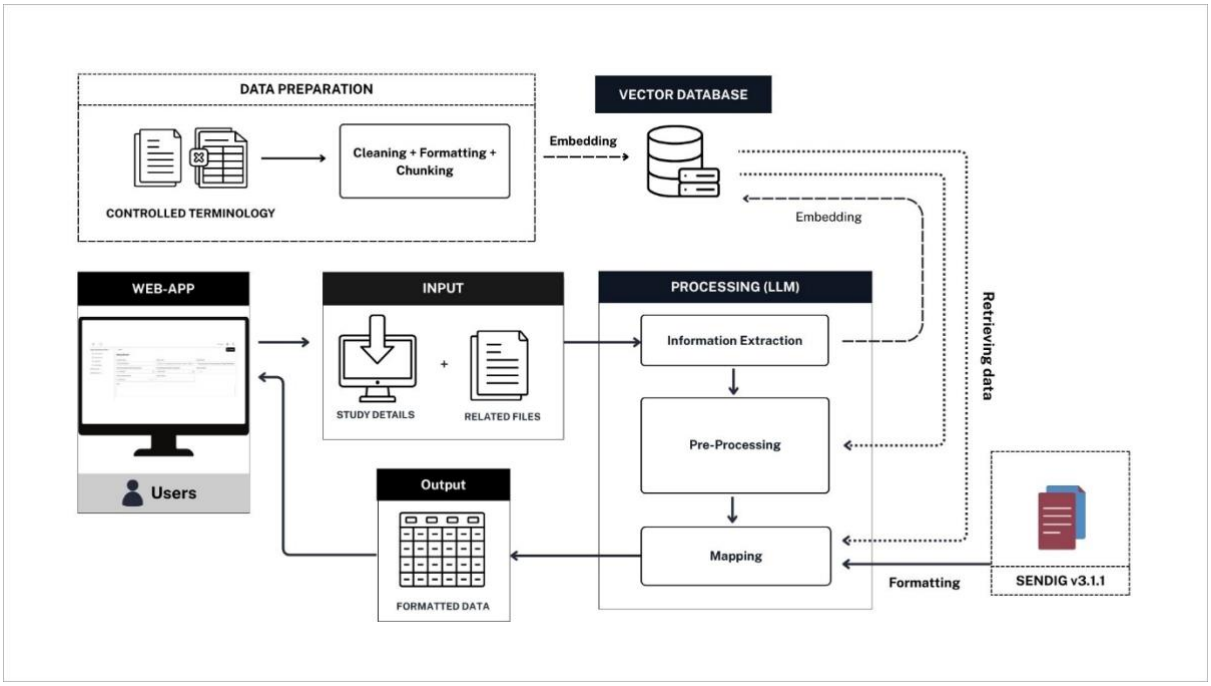


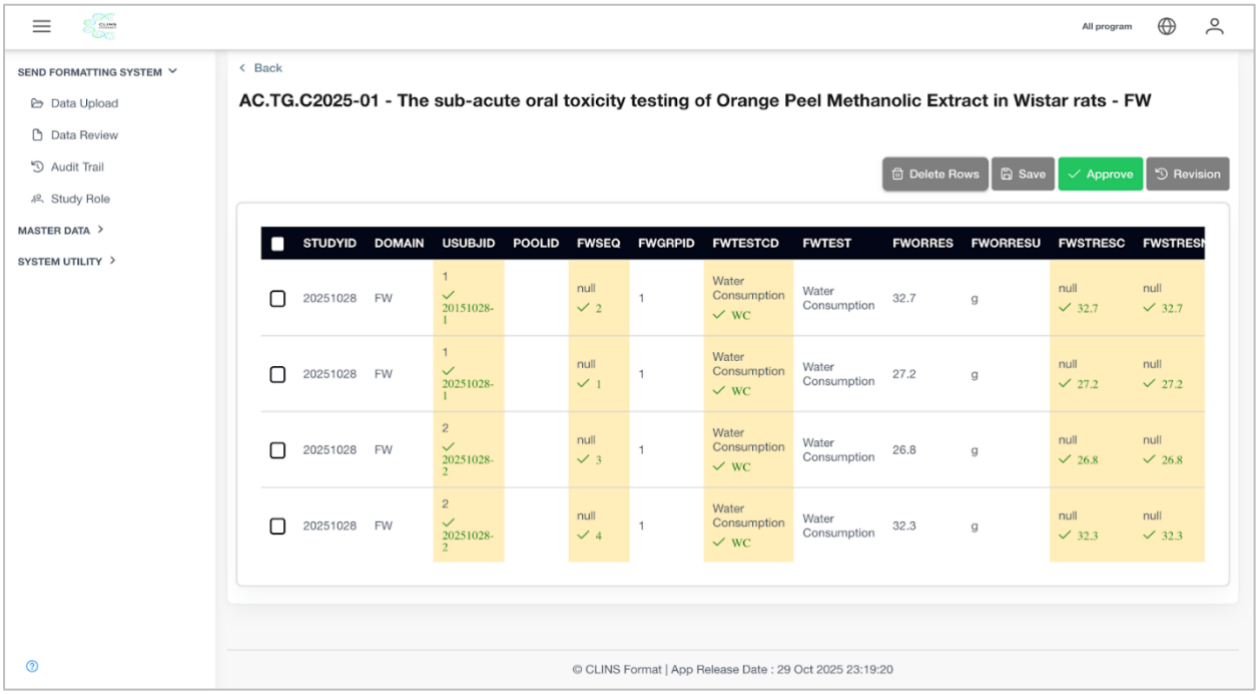
Figure 2. LLM-RAG Processing Framework – *The figure depicts how the CDISC SEND platform transforms unstructured nonclinical study content into SEND-compliant data sets through a governed sequence of normalization, pre-processing, and mapping steps. Large language models interpret study data within schema-constrained boundaries defined by SENDIG v3.1.1, while retrieval-augmented generation (RAG) supplies controlled terminology to validate variables and ensure traceable, auditable outputs ready for expert review.*

MODEL RESULTS

The controlled LLM-RAG processing framework demonstrates the platform's ability to automatically interpret, structure, and map unstructured nonclinical study data into SEND-compliant formats with high precision and consistency. For this evaluation, a PDF containing water consumption data was used as an example (Figure 3). The system successfully extracted relevant study information, including variables, units, and measured values, and then structured these elements according to the SENDIG v3.1.1. The resulting output, displayed in the platform interface, shows an automatically generated SEND-formatted data set accompanied by suggested controlled terminology and standardized values (Display 2). Suggested terms are visually marked with green check symbols below the original values, allowing reviewers to accept, modify, or reject the recommendations. This process confirms the framework's capability to maintain data traceability, ensure standard conformance, and align with regulatory requirements.

Group	Sex	No.	Water comsumption (g)						
			11/4/2024	12/4/2024	13/4/2024	14/4/2024	15/4/2024	16/4/2024	17/4/2024
			Day1	Day2	Day3	Day4	Day5	Day6	Day7
1	Male	1	27.2	32.7	28.3	25.7	27.3	26.8	25.8
		2	26.8	32.3	26.7	24.3	26.7	26.2	25.2
		3	30.1	30.9	32.8	33.6	17.7	22.6	20.2
		4	29.9	30.1	33.2	33.4	18.3	23.4	20.8
		5	32.0	24.2	26.2	35.5	26.6	25.6	26.6
		6	32.0	23.8	25.8	35.5	26.4	25.4	26.4
		7	29.8	25.2	34.3	37.4	29.3	31.8	32.8
		8	29.2	24.8	33.7	36.6	29.7	32.2	33.2
		9	21.9	24.7	27.5	27.7	26.6	25.6	24.2
		10	23.1	24.3	28.5	29.3	27.4	26.4	24.8
		113	26.8	27.2	25.2	36.4	27.9	26.9	37.0
		114	26.2	26.8	24.8	35.6	27.1	26.1	36.0

Figure 3. Input PDF Containing Water Consumption Data



Display 2. Output Interface Displaying SEND-Mapped Data and Value Suggestions

RESULT EVALUATION

The Pinnacle 21 Community Validator evaluated conformance of the platform outputs to CDISC SEND standards under two scenarios to assess the impact of automated terminology and value corrections. In the first scenario, data sets were exported without applying the platform's suggestions. As shown in Table 1. Pinnacle 21 Community Issue Summary before Applying Platform Suggestions, the validator reported FW-domain findings including CT2002 ("FWTESTCD value not found in 'Food and Water Consumption Test Code' extensible codelist," terminology category), CT2002 ("FWTESTCD and FWTEST values do not have the same code in CDISC CT," format category), and SD0018 ("Invalid value for FWTESTCD variable," metadata category). These findings indicate inconsistencies in controlled terminology, formatting, and metadata that prevented conformance to CDISC standards. In the second scenario, reviewers applied the platform's recommendations for controlled terminology, variable codes, and standardized values, as shown in Table 2. Pinnacle 21 Community Issue Summary

after Applying Platform Suggestions, validation reported no domain-level findings, confirming resolution of terminology and structural issues. Collectively, these results show that the controlled LLM-RAG process, supported by expert verification, improves codelist consistency and domain conformance, thereby enhancing data set quality, structural alignment, and readiness for regulatory submission.

Table 1. Pinnacle 21 Community Issue Summary before Applying Platform Suggestions

Issue Summary of FW Domain from Pinnacle Validator Report			
Pinnacle 21 ID	**Message**	**Severity**	**Found**
CT2002	FWTESTCD value not found in 'Food and Water Consumption Test Code' extensible codelist	-	84
CT2002	FWTESTCD and FWTEST values do not have the same Code in CDISC CT	-	84
SD0018	Invalid value for FWTESTCD variable	-	84

Table 2. Pinnacle 21 Community Issue Summary after Applying Platform Suggestions

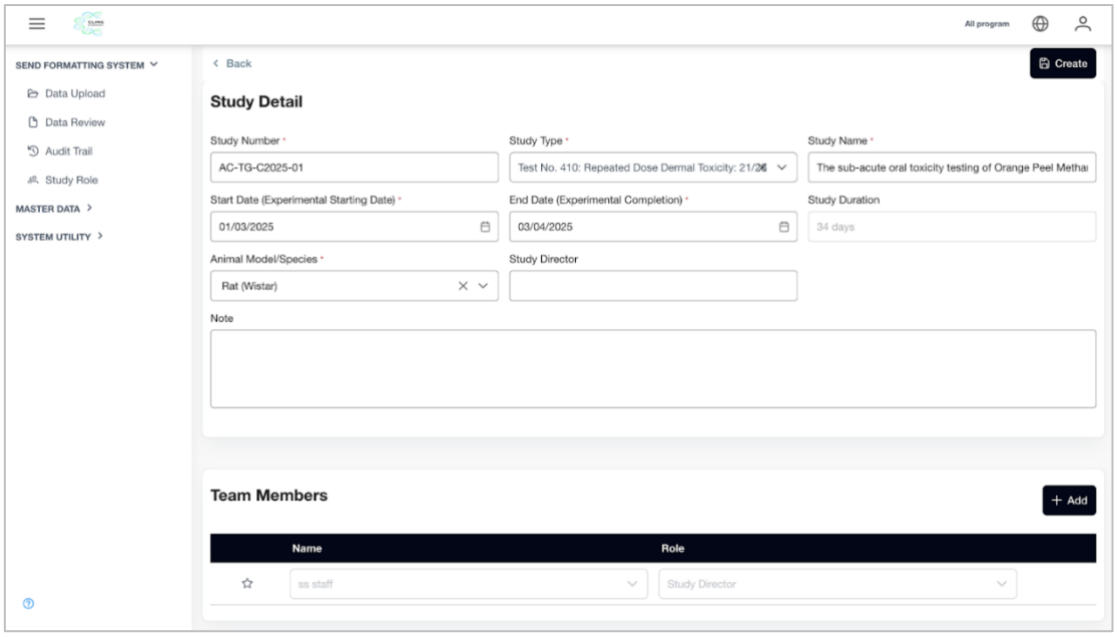
Issue Summary of FW Domain from Pinnacle Validator Report			
Pinnacle 21 ID	**Message**	**Severity**	**Found**
-	-	-	-

PLATFORM USAGE

This section outlines the end-to-end workflow, from secure project setup to automated formatting, expert review, and final export. The following steps outline the key stages of this process:

1. Project Setup

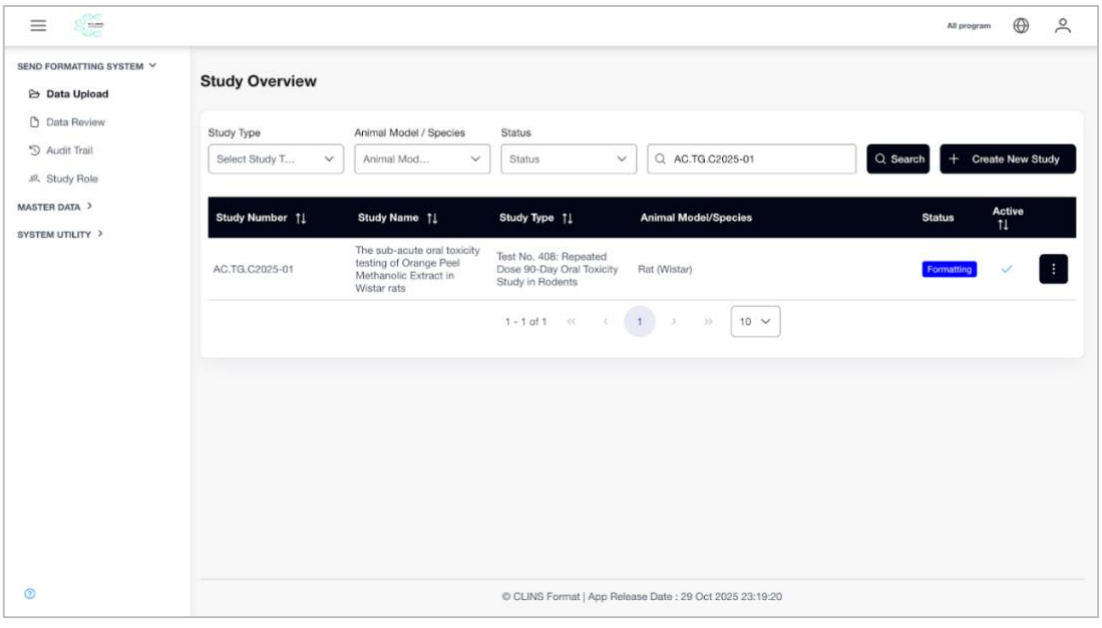
   Users start by accessing the platform through a secure login to ensure authorized use and data protection. After signing in, a new project is created within the system to manage study configurations and data sets. The Study Detail page allows users to configure essential study information such as study number, study name, type of study, start and end dates, animal model or species, and the study director. Team members and their roles can also be defined within the same interface to establish project-level accountability. Once the setup is complete, users upload the corresponding nonclinical study data, including laboratory results, reports, and other relevant documentation. This organized setup ensures that each data set is properly associated with its study configuration, providing a structured foundation for subsequent CDISC-SEND conversion process. (Display 3)



Display 3. Data Upload and Project Setup Interface

2. Automated SEND Formatting Process

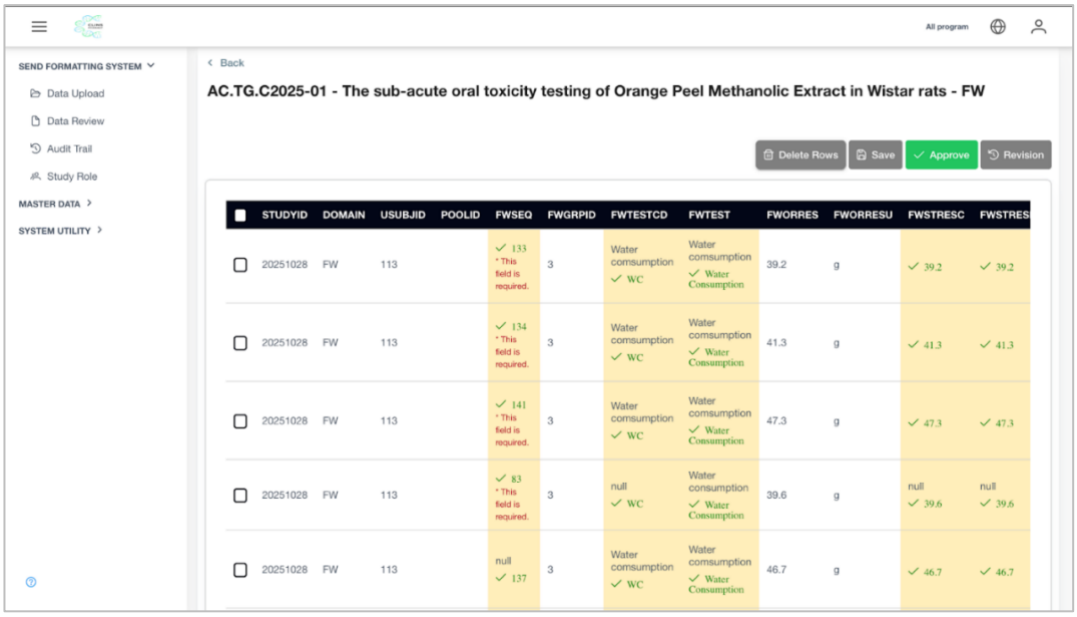
After users upload study data and project details, the system processes the files under the Formatting status. The SEND module runs an LLM-RAG processing framework that extracts and structures content, then generates domain and variable mapping suggestions aligned with SENDIG v3.1.1. and controlled terminology. The dashboard provides real-time visibility into each study, displaying progress indicators and showing the current formatting stage. (Display 4)



Display 4. SEND Formatting Process Interface

3. Specialized Review and Approval Interface

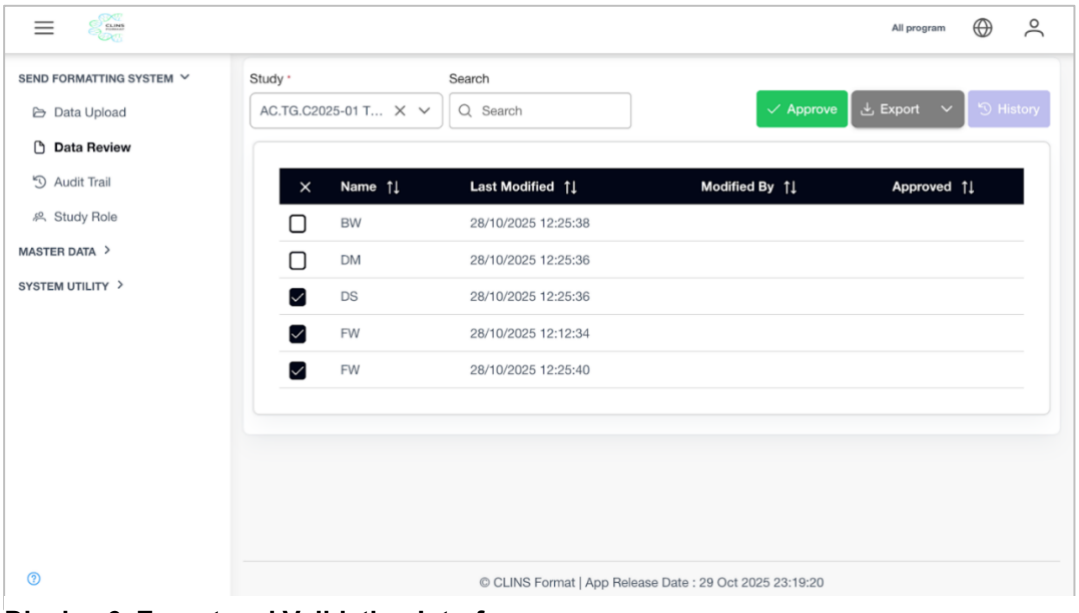
In this workspace, experts evaluate system-generated SEND mappings and value assignments at the record level. The tabular view presents key variables for verification and flags required fields or data gaps, for example “FWSEQ This field is required.” Controlled terminology suggestions appear beneath the original entries with green check marks, such as WC for FWTESTCD and Water Consumption for FWTEST. Reviewers can accept, modify, or reject suggestions, edit cells directly, and apply batch actions using Save, Approve, Revision, and Delete Rows. The system records all edits and decisions in an audit trail. After approval, the curated records are finalized for export as structured SEND-compliant outputs. (Display 5)



Display 5. Data Review and Approval Interface

4. Export and Validation

Users can review domain-level data sets such as BW, DM, DS, and FW, check modification details, and confirm approval status before exporting. Approved data sets are converted into structured JSON files. These files can be securely downloaded or transferred for further validation using external tools to ensure rule-based compliance, variable consistency, and structural integrity. (Display 6)



Display 6. Export and Validation Interface

CONCLUSION

The CDISC-SEND conversion platform demonstrates that integrating a large language model with retrieval-augmented generation and established data standards modernizes nonclinical data management. Automating the transformation of study data into SEND-formatted data sets reduces manual workload, limits mapping variability, and improves consistency. A modular microservices architecture supports scalability and traceability while maintaining alignment with the SEND Implementation Guide version 3.1.1 and controlled terminology. An expert review process ensures that automated decisions remain consistent with sponsor conventions and regulatory expectations. The results evaluation using Pinnacle 21 community shows that incorporating controlled LLM-RAG processing framework together with expert verification enhances code lists consistency and domain conformance, leading to improved data set quality and stronger readiness for regulatory submission. The same architectural pattern extends to clinical SDTM with minimal reconfiguration, enabling a unified AI-driven approach to data standardization that enhances quality, promotes cross-study consistency, and improves readiness for regulatory submission across nonclinical and clinical data sets.

REFERENCES

Brinkmann, A., Baumann, N., and Bizer, C. (2024). Using LLMs for the Extraction and Normalization of Product Attribute Values. arXiv preprint arXiv:2403.02130. https://arxiv.org/abs/2403.02130

CDISC. 2021. “Standard for Exchange of Nonclinical Data Implementation Guide (SENDIG) v3.1.1.” Available at https://www.cdisc.org/standards/foundational/send/sendig-v3-1-1.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., and Jain, A. (2024). Structured Information Extraction from Scientific Text with Large Language Models. Nature Communications, 15(1), 1418. https://doi.org/10.1038/s41467-024-45563-x

Dai, S., Zhang, L., Chen, R., Huang, J., Li, X., and Zhao, K. 2025. “OnePiece: Bringing Context Engineering and Reasoning to Industrial Cascade Ranking Systems.” arXiv preprint. Available at https://arxiv.org/abs/2509.18091.

FDA. 2024. “Study Data Technical Conformance Guide.” Available at https://www.fda.gov/regulatory-information/search-fda-guidance-documents/study-data-technical-conformance-guide-technical-specifications-document.

FDA. 2025. “Specifications for eCTD v4.0 Validation Criteria.” Available at https://www.fda.gov/drugs/electronic-regulatory-submission-and-review/ectd-submission-standards-ectd-v40-and-regional-m1.

Gupta, A. 2020. “Pinnacle 21 Community v3.0 – A User’s Perspective.” PPD, Morrisville, NC. Available at https://www.researchgate.net/publication/341654258_Pinnacle_21_Community_v30_-A_Users_Perspective.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, S., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. 2020. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” Advances in Neural Information Processing Systems (NeurIPS 2020). Available at https://arxiv.org/abs/2005.11401.

Mei, Y., Chen, X., Zhao, F., Lin, H., and Wang, Y. 2025. “A Survey of Context Engineering for Large Language Models.” arXiv preprint. Available at https://arxiv.org/abs/2507.13334.

National Cancer Institute (NCI). 2025. “CDISC Controlled Terminology for SEND.” Available at https://evs.nci.nih.gov/ftp1/CDISC/SEND/SEND%20Terminology.html.

Tibbs-Slone, E., and Walker, A. 2021. “SEND Implementation and Challenges.” Tanigaku, 2021(23):13–16. Available at https://www.jstage.jst.go.jp/article/tanigaku/2021/23/2021_13/_pdf/-char/en.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. 2022. “Emergent Abilities of Large Language Models.” arXiv preprint. Available at https://arxiv.org/abs/2206.07682.

Zorn, P. 2023. “SEND Dataset Quality Control: Best Practices and Recommendations.” Certara Blog. Accessed November 2025. Available at https://www.certara.com/blog/send-dataset-quality-control-best-practices-recommendations/.

ACKNOWLEDGEMENT

We gratefully acknowledge the Thailand Center of Excellence for Life Sciences (TCELS), recently rebranded as the Technology and Innovation in Life Sciences National Agency (TILSNA), for funding the development of this system and for their support in advancing life science technology and innovation in Thailand.

We also extend our appreciation to the National Laboratory Animal Center, Mahidol University (NLAC) a leading facility in Thailand for rodent animal model testing and fully compliant with OECD-GLP standards for their invaluable expertise in evaluating our model results. Special thanks go to Passaraporn Srimangkornkaew, Ph.D. (Research Assistant, Senior Professional Level) and Mr. Korakoch Arcomsilp (Computer Programmer, Senior Professional Level) for their expert guidance and technical support throughout this project.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Nattawit Pewngam
Ravis Technology Co., Ltd.
Nattawit_p@ravistechnology.com
www.ravistechnology.com

Chotika Chatgasem
Ravis Technology Co., Ltd.
Chotika_c@ravistechnology.com
www.ravistechnology.com