

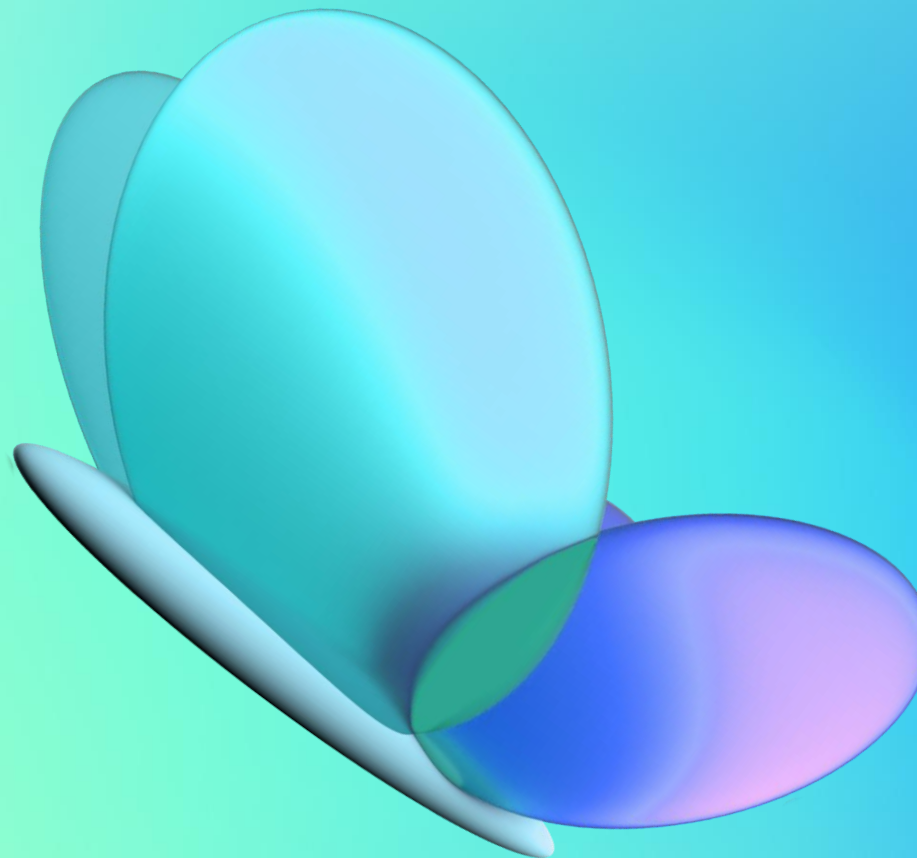


ML14

Enhancing Clinical Data Standardization Through Intelligent Mapping of eCRF Forms and Items to SDTM

Saransh Khandelwal, Lifio

Kedar Deshpande, Lifio



Problem Statement & Context



Why This Problem matters ?

- SDTM standardization is mandatory for **regulatory submissions** and cross-study analysis
- **eCRF designs** vary widely across studies, vendors, and therapeutic areas
- Manual mapping consumes significant time and drives rework during validation cycles



Core Mapping Challenges

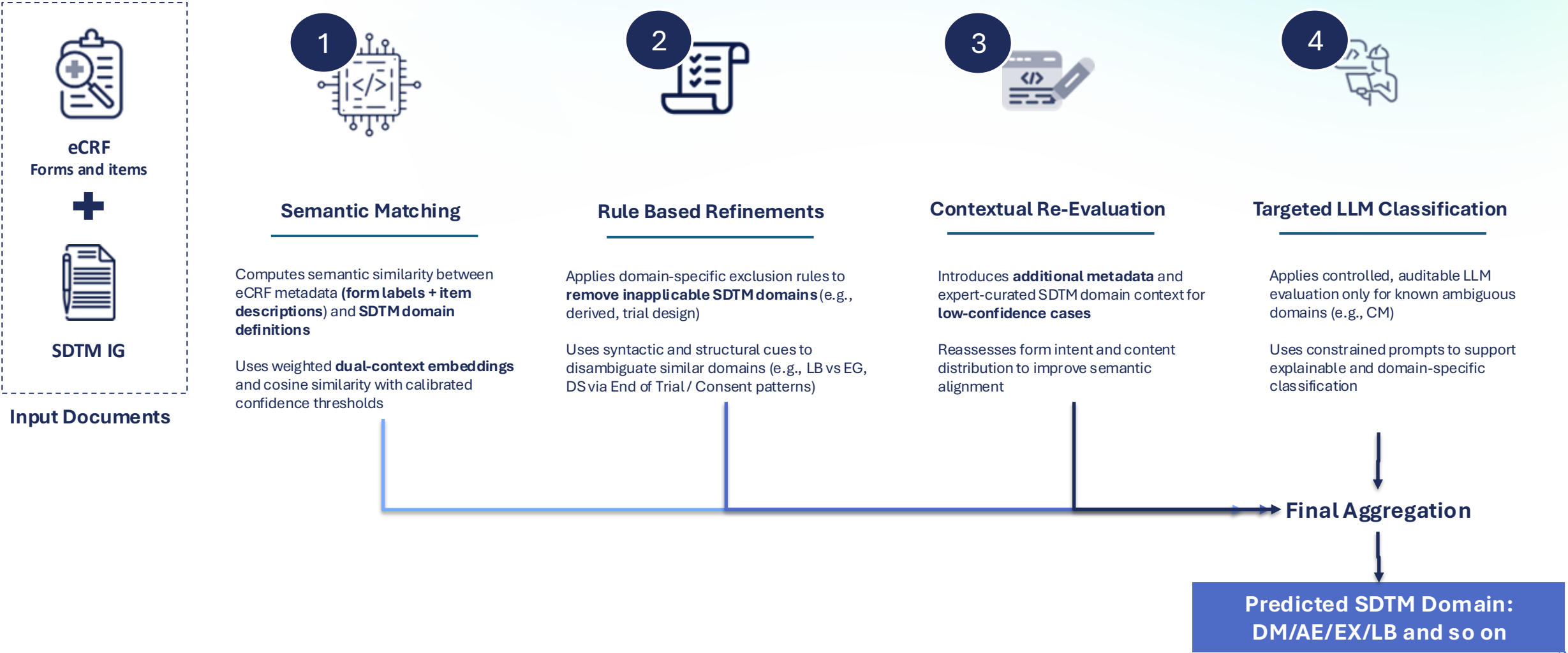
- High variability in form structures
 - ✓ *Example: Exposure data recorded in Exposure forms, Dose Administration forms, or embedded in Visit forms*
- Item-level mappings depend on context, domain intent, and CDISC conventions
 - ✓ *"Start Date" mapping to EXSTDTC, CMSTDTC, or AESTDTC depending on context*
- Rule-based or name-based automation fails under real-world heterogeneity



Need / Research Gaps

- **Pressing need** for a **hybrid, intelligent framework** integrating:
 - Semantic content mapping from supporting clinical documents with rule-based domain knowledge
- **Goal**
 - End-to-end automation of eCRF **form-to-SDTM domain classification** and,
 - Context-aware **eCRF item-to-SDTM variable** mapping within the assigned domain.

Form → SDTM Domain Mapping



Item → SDTM Variable Mapping



Predicted SDTM Domain



Human-in-the-Loop Safeguard

Preventing Error Propagation



Expert Review & Validation
Domain-Level Validation

Automation Stops Here

- ✓ Reduces cascading errors
- ✓ Preserves regulatory confidence



Item → SDTM Variable Mapping

Hybrid Semantic + Syntactic Strategy

Preprocessing & Context Enrichment

Semantic Matching
(Domain-Constrained)

Syntactic Matching
(CDISC Patterns)

Rule-Based Post-Processing

QVAL Fallback for Low Confidence

Item	CDISC Variable	Assignment
#QVAL01	--QVAL	- -QVAL

Fallback Variable Mapping
Assign #QVAL01 → --QVAL

Predicted SDTM Variable

Dual sentence embedding models are used in an ensemble to encode enriched eCRF question text as well as SDTM variable labels and associated CDISC notes

Co-occurrence rules ensure complementary variables are assigned together, such as pairing **TEST** with **TESTCD** when either is detected



Form	Item	Predicted Domain	Predicted Variable	Correct
EG	VR_EGRES	EG	[EGORRES, EGTEST, EGTESTCD]	1
PRBIOPSY	PRBIOPTYPE	PR	PRTRT	1
AE	AESTDAT	AE	AESTDTC	1
PCMCTX	TRTEND	CM	CMRSDISC	1
PM	PMROUTE	RP	QVAL	0

SDTM Mapping Algo Results



- Evaluated on **10 real-world clinical studies** with expert-annotated form- and item-level SDTM mappings across multiple scenarios
- Achieved **~90% item-level accuracy with expert-provided domains** and **~75% accuracy using predicted domains**, highlighting controlled error propagation
- Delivered **~80% form-to-domain accuracy** and automated >80% routine mappings, reducing manual effort by 60–70% with full explainability

Business & Operational Impact



Operational Efficiency



- Significantly accelerates study build and SDTM conversion timelines by reducing the reliance on manual, repetitive mapping activities
- Enables SDTM programmers to focus on complex mappings, validation, and oversight



Quality, Consistency & Traceability



- Improves **consistency and explainability** of mappings aligned with CDISC expectations
- Delivers consistent SDTM mappings across programmers, studies, and vendors



Foundation for Downstream Standards



- Provides a reliable base for SDTM transformation logic, **SAS code generation**, and mapping specifications
- Enables seamless extension to **ADaM data standards** and downstream analytical workflows

Future Scope & Conclusion



Production-Ready Automation

Demonstrates a scalable, production-ready framework for automating eCRF-to-SDTM mapping at form and item levels.



Hybrid Intelligence Approach

Combines semantic similarity, deterministic rules, and controlled AI to reduce manual effort while maintaining explainability.



Human-Governed Standardization

Integrates expert oversight to prevent error propagation and ensure regulatory confidence.



Expanded Domain Coverage

Future work will extend validation datasets across additional SDTM domains to improve robustness and generalizability.



End-to-End Standardization Vision

Framework will be expanded to support ADaM and downstream standards with improved confidence calibration and metrics.

