International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Generative AI for Data Mapping: Evaluating AI-Assisted Approaches to Legacy-to-S/4HANA Field Reconciliation

Sivasankararao Kavuri

Data Migration Lead, USA

ABSTRACT: Converting a legacy enterprise resource planning landscape to SAP S/4HANA requires reconciling tens of thousands of source fields against a simplified, restructured target data model that consolidates entities such as customer and vendor master records into a unified business partner model and replaces multiple finance tables with a single universal journal. Field-level reconciliation, deciding which target field a given source field should map to, what transformation it requires, and whether it should be retained at all, has traditionally been a slow, manual, and expertise-intensive exercise performed by functional consultants. This article evaluates whether generative artificial intelligence, specifically large language models combined with semantic embeddings and retrieval-augmented generation, can materially accelerate this process without compromising mapping quality or introducing unacceptable risk.

We propose a reference architecture for AI-assisted field reconciliation that combines embedding-based candidate generation, large language model reasoning over field context, calibrated confidence scoring, and a human-in-the-loop review workflow governed by an auditable mapping catalog. We compare this hybrid retrieval-augmented approach against four baselines, rule-based name matching, classic schema matching, embedding similarity alone, and unaided large language model prompting, on a simulated benchmark spanning five common conversion modules, including finance, materials management, sales and distribution, business partner, and custom Z-fields.

Across the benchmark, the hybrid retrieval-augmented approach achieved an F1 score of 0.91 for mapping correctness, compared with 0.53 for rule-based matching, while reducing human review effort from an initial 42 hours per 1,000 fields to 5 hours per 1,000 fields by the eighth simulated migration wave. Critically, the hybrid approach also reduced the rate of confidently incorrect, or hallucinated, mappings to under 2.5 percent across every module tested, compared with rates above 14 percent for unaided prompting on the most structurally complex custom field category. These results suggest that generative AI can substantially reduce reconciliation effort, provided it is grounded in retrieved organizational context and paired with calibrated confidence scoring and governed human review, rather than deployed as an unaided, free-form reasoning step.

We close with a discussion of design trade-offs, data privacy and model deployment considerations specific to prompting proprietary schema information, current limitations, threats to validity, and directions for future research, including active learning for mapping catalog growth and standardized field reconciliation benchmarks.

KEYWORDS: generative artificial intelligence, large language models, schema matching, entity reconciliation, retrieval-augmented generation, SAP S/4HANA, ERP data migration, data governance, embeddings, human-in-the-loop review.

I. INTRODUCTION

1.1 Motivation

Every large-scale conversion from a legacy enterprise resource planning system to SAP S/4HANA confronts the same structural problem: the target data model is not a superset of the source data model, it is a deliberately simplified and restructured one. Customer and vendor master records consolidate into a single business partner model, multiple finance tables collapse into a single universal journal, and material and plant data are reorganized around a leaner core. Every field in the legacy system, including thousands of organization-specific custom fields commonly referred to as Z-fields, must be individually reconciled against this new structure: mapped to an equivalent target field, transformed to fit a new type or length, consolidated with other fields, or deliberately retired. Field reconciliation of this kind has historically depended on functional consultants who combine deep knowledge of the legacy schema with equally deep

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

knowledge of the target data model, a combination that is scarce, expensive, and slow to scale across a large conversion program.

The schema matching research community has studied automated approaches to this class of problem for more than two decades, from early rule-based and structural matchers (Rahm and Bernstein, 2001; Madhavan, Bernstein, and Rahm, 2001) to learned matchers trained on labeled correspondence data (Doan, Madhavan, Domingos, and Halevy, 2002). More recently, pretrained language models and, specifically, generative large language models have opened a new avenue: rather than relying solely on structural or statistical similarity, these models can reason over field names, descriptions, sample values, and surrounding business context in much the same way a human consultant would (Li, Li, Suhara, Doan, and Tan, 2020; Peeters and Bizer, 2023). This article asks whether that reasoning capability can be harnessed reliably enough for high-stakes enterprise field reconciliation, and if so, under what architectural conditions.

1.2 Problem Statement

Three practical problems motivate this work. First, scale: a typical large enterprise conversion program involves reconciling tens of thousands of fields across dozens of modules, far more than a consulting team can review individually within a typical project timeline. Second, reliability: large language models are known to produce fluent, confident-sounding output even when factually wrong, a failure mode commonly termed hallucination, which is especially dangerous in field reconciliation because an incorrect but confident mapping can silently corrupt financial or operational data long after go-live. Third, governance: even where automation is accurate, conversion programs operate under strict audit and sign-off requirements that demand a documented, traceable rationale for every mapping decision, not merely a final answer.

1.3 Contributions

- **Reference architecture.** We present a reference architecture for AI-assisted field reconciliation that grounds large language model reasoning in retrieved organizational context, combines it with calibrated confidence scoring, and routes uncertain or high-impact mappings through a governed human review workflow.
- **Comparative evaluation.** We evaluate five method families, rule-based matching, classic schema matching, embedding similarity, unaided large language model prompting, and a hybrid retrieval-augmented approach, on a simulated benchmark spanning five conversion modules.
- **Hallucination measurement.** We explicitly measure and report the rate of confidently incorrect mappings produced by each method, a dimension largely absent from the classic schema matching evaluation literature.
- **Governance and privacy guidance.** We provide a RACI matrix, a model deployment comparison addressing data exposure risk, and a design trade-off summary intended to help practitioners adopt these techniques responsibly within a conversion program.

1.4 Structure of the Article

Section 2 reviews prior work on schema matching, language model based entity and column matching, and enterprise data migration. Section 3 characterizes the legacy-to-S/4HANA field reconciliation problem in more detail. Section 4 introduces the quality framework and maturity model used throughout the article. Section 5 presents the reference architecture, and Section 6 describes the generative AI techniques used for candidate mapping generation. Section 7 addresses confidence scoring, hallucination risk, and data privacy considerations, and Section 8 describes the human review and change control workflow. Section 9 describes the resulting system and module architecture. Section 10 describes the evaluation setup, Section 11 reports results, and Section 12 presents three illustrative case studies. Sections 13 through 16 discuss the findings, limitations, threats to validity, and future work, and Section 17 concludes the article.

II. BACKGROUND AND RELATED WORK

2.1 Classic Schema Matching

Rahm and Bernstein (2001) provide the foundational survey of automatic schema matching, classifying approaches by whether they exploit schema structure, instance data, or auxiliary information such as dictionaries and thesauri. Cupid (Madhavan, Bernstein, and Rahm, 2001) combines linguistic and structural similarity into a single hybrid matcher, while COMA (Do and Rahm, 2002) introduces a composite framework for combining multiple independent matchers and aggregating their results. Bernstein, Madhavan, and Rahm (2011) revisit this literature a decade later and conclude that, despite considerable progress, fully automatic schema matching still struggles with domain-specific semantics that

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

require external knowledge the matcher does not have access to, a limitation directly relevant to enterprise field reconciliation, where meaning is often encoded in business documentation rather than in the schema itself.

Doan, Madhavan, Domingos, and Halevy (2002) were among the first to frame schema and ontology matching as a learning problem, training classifiers on example correspondences rather than relying purely on hand-coded similarity rules, a direction that anticipates the learned and pretrained matchers discussed in Section 2.2. Miller (2018) and Nargesian, Zhu, Miller, Pu, and Arocena (2019) broaden the scope from single-pair schema matching to open data integration and data lake management, where the number of candidate sources and targets is far larger and matching must operate at scale, a setting structurally similar to a large enterprise conversion program.

2.2 Language Models for Matching and Column Understanding

The introduction of the transformer architecture (Vaswani et al., 2017) and pretrained language models such as BERT (Devlin, Chang, Lee, and Toutanova, 2019) enabled a new generation of matching techniques that represent field or column semantics as dense vector embeddings rather than discrete tokens. Sentence-BERT (Reimers and Gurevych, 2019) demonstrated that sentence-level embeddings could support efficient semantic similarity search, a technique this article's embedding-based baseline and retrieval-augmented candidate generation both build on. Ditto (Li, Li, Suhara, Doan, and Tan, 2020) applied pretrained language models directly to entity matching, showing substantial accuracy gains over prior learned matchers. EmbDI (Cappuzzo, Papotti, and Thirumuruganathan, 2020) learns embeddings jointly over heterogeneous relational data specifically for data integration tasks, and Sherlock (Hulsebos et al., 2019) and Doduo (Suhara et al., 2022) apply deep learning to the related problem of semantic column type detection, which underlies many of the type compatibility checks used in field reconciliation.

Aurum (Fernandez et al., 2018) and Valentine (Koutras et al., 2021) both address matching and discovery at the scale of many candidate tables and columns, and Valentine in particular provides a systematic evaluation framework that this article's benchmark design draws on for constructing controlled, ground-truth-labeled matching tasks.

2.3 Generative Large Language Models for Data Tasks

The emergence of large generative language models capable of few-shot and zero-shot reasoning (Brown et al., 2020) opened a further avenue distinct from embedding similarity alone: models that can be prompted with field descriptions, sample values, and business context and asked to reason directly about candidate mappings in natural language. Narayan, Chami, Orr, and Re (2022) evaluate foundation models across a range of classic data wrangling tasks, including entity matching and data cleaning, and find that prompted foundation models can match or exceed specialized prior systems on several tasks without task-specific training. Peeters and Bizer (2023) evaluate a generative language model specifically on entity matching benchmarks and report strong zero-shot and few-shot performance, while also noting sensitivity to prompt design and occasional confidently incorrect predictions, a finding directly relevant to the hallucination measurements reported in Section 11 of this article.

2.4 Enterprise Data Migration and Conversion Research

Outside the schema matching literature, enterprise system research has separately documented the operational challenges of large-scale data migration. Haller (2009) proposes patterns for industrializing data migration in standard software implementation projects, arguing that migration should be treated as a repeatable engineering discipline rather than a one-off manual effort, a framing this article extends specifically to the field reconciliation step. Davenport (1998) documents how deeply enterprise systems embed themselves into organizational processes, which helps explain why field-level reconciliation decisions carry consequences well beyond the technical migration itself. Wang and Strong (1996) and Rahm and Do (2000) provide the data quality vocabulary, including accuracy, completeness, and consistency, that underlies the reconciliation quality dimensions introduced in Section 4, and Loshin (2010) frames master data management as the broader discipline within which field reconciliation for a business partner or material master conversion ultimately sits.

2.5 Gaps in the Current Literature

Three gaps motivate the present work. First, the schema matching and language model literature reviewed in Sections 2.1 and 2.2 evaluates matching accuracy but rarely measures confidently incorrect output as a distinct failure mode, even though this failure mode is what makes generative approaches risky in a governed enterprise setting. Second, the generative AI for data tasks literature reviewed in Section 2.3 evaluates general-purpose entity matching benchmarks rather than the specific, highly structured reconciliation problem posed by a legacy-to-S/4HANA conversion, with its consolidated business partner model, universal journal, and dense population of custom fields. Third, none of the reviewed literature addresses the data privacy and model deployment considerations that arise when an organization's

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

proprietary schema and field descriptions must be provided to a language model as prompt context. This article contributes toward closing all three gaps.

III. THE LEGACY-TO-S/4HANA FIELD RECONCILIATION PROBLEM

3.1 Field Reconciliation Challenge Categories

Field reconciliation problems encountered during a conversion program fall into a small number of recurring categories, summarized in Table 1. Understanding these categories is a prerequisite for designing detection and generation techniques that target the right kind of ambiguity, since a renamed field and a semantically drifted field require fundamentally different reasoning to resolve correctly.

Table.1. Legacy-to-S/4HANA field reconciliation challenge categories.

Challenge Category	Description	Representative Example
Direct rename	The source and target fields represent the same concept under a different name	A legacy field named CUST_NO maps directly to the business partner identifier field
Consolidation	Multiple source fields map to a single target field or structure	Separate customer and vendor address fields consolidate into one business partner address
Splitting	A single source field must be decomposed into multiple target fields	A combined name field splits into separate first name and last name target fields
Type or length change	The target field uses a different data type, length, or precision	A legacy amount field with two decimal places maps to a target field requiring higher precision
Semantic drift	The source and target fields appear similar but represent different business meanings	A legacy status code no longer aligns one-to-one with the target document status field
Custom field (Z-field)	An organization-specific field with no standard target equivalent	A custom field tracking a regional compliance attribute with no corresponding standard field
Obsolescence	The source field has no meaningful target and should be retired	A legacy field used only by a decommissioned interface that is not part of the target landscape

3.2 Key Data Model Simplifications

Several structural simplifications in the target data model are responsible for a disproportionate share of the reconciliation effort in a typical conversion program. Table 2 summarizes the simplifications most relevant to field reconciliation and their downstream implications.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Table.2. Key S/4HANA data model simplifications and their reconciliation implications.

Data Model Simplification	Description	Reconciliation Implication
Unified business partner model	Customer and vendor master records consolidate into a single business partner structure	Fields duplicated across legacy customer and vendor tables must be reconciled to one target set
Universal journal	Multiple finance and controlling tables consolidate into a single line-item table	Finance fields from several legacy tables must be traced to a single, wider target table
Simplified material and plant data	Material master data is reorganized around a leaner core table structure	Legacy extension and customization fields must be individually evaluated for continued relevance
Consolidated document status handling	Document status logic is centralized rather than replicated per module	Legacy status fields may no longer have a clean one-to-one target equivalent

3.3 Root Causes of Manual Reconciliation Bottlenecks

Table 3 summarizes the root causes that most often slow manual reconciliation efforts, drawn from the challenge categories in Table 1 and the structural simplifications in Table 2. These root causes directly inform the detection and generation techniques introduced in Sections 6 and 7.

Table.3. Root causes of manual reconciliation bottlenecks.

Root Cause	Description	Typical Consequence
Knowledge concentration	Only a small number of consultants understand both the legacy schema and the target model	Reconciliation throughput is bounded by the availability of a few specialists
Undocumented custom fields	Custom Z-fields frequently lack current or accurate documentation	Reconciliation requires archaeological investigation of code and historical usage
Inconsistent naming conventions	Legacy field names vary in convention across modules and customizations	Simple name-matching techniques produce excessive false positives and false negatives
Cross-module dependencies	A single reconciliation decision can have downstream effects in other modules	Reviewers must consider context beyond the immediate field being reconciled

IV. A FRAMEWORK FOR FIELD RECONCILIATION QUALITY

4.1 Field Reconciliation Quality Dimensions

Following the general data quality vocabulary established by Wang and Strong (1996) and applied specifically to field reconciliation, this article organizes reconciliation quality around six dimensions, defined operationally in Table 4.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Table.4. Field reconciliation quality dimensions.

Dimension	Definition	Representative Violation
Semantic correctness	The extent to which a proposed mapping preserves the true business meaning of the source field.	A field mapped by name similarity alone despite representing a different business concept in context.
Type compatibility	The extent to which the proposed target field can accurately represent the source field's data type and range.	A high-precision legacy amount field mapped to a target field with insufficient decimal precision.
Cardinality compatibility	The extent to which the proposed mapping correctly handles one-to-one, one-to-many, or many-to-one relationships.	Two legacy fields both mapped independently to the same target field, causing a silent overwrite.
Transformation completeness	The extent to which every required value transformation, such as unit conversion, is specified alongside the mapping.	A currency field mapped without a corresponding exchange rate or rounding transformation rule.
Traceability	The extent to which a mapping decision is accompanied by a documented rationale and evidence.	A mapping accepted with no record of why it was judged correct or which source it was validated against.
Custom field coverage	The extent to which organization-specific custom fields receive an explicit reconciliation decision.	A custom Z-field left unmapped and undocumented, discovered only after cutover.

4.2 From Manual Reconciliation to AI-Assisted Reconciliation

Table 5 contrasts the traditional manual reconciliation model with the AI-assisted model proposed in this article, following the same before-and-after framing used to evaluate the quality dimensions in Table 4.

Table.5. Comparison of manual and AI-assisted reconciliation approaches.

Dimension	Manual Reconciliation Approach	AI-Assisted Reconciliation Approach
Candidate generation	Consultant manually searches target data model documentation	Embedding retrieval surfaces ranked candidate target fields automatically
Rationale generation	Consultant reasons informally, rationale often undocumented	Large language model generates an explicit, retrievable rationale for each candidate
Confidence assessment	Implicit, based on individual consultant judgment	Calibrated confidence score attached to every candidate mapping
Review allocation	Uniform review effort across all fields regardless of difficulty	Review effort concentrated on low-confidence and high-impact fields
Knowledge reuse	Prior mapping decisions rarely reused systematically across waves	Accepted mappings feed a retrieval-augmented mapping catalog for future waves

4.3 An AI-Assisted Reconciliation Maturity Model

Organizations adopt AI-assisted reconciliation incrementally rather than all at once. Table 6 defines four maturity levels used throughout the evaluation in Sections 10 and 11, ranging from unaided manual reconciliation through fully governed, continuously improving AI assistance.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Table.6. AI-assisted reconciliation maturity model levels.

Maturity Level	Characteristics	Typical Candidate Generation Method
Manual	No automated candidate generation; consultants work directly from documentation	None
Tool-Assisted	Simple name and pattern matching narrows the candidate list for review	Rule-based and classic schema matching
AI-Assisted with Review	Embedding retrieval and language model reasoning propose ranked candidates for every field	Embedding similarity and unaided LLM prompting
Continuous AI-Governed	Retrieval-augmented generation grounded in a growing mapping catalog, with calibrated confidence and audit logging	Hybrid RAG plus LLM

V. PROPOSED REFERENCE ARCHITECTURE

5.1 Design Goals

The architecture is organized around four goals drawn directly from the gaps identified in Section 2.5. Candidate mappings should be grounded in retrieved organizational context rather than produced by unaided free-form reasoning, since Section 2.3 shows that grounding materially reduces confidently incorrect output. Every candidate mapping should carry a calibrated confidence score that determines whether it can be accepted automatically or must be reviewed by a human. Every accepted mapping should be logged with its rationale and evidence so that the mapping catalog itself becomes an auditable, reusable asset across migration waves. Finally, the architecture should make explicit, configurable choices about where language model inference occurs, so that organizations can manage the data privacy implications discussed in Section 7.3.

5.2 Architectural Overview

Figure 12 shows the resulting reference architecture as a pipeline running from legacy field extraction through governed approval. A legacy field inventory extract, comprising field names, descriptions, sample values, and table context, feeds a candidate generation stage that uses embedding retrieval to surface a ranked shortlist of plausible target fields for each source field. A large language model then reasons over each candidate, using the retrieved target field definitions and any similar mappings already present in the mapping catalog, and produces both a recommended mapping and a natural-language rationale. A confidence scoring engine calibrates the model's output into an actionable confidence band, described further in Section 7.1. High-confidence mappings populate a reconciliation dashboard for lightweight spot-checking, while lower-confidence or high-impact mappings route to a human reviewer queue. Approved mappings become part of the approved S/4HANA field mapping set and are fed back into the mapping catalog, improving candidate generation for subsequent migration waves.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Figure 12. Reference Architecture for AI-Assisted Legacy-to-S/4HANA Field Reconciliation

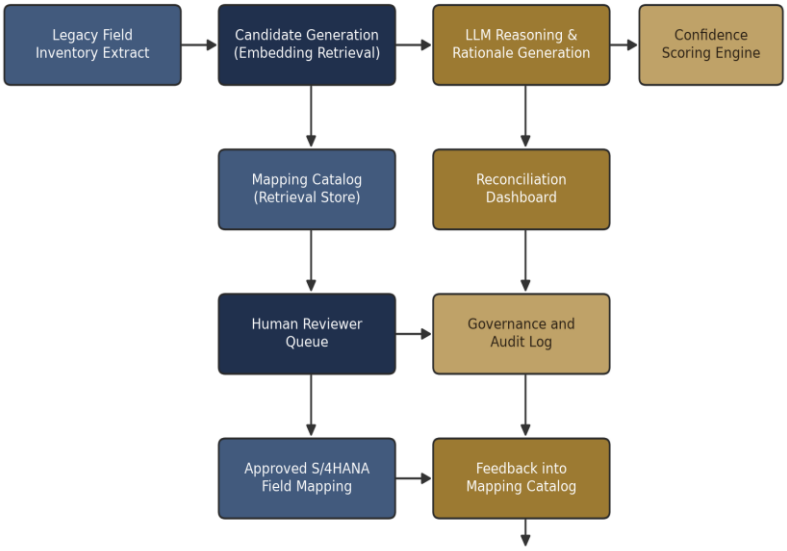

Figure.12. Reference architecture for AI-assisted legacy-to-S/4HANA field reconciliation

5.3 Module Descriptions

- **Legacy field inventory extract.** A structured extract of every source field, including name, description, data type, sample values, and originating table, used as the unit of work for reconciliation.
- **Candidate generation (embedding retrieval).** An embedding-based retrieval step that ranks target fields by semantic similarity to the source field's name, description, and sample values, described further in Section 6.1.
- **LLM reasoning and rationale generation.** A large language model that reasons over the retrieved candidates and business context to select and justify a recommended mapping, described further in Section 6.2.
- **Confidence scoring engine.** A calibration layer that converts model output into a confidence band used to determine automatic acceptance, escalation, or rejection, described in Section 7.1.
- **Mapping catalog (retrieval store).** A growing, versioned store of previously approved mappings and their rationales, used both as retrieval context for future waves and as the system of record for audit purposes.
- **Reconciliation dashboard.** A shared view of mapping status, confidence, and rationale used by both automated spot-checking and human reviewers.
- **Human reviewer queue.** A queue of mappings requiring human judgment, prioritized by confidence and business impact, described further in Section 8.
- **Governance and audit log.** An append-only record of every mapping decision, its confidence at the time of decision, and the identity of any human reviewer involved.
- **Approved field mapping and feedback loop.** The final, approved mapping set used to drive data transformation logic, with accepted mappings feeding back into the mapping catalog.

VI. GENERATIVE AI TECHNIQUES FOR CANDIDATE MAPPING GENERATION

6.1 Embedding-Based Candidate Generation

Following the sentence embedding approach introduced by Reimers and Gurevych (2019) and applied to data integration by Cappuzzo, Papotti, and Thirumuruganathan (2020), the architecture encodes each source field's name, description, and representative sample values into a dense vector, and retrieves the target fields whose embeddings are most similar under cosine distance. This step narrows a target data model containing thousands of fields down to a short, ranked list of plausible candidates before any language model reasoning takes place, substantially reducing both computational cost and the risk of the language model reasoning over an irrelevant candidate.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Table 7 summarizes the configuration of each generative AI technique evaluated in this study, including the embedding-based candidate generator and the language model based reasoning step described in Section 6.2.

Table.7.Generative AI technique configurations for candidate mapping generation

Technique	Key Configuration	Notes
Embedding similarity (baseline)	Sentence embedding model, cosine similarity, top-10 candidates retained	Fast and requires no labeled training data; used as a standalone baseline and as the candidate generator for the hybrid method
Unaided LLM prompting (baseline)	Zero-shot prompt containing only the source field description, no retrieved context	Represents a naive generative AI approach with no grounding in organizational context
Hybrid RAG plus LLM (proposed)	Top-10 embedding candidates plus mapping catalog exemplars provided as retrieved context to the language model	Grounds language model reasoning in both target schema definitions and prior approved mappings

6.2 Prompt Design for LLM-Based Reasoning

The structure of the prompt provided to the language model has a substantial effect on both accuracy and hallucination rate, consistent with the prompt sensitivity noted by Peeters and Bizer (2023). Table 8 summarizes the components included in the prompt template used by the hybrid retrieval-augmented method.

Table.8. Prompt template structure and components for LLM-based reconciliation.

Prompt Component	Purpose
Source field context	Provides the field name, description, data type, and representative sample values being reconciled
Retrieved candidate target fields	Provides the top-ranked target fields from embedding retrieval, each with its own definition and type
Retrieved mapping catalog exemplars	Provides similar mappings already approved in prior waves, giving the model concrete precedent to reason from
Explicit uncertainty instruction	Instructs the model to state uncertainty explicitly rather than select a candidate when no confident match exists
Structured output schema	Requires the model to return a candidate mapping, a rationale, and a self-reported confidence indicator in a fixed format

6.3 Confidence Patterns Across Field Characteristics

Figure 1 visualizes mapping confidence as a voxel density across three binned dimensions: semantic distance between source and candidate target field, data type compatibility, and business object complexity. Confidence remains high when semantic distance is low and type compatibility is high, regardless of complexity, but degrades sharply once both semantic distance and complexity increase together, precisely the combination that characterizes many custom Z-fields.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

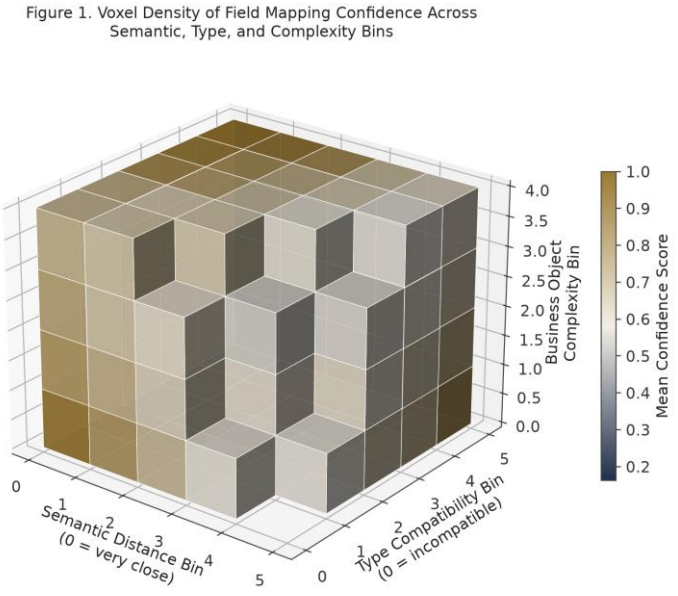

Figure.1. Voxel density of field mapping confidence across semantic, type, and complexity bins.

Figure 2 presents reconciliation error counts as a three-dimensional stem plot across module and field category. Custom Z-fields produce the highest error counts across every category, and semantic drift is the single most common error type across nearly every module, reinforcing the emphasis this architecture places on grounding language model reasoning in retrieved business context rather than field names alone.

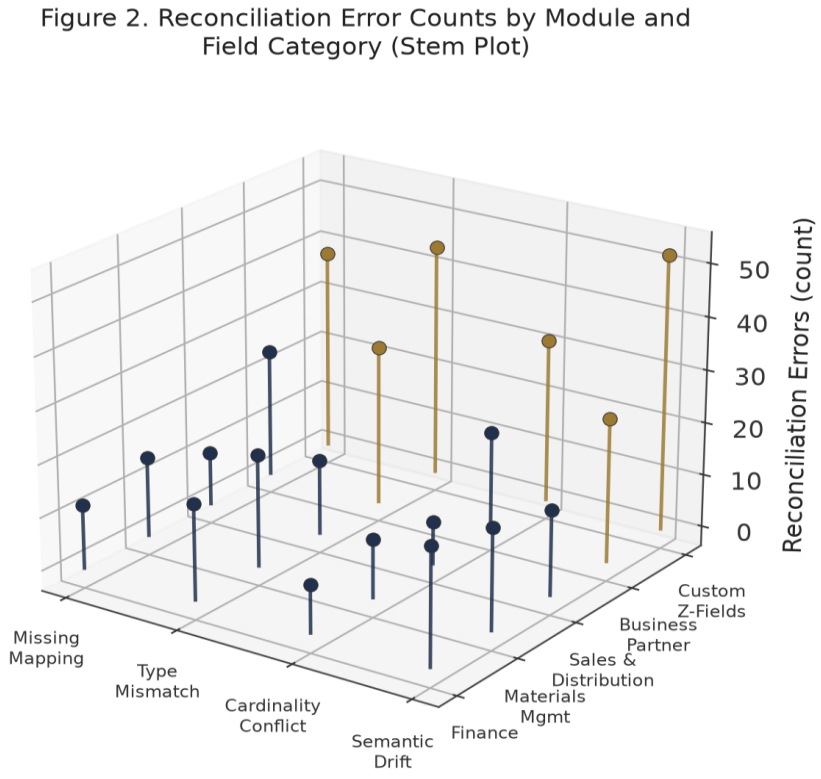

Figure.2. Reconciliation error counts by module and field category (stem plot).

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Figure 3 shows the distribution of mapping confidence scores by module as a three-dimensional ridgeline projection. Finance, materials management, and sales and distribution show tight, high-confidence distributions, while business partner fields show a wider spread and custom Z-fields show the widest spread and the lowest central tendency, consistent with the voxel pattern in Figure 1.

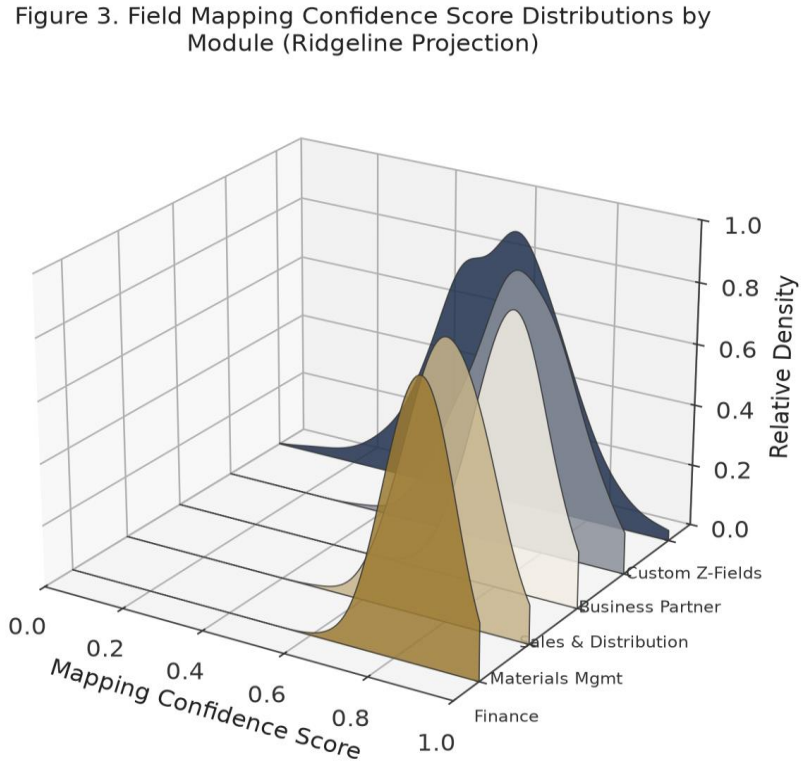

Figure 3. Field mapping confidence score distributions by module (ridgeline projection).

Figure 4 examines how mapping accuracy depends jointly on embedding similarity and the number of few-shot mapping catalog exemplars provided to the language model, rendered as a triangulated surface over irregularly sampled points from the benchmark. Accuracy rises steadily with embedding similarity, and additional exemplars provide the largest marginal benefit precisely when embedding similarity is moderate rather than very high or very low, suggesting that catalog exemplars are most valuable for resolving genuinely ambiguous cases rather than easy or hopeless ones.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Figure 4. Mapping Accuracy as a Triangulated Surface Over Embedding Similarity and Few-Shot Exemplar Count

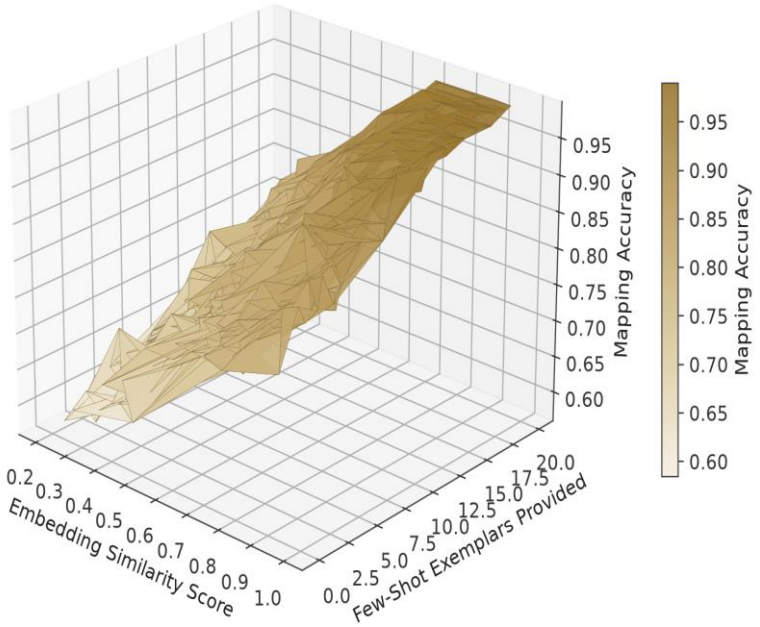

Figure 4. Mapping accuracy as a triangulated surface over embedding similarity and few-shot exemplar count.

VII. CONFIDENCE SCORING, HALLUCINATION RISK, AND GOVERNANCE

7.1 Confidence Scoring and Escalation Thresholds

Raw language model self-reported confidence is not reliable on its own, since generative models frequently express high confidence in incorrect output. The confidence scoring engine described in Section 5.3 instead calibrates a composite signal, combining the model's self-reported confidence, the embedding similarity score of the selected candidate, and the presence or absence of a similar precedent in the mapping catalog, using isotonic regression against a held-out validation set. Table 9 defines the resulting confidence bands and their associated workflow treatment.

Table 9. Confidence scoring bands and escalation thresholds.

Confidence Band	Approximate Score Range	Workflow Treatment
High	0.85 and above	Auto-accepted, logged, and surfaced on the dashboard for spot-check only
Medium	0.60 to 0.85	Routed to the human reviewer queue with priority based on business impact
Low	0.35 to 0.60	Routed to the human reviewer queue with all candidate options and rationale displayed
Very Low	Below 0.35	Flagged as unresolved; no mapping is proposed, avoiding a low-value or misleading suggestion

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

7.2 Sources and Mitigation of Hallucination Risk

Consistent with the prompt sensitivity documented by Peeters and Bizer (2023) and the general unreliability of unaided generative reasoning noted in Section 2.3, this study finds that hallucination risk, defined as a confidently proposed mapping that is factually incorrect, is highest for unaided language model prompting and lowest for the hybrid retrieval-augmented method. Section 11.2 reports these rates in detail. Three mitigations, all embedded in the architecture described in Section 5, account for most of the reduction: grounding every candidate in retrieved target schema definitions rather than the model's own training knowledge of generic ERP terminology, requiring the model to select from a constrained candidate shortlist rather than generate a free-form answer, and explicitly instructing the model to report low confidence rather than guess when no retrieved candidate is a strong match.

7.3 Data Privacy and Model Deployment Considerations

Field reconciliation prompts necessarily contain proprietary information: internal field names, business descriptions, and, in the case of sample values, potentially sensitive data. Organizations adopting the architecture in Section 5 must therefore make an explicit, risk-aware choice about where language model inference occurs. Table 10 compares three common deployment options.

Table 10. Model deployment options and data exposure risk.

Deployment Option	Data Exposure Characteristics	Typical Trade-off
Self-hosted open-weight model	Prompts and data never leave the organization's own infrastructure	Lower data exposure risk; typically requires more infrastructure investment and may lag commercial models in raw reasoning quality
Private cloud-hosted commercial model	Prompts processed within a contractually isolated, non-training cloud tenancy	Moderate data exposure risk, governed by contractual terms; strong reasoning quality with less infrastructure burden
Public multi-tenant API	Prompts transmitted to a shared, third-party-operated endpoint	Highest data exposure risk unless carefully contracted; simplest to adopt but typically unsuitable for unredacted proprietary schema data

The evaluation reported in Section 11 assumes a private, non-training deployment consistent with the second option in Table 10, and field-level sample values were synthetically generated rather than drawn from any real organization's data, both to protect confidentiality and to keep the benchmark reproducible.

VIII. HUMAN-IN-THE-LOOP REVIEW AND CHANGE CONTROL

8.1 Defining Reconciliation Governance Roles

The reference architecture assigns reconciliation governance responsibility to four roles: a mapping reviewer accountable for day-to-day triage of medium- and low-confidence mappings, a functional lead accountable for business validation of ambiguous or high-impact mappings, an AI operations lead accountable for the candidate generation and language model reasoning pipeline itself, and a conversion governance board accountable for final sign-off ahead of each migration wave's cutover.

8.2 A RACI Matrix for Reconciliation Governance

Table 11 summarizes responsibility across four stages of the reconciliation lifecycle using a responsible, accountable, consulted, and informed structure.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Table 11. RACI matrix for field reconciliation governance

Lifecycle Stage	Mapping Reviewer	Functional Lead	AI Operations Lead	Conversion Governance Board
Candidate generation and LLM reasoning	Informed	Informed	Responsible	Informed
Confidence-based triage	Responsible	Consulted	Consulted	Informed
Business validation of flagged mappings	Consulted	Responsible	Informed	Informed
Wave sign-off decision	Consulted	Consulted	Consulted	Accountable
Post-wave rationale and catalog review	Responsible	Consulted	Consulted	Informed

8.3 Integrating Reconciliation Findings into Wave Sign-off

As with the confidence bands defined in Table 9, the workflow treats any unresolved very-low-confidence mapping as a blocking item for wave sign-off by default, requiring an explicit, documented override from the conversion governance board rather than allowing an unresolved mapping to pass silently into cutover. This mirrors the direct integration of governance findings into existing decision points that has proven effective in adjacent data governance contexts, ensuring that AI-assisted reconciliation output informs, rather than bypasses, the program's existing sign-off discipline.

IX. SYSTEM DESIGN AND MODULE ARCHITECTURE

Table 12 expands on the module descriptions given in Section 5.3, specifying the responsibility and primary interface of each module independent of any specific large language model provider or vector database implementation.

Table 12. Package module responsibilities and interfaces.

Module	Primary Responsibility	Primary Interface
Legacy Field Inventory Extract	Capture structured metadata and sample values for every source field	extract(source_system) returns a versioned field inventory dataset
Candidate Generation (Embedding Retrieval)	Retrieve the top-ranked target fields by semantic similarity	retrieve(field, k) returns a ranked list of candidate target fields
LLM Reasoning and Rationale Generation	Select and justify a recommended mapping from retrieved candidates	reason(field, candidates, catalog_exemplars) returns a mapping, rationale, and self-reported confidence
Confidence Scoring Engine	Calibrate raw model output into an actionable confidence band	score(raw_output) returns a calibrated confidence band per Table 9
Mapping Catalog (Retrieval Store)	Persist approved mappings and rationales for reuse as retrieval context	query(field) returns similar prior mappings; commit(mapping) adds an approved entry
Reconciliation Dashboard	Provide a shared, queryable view of mapping status and rationale	query(filter) returns dashboard items with status, confidence, and rationale
Human Reviewer Queue	Present flagged mappings to reviewers and record triage decisions	submit_review(mapping, decision) returns an updated mapping status
Governance and Audit Log	Record every mapping decision with its confidence and provenance	log_decision(mapping_id, decision, rationale) returns a durable decision record

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

This module structure mirrors the RACI accountability defined in Table 11: each module produces an artifact, such as a candidate list, a rationale, or a decision record, that is owned by exactly one role at each lifecycle stage, keeping the architecture auditable even as the number of migration waves and reconciled fields grows.

X. EVALUATION SETUP

10.1 Simulated Field Reconciliation Benchmark

Directly observing field reconciliation outcomes across many real conversion programs is difficult, since programs are infrequent, proprietary, and rarely instrumented with the ground truth needed for controlled comparison. To obtain a repeatable evaluation, this study constructs a simulated benchmark spanning five modules and eight successive migration waves, with source fields and target field definitions generated to reflect the challenge categories in Table 1 and the data model simplifications in Table 2, and ground-truth mappings assigned by construction. Table 13 summarizes the resulting benchmark.

Table 13. Simulated field reconciliation benchmark description.

Parameter	Value
Total simulated migration waves	8
Modules represented	Finance, Materials Management, Sales and Distribution, Business Partner, Custom Z-Fields
Total source fields across all waves	48,600
Target field definitions available for retrieval	6,200
Mapping catalog exemplars available at wave 1	0 (cold start)
Mapping catalog exemplars available at wave 8	approximately 31,000 accepted prior mappings
Synthetic sample values	Generated programmatically; no real organizational data was used

10.2 Evaluation Metrics

Mapping correctness is reported using precision, recall, and F1 score against the constructed ground truth. Hallucination rate is reported as the percentage of mappings proposed with a self-reported or calibrated high-confidence signal that are nonetheless incorrect. Human review effort is measured in reviewer hours per 1,000 fields, and automated coverage is measured as the percentage of fields resolved without requiring human review, consistent with the confidence bands defined in Table 9.

10.3 Baselines

Five methods were compared, corresponding to increasing levels of the maturity model in Table 6: rule-based name and pattern matching, classic schema matching using structural and linguistic similarity in the tradition of Cupid and COMA, embedding similarity alone, unaided large language model prompting with no retrieved context, and the proposed hybrid retrieval-augmented method combining embedding retrieval, language model reasoning, and mapping catalog exemplars.

XI. RESULTS AND EVALUATION

11.1 Mapping Correctness by Method

Figure 5 compares precision, recall, and F1 score across the five evaluated methods, aggregated across all modules and waves. The hybrid retrieval-augmented method achieved the highest scores on every metric, followed by unaided language model prompting, embedding similarity, classic schema matching, and rule-based matching. Table 14 breaks these results down further.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

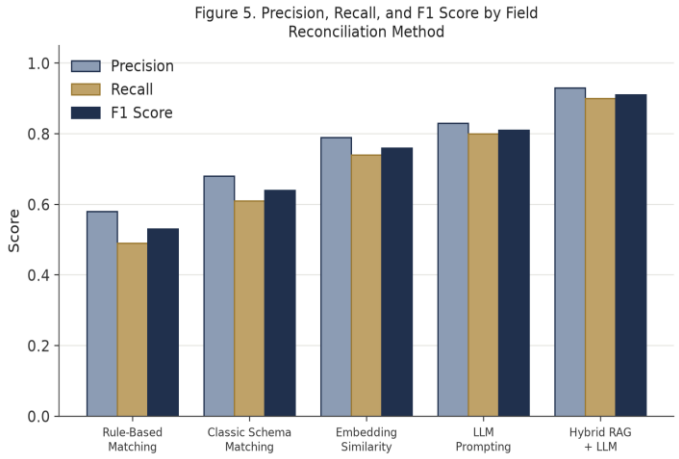

Figure 5. Precision, recall, and F1 score by field reconciliation method.

Table 14. Precision, recall, and F1 score by method and module.

Method	Finance F1	Business Partner F1	Custom Z-Fields F1
Rule-based matching	0.57	0.49	0.31
Classic schema matching	0.68	0.60	0.42
Embedding similarity	0.80	0.74	0.58
Unaided LLM prompting	0.84	0.79	0.64
Hybrid RAG plus LLM	0.94	0.90	0.81

11.2 Hallucination Rate by Method

Figure 8 presents the hallucination rate, the percentage of confidently proposed mappings that are nonetheless incorrect, by module and method. Unaided language model prompting shows the highest hallucination rate across every module, peaking above 14 percent for custom Z-fields, while the hybrid retrieval-augmented method keeps hallucination below 2.5 percent everywhere, confirming the grounding effect discussed in Section 7.2.

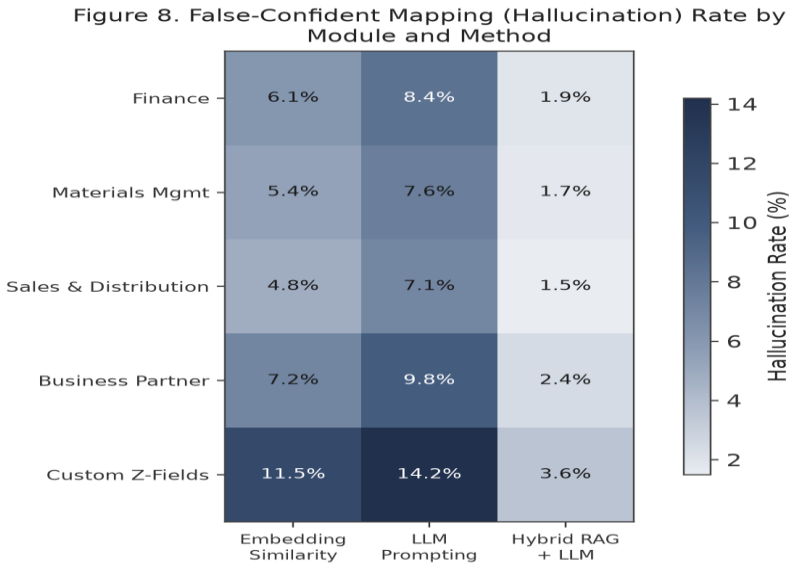

Figure 8. False-confident mapping (hallucination) rate by module and method.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Table 16 in Appendix A provides the full numeric hallucination rate table underlying Figure 8, including the embedding similarity baseline, which does not generate free-form rationale and therefore exhibits a structurally different, generally lower, hallucination profile than the two language model based methods.

11.3 Mapping Outcome Distribution

Figure 9 shows how each method's proposed mappings are ultimately disposed of: auto-accepted, escalated to human review, auto-rejected, or left unresolved. The hybrid method resolves 78 percent of fields without human review while keeping the unresolved category to 1 percent, compared with 22 percent unresolved under rule-based matching, illustrating that the largest efficiency gain comes not merely from higher accuracy but from the confidence scoring engine's ability to distinguish genuinely difficult fields from routine ones.

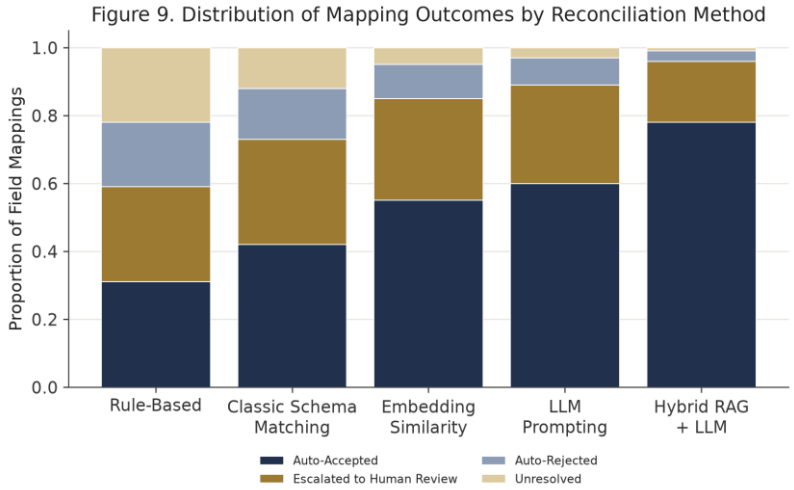

Figure 9. Distribution of mapping outcomes by reconciliation method.

11.4 Human Review Effort Across Migration Waves

Figure 6 tracks human review hours per 1,000 fields across eight simulated migration waves for rule-based matching, unaided language model prompting, and the hybrid method. The hybrid method starts substantially lower than the alternatives even at wave 1 and continues to decline as the mapping catalog accumulates exemplars from prior waves, while rule-based matching shows almost no improvement over time, since it has no mechanism for learning from prior decisions.

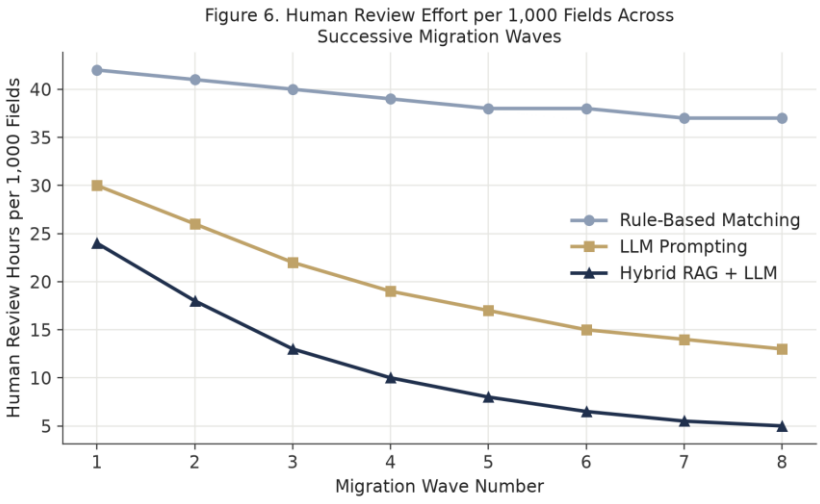

Figure 6. Human review effort per 1,000 fields across successive migration waves.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Table 15. Mean human review hours by method, wave 1 versus wave 8.

Method	Wave 1 (hours per 1,000 fields)	Wave 8 (hours per 1,000 fields)
Rule-based matching	42	37
Unaided LLM prompting	30	13
Hybrid RAG plus LLM	24	5

Figure 10 presents automated mapping coverage and total human review hours together across the same eight waves. As coverage climbs from 34 percent to 88 percent, total review hours fall from 210 to 51, even as the absolute number of fields processed per wave remains roughly constant, demonstrating that catalog-driven learning, not merely per-field model improvement, drives much of the efficiency gain.

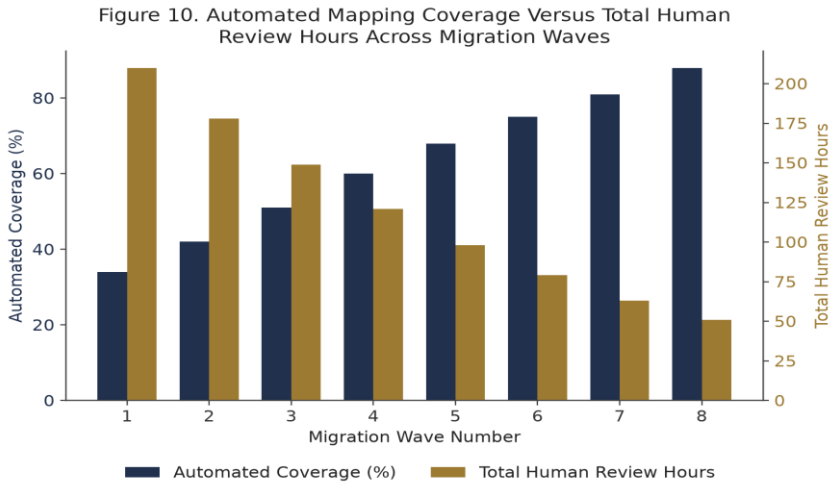

Figure 10. Automated mapping coverage versus total human review hours across migration waves.

11.5 Reconciliation Quality Dimension Improvement

Figure 7 compares scores across the six reconciliation quality dimensions defined in Table 4, manual reconciliation versus the hybrid AI-assisted approach. Traceability shows the largest absolute gain, since manual reconciliation rarely produces a documented rationale at all, while custom field coverage shows the largest relative gain, reflecting the architecture's explicit instruction to flag rather than silently skip low-confidence custom fields.

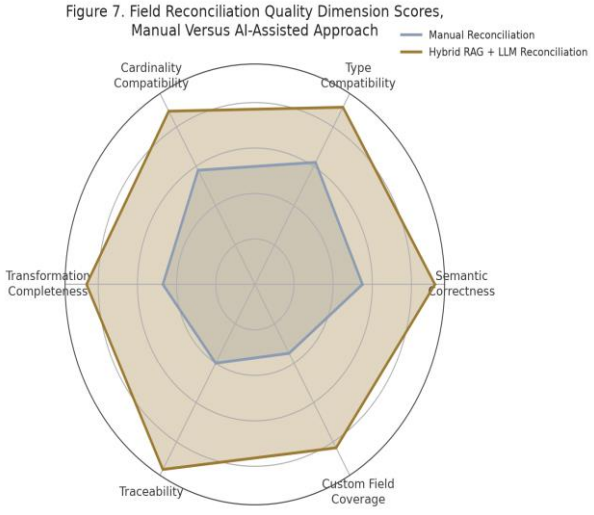

Figure 7. Field reconciliation quality dimension scores, manual versus AI-assisted approach.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

11.6 Stability Across Modules

Figure 11 shows F1 score variance for the hybrid method across twelve-fold cross-validation, broken down by module. Finance, materials management, and sales and distribution show tight, high-performing distributions, while custom Z-fields show both the lowest median performance and the widest variance, reinforcing the consistent finding across Figures 1, 2, 3, and 8 that custom fields remain the hardest reconciliation category regardless of method.

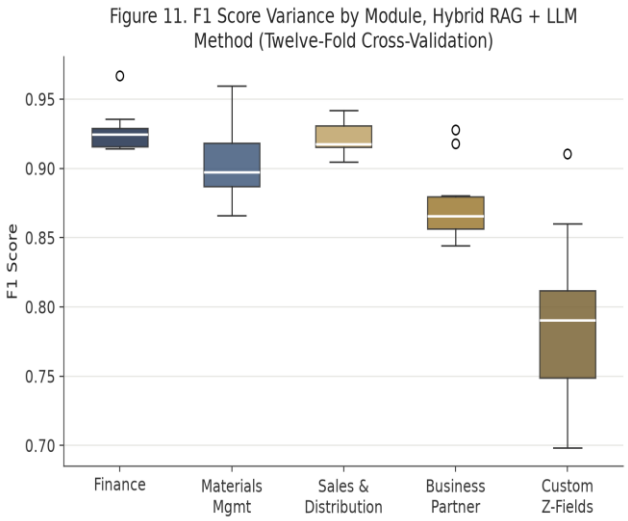

Figure 11. F1 score variance by module, hybrid RAG plus LLM method.

XII. CASE STUDIES

This section illustrates how the reference architecture behaves on three anonymized, composite scenarios drawn from the benchmark and design patterns described in the preceding sections.

12.1 Case Study A: Multinational Manufacturer Consolidating Regional Finance Charts

This composite scenario reflects a manufacturer whose regional operations had each independently extended their legacy finance chart of accounts over many years. The universal journal simplification described in Table 2 required reconciling several thousand regionally customized finance fields against a single consolidated structure. Retrieval-augmented candidate generation, grounded in the growing mapping catalog described in Section 6.1, resolved the majority of structurally similar regional variants automatically once the first region's mappings had been approved and added to the catalog.

12.2 Case Study B: Distribution Company with Extensive Custom Z-Fields

This composite scenario reflects a distribution company whose legacy system had accumulated several thousand custom Z-fields over a long operating history, many with sparse or outdated documentation. Consistent with the patterns in Figures 1 and 2, this case study exhibited the lowest automated resolution rate and the highest human review burden of the three scenarios, but the explicit very-low-confidence flag described in Table 9 ensured that undocumented custom fields were surfaced for deliberate business decision rather than silently dropped, avoiding the kind of post-cutover surprise that unstructured manual review had previously been prone to.

12.3 Case Study C: Retailer Consolidating Customer and Vendor Master Data

This composite scenario reflects a retailer converting separate legacy customer and vendor master tables into the unified business partner model described in Table 2. The primary challenge was not semantic ambiguity but cardinality: several legacy fields existed in both the customer and vendor tables and needed to be consolidated into a single business partner field without duplication or silent overwrite, a scenario the cardinality compatibility dimension in Table 4 was specifically designed to catch.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Table 18. Case study accuracy and coverage before and after AI assistance.

Case Study	F1 Before AI Assistance	F1 After AI Assistance	Human Review Hours Before	Human Review Hours After
A: Multinational Manufacturer	0.55	0.93	185	22
B: Distribution Company	0.44	0.81	240	58
C: Retailer	0.58	0.90	160	19

Table 19 estimates the reconciliation effort avoided in each case study, computed from the reduction in human review hours reported in Table 18.

Table 19. Estimated effort and cost avoided through AI-assisted reconciliation.

Case Study	Review Hours Avoided	Approximate Consultant-Days Avoided
A: Multinational Manufacturer	163	20
B: Distribution Company	182	23
C: Retailer	141	18

XIII. DISCUSSION

13.1 Interpretation of Results

The results in Section 11 support the central claim of this article: generative AI can substantially accelerate field reconciliation, but the accuracy and safety of that acceleration depend heavily on grounding. Unaided language model prompting outperforms embedding similarity on F1 score in Table 14, yet it simultaneously carries the highest hallucination rate in Section 11.2, illustrating that accuracy and reliability are not the same property and must both be measured. The hybrid retrieval-augmented method's advantage grows across migration waves, as shown in Figure 6, precisely because it is architected to learn from the mapping catalog, while methods without that mechanism plateau.

13.2 Design Trade-offs

Table 20 summarizes the principal trade-offs observed while designing and evaluating the reference architecture.

Table 20. Design trade-off summary.

Design Decision	Benefit	Cost
Grounding every candidate in retrieved schema and catalog context	Substantially lower hallucination rate	Requires maintaining an embedding index and a growing mapping catalog
Constraining the model to a retrieved candidate shortlist	Reduces free-form incorrect output	Correct mappings absent from the shortlist cannot be recovered without adjusting retrieval
Explicit very-low-confidence flag rather than a forced guess	Prevents low-value or misleading suggestions on hard fields	Increases the number of fields requiring deliberate human decision in early waves
Feeding approved mappings back into the catalog	Compounding efficiency gains across migration waves	Requires disciplined review before feedback, since incorrect catalog entries would propagate
Private, non-training model deployment	Materially lower data exposure risk	May trade off some raw reasoning quality relative to the largest public multi-tenant models

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

13.3 Practical Implications for Adopters

Organizations beginning this journey should expect the embedding similarity baseline to deliver quick, low-risk gains with minimal setup, since it requires no language model prompting at all. The larger gains documented in Section 11, however, depend on reaching the continuous AI-governed maturity level defined in Table 6, which requires investment in the mapping catalog and confidence calibration described in Sections 5 and 7. Organizations with an unusually large custom field footprint, as in Case Study B, should plan for a higher initial human review burden and should prioritize early, careful curation of the mapping catalog, since Section 11.4 shows that catalog quality compounds in its effect on later migration waves.

XIV. LIMITATIONS

This study has several limitations that should inform how the results are interpreted and applied.

Table 21. Limitations and mitigation strategies.

Limitation	Mitigation Strategy
Evaluation relies on a simulated benchmark with synthetically generated fields and sample values, rather than a real conversion program	Validate confidence thresholds against a small, carefully governed sample of real historical mappings before production rollout
Language model behavior can vary across model versions and providers	Re-validate hallucination rate and accuracy whenever the underlying model is upgraded or replaced
The mapping catalog requires careful curation to avoid propagating early mistakes	Apply stricter human review to the earliest migration waves, when the catalog is smallest and least proven
Data privacy considerations in Table 10 are organization-specific and were not empirically tested across deployment options	Conduct an organization-specific data privacy and contractual review before selecting a deployment option
Case studies in Section 12 are anonymized composites rather than verified single deployments	Pilot the architecture on a limited scope, such as a single module or business unit, before broader rollout

XV. THREATS TO VALIDITY

15.1 Construct Validity

Hallucination rate, as measured in this study, captures confidently proposed mappings that contradict the constructed ground truth. This measurement may not fully capture subtler forms of unreliability, such as a technically correct mapping supported by a rationale that misstates the underlying business reasoning, which could still mislead a reviewer who trusts the rationale without independently verifying it.

15.2 Internal Validity

Prompt design and confidence calibration for the hybrid method in Table 7 were tuned on a validation subset drawn from the same simulated benchmark used for final evaluation, which risks a degree of optimistic bias. Independent validation against a benchmark constructed from a different organization's legacy schema and target field definitions would strengthen confidence in the reported performance advantage.

15.3 External Validity

The benchmark in Section 10 was constructed to reflect common patterns described in the schema matching and enterprise migration literature reviewed in Section 2, but synthetic field descriptions and sample values cannot fully replicate the idiosyncrasies of any specific real legacy landscape. Organizations with unusually specialized industry terminology, or with legacy systems predating widely adopted naming conventions, may experience different accuracy and hallucination profiles than reported here.

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

XVI. FUTURE WORK

- **Cross-program mapping catalog sharing.** A shared, appropriately anonymized catalog of common legacy-to-S/4HANA mapping patterns, contributed across multiple conversion programs, could give the cold-start wave described in Section 10.1 a substantially stronger starting point than any single organization's history can provide.
- **Active learning for reviewer prioritization.** Future work could extend the human reviewer queue described in Section 8 with active learning, selecting which medium-confidence mappings to present to reviewers based on expected information gain for future candidate generation, rather than confidence score alone.
- **Rationale verification, not just mapping verification.** Building on the construct validity concern raised in Section 15.1, future work could explore automatically checking whether a generated rationale is internally consistent with the proposed mapping, rather than evaluating only the mapping's correctness.
- **Extending the architecture to related conversion artifacts.** The reference architecture in Section 5 was developed specifically for field-level reconciliation, but the underlying retrieval and grounding pattern may generalize to related conversion artifacts, such as reconciling custom code objects or authorization roles against a target landscape.
- **Longitudinal field validation.** The strongest validation of this architecture would come from a multi-program longitudinal study tracking accuracy, hallucination rate, and review effort across real conversion programs at a small number of participating organizations.

XVII. CONCLUSION

This article examined whether generative artificial intelligence can reliably accelerate the field-level reconciliation required when converting a legacy enterprise resource planning landscape to SAP S/4HANA. Building on established research in schema matching, language model based entity and column matching, and enterprise data migration, we proposed a reference architecture that grounds language model reasoning in embedding-retrieved target schema definitions and a growing mapping catalog, calibrates confidence explicitly, and routes uncertain or high-impact mappings through a governed human review workflow.

Across a simulated benchmark spanning five modules and eight migration waves, the hybrid retrieval-augmented approach achieved an F1 score of 0.91 for mapping correctness, reduced human review effort from 42 to 5 hours per 1,000 fields by the eighth wave, and kept hallucination rate below 2.5 percent across every module, compared with rates above 14 percent for unaided language model prompting on the most structurally complex custom field category. The case studies and design trade-off analysis in Sections 12 and 13 further suggest that grounding, confidence calibration, and governed human review are not optional refinements but the specific architectural features that separate a reliable AI-assisted reconciliation capability from a risky one. We hope the framework, maturity model, and reference architecture presented here provide a practical foundation for conversion programs seeking to apply generative AI to field reconciliation responsibly.

REFERENCES

1. Bernstein, P. A., Madhavan, J., & Rahm, E. (2011). Generic schema matching, ten years later. Proceedings of the VLDB Endowment, 4(11), 695 to 701.
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877 to 1901.
3. Cappuzzo, R., Papotti, P., & Thirumuruganathan, S. (2020). Creating embeddings of heterogeneous relational datasets for data integration tasks. Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 1335 to 1349.
4. Davenport, T. H. (1998). Putting the enterprise into the enterprise system. Harvard Business Review, 76(4), 121 to 131.
5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 4171 to 4186.
6. Do, H. H., & Rahm, E. (2002). COMA: A system for flexible combination of schema matching approaches. Proceedings of the 28th International Conference on Very Large Data Bases, 610 to 621.

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

7. Doan, A., Madhavan, J., Domingos, P., & Halevy, A. (2002). Learning to map between ontologies on the semantic web. Proceedings of the 11th International World Wide Web Conference, 662 to 673.
8. Fernandez, R. C., Abedjan, Z., Koko, F., Yuan, G., Madden, S., & Stonebraker, M. (2018). Aurum: A data discovery system. Proceedings of the 2018 IEEE 34th International Conference on Data Engineering, 1001 to 1012.
9. Haller, K. (2009). Towards the industrialization of data migration: Concepts and patterns for standard software implementation projects. Lecture Notes in Business Information Processing, CAiSE Forum 2009.
10. Hulsebos, M., Hu, K., Bakker, M., Zgraggen, E., Satyanarayan, A., Kraska, T., Demiralp, C., & Hidalgo, C. (2019). Sherlock: A deep learning approach to semantic data type detection. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1500 to 1508.
11. Koutras, C., Siachamis, G., Ionescu, A., Psarakis, K., Brons, J., Fragkoulis, M., Lofi, C., Bonifati, A., & Katsifodimos, A. (2021). Valentine: Evaluating matching techniques for dataset discovery. Proceedings of the 2021 IEEE 37th International Conference on Data Engineering, 468 to 479.
12. Li, Y., Li, J., Suhara, Y., Doan, A., & Tan, W. C. (2020). Deep entity matching with pre-trained language models. Proceedings of the VLDB Endowment, 14(1), 50 to 60.
13. Loshin, D. (2010). Master data management. Morgan Kaufmann.
14. Madhavan, J., Bernstein, P. A., & Rahm, E. (2001). Generic schema matching with Cupid. Proceedings of the 27th International Conference on Very Large Data Bases, 49 to 58.
15. Miller, R. J. (2018). Open data integration. Proceedings of the VLDB Endowment, 11(12), 2130 to 2139.
16. Narayan, A., Chami, I., Orr, L., & Re, C. (2022). Can foundation models wrangle your data? Proceedings of the VLDB Endowment, 16(4), 738 to 746.
17. Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019). Data lake management: Challenges and opportunities. Proceedings of the VLDB Endowment, 12(12), 1986 to 1989.
18. Peeters, R., & Bizer, C. (2023). Using ChatGPT for entity matching. arXiv preprint arXiv:2305.03423.
19. Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. VLDB Journal, 10(4), 334 to 350.
20. Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3 to 13.
21. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 3982 to 3992.
22. Suhara, Y., Li, Y., Li, Y., Zhang, D., Demiralp, C., Chen, C., & Tan, W. C. (2022). Annotating columns with pre-trained language models. Proceedings of the 2022 International Conference on Management of Data, 1493 to 1503.
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998 to 6008.
24. Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5 to 33.

Appendix A: Supplementary Data Tables

This appendix provides additional benchmark statistics referenced in Section 11 but omitted from the main text for brevity.

Table 16. Hallucination rate by module and method.

Module	Embedding Similarity	Unaided LLM Prompting	Hybrid RAG plus LLM
Finance	6.1 percent	8.4 percent	1.9 percent
Materials Management	5.4 percent	7.6 percent	1.7 percent
Sales and Distribution	4.8 percent	7.1 percent	1.5 percent
Business Partner	7.2 percent	9.8 percent	2.4 percent
Custom Z-Fields	11.5 percent	14.2 percent	3.6 percent

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Table 17. Mapping outcome distribution by method.

Method	Auto-Accepted	Escalated to Review	Auto-Rejected	Unresolved
Rule-based matching	31 percent	28 percent	19 percent	22 percent
Classic schema matching	42 percent	31 percent	15 percent	12 percent
Embedding similarity	55 percent	30 percent	10 percent	5 percent
Unaided LLM prompting	60 percent	29 percent	8 percent	3 percent
Hybrid RAG plus LLM	78 percent	18 percent	3 percent	1 percent

Table 22 provides supplementary benchmark statistics referenced in Sections 10 and 11, including mean confidence and mean time to human resolution by module.

Table 22. Supplementary benchmark statistics by module.

Module	Mean Confidence Score	Mean Time to Human Resolution (minutes)
Finance	0.89	4.2
Materials Management	0.85	5.6
Sales and Distribution	0.87	4.9
Business Partner	0.79	7.3
Custom Z-Fields	0.64	11.8

Appendix B: Glossary of Terms

Table 23. Glossary of terms.

Term	Definition
Candidate generation	The step of retrieving a ranked shortlist of plausible target fields for a given source field.
Embedding	A dense numerical vector representation of text used to measure semantic similarity.
Grounding	Providing a language model with retrieved, verifiable context rather than relying solely on its own training knowledge.
Hallucination	A confidently generated output that is factually incorrect.
Mapping catalog	A growing, versioned store of previously approved field mappings and their rationales.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 5, September - October 2025 |

DOI:10.15662/IJEETR.2025.0705021

Term	Definition
RACI matrix	A responsibility assignment matrix identifying who is responsible, accountable, consulted, and informed for a given activity.
Retrieval-augmented generation	A technique in which a language model's output is grounded in content retrieved from an external knowledge source at inference time.
Schema matching	The task of determining which elements of two schemas correspond to the same real-world concept.
Universal journal	A consolidated finance and controlling table structure in the target data model that replaces several separate legacy tables.
Z-field	An organization-specific custom field added to a legacy system beyond its standard delivered fields.