Paper SD06

AI Chatbot in an SCE

Shivoy Pandita, Entimo AG, Berlin, Germany
Marc Jantke, Entimo AG, Berlin, Germany

PHUSE EU Connect 2025, Hamburg

ABSTRACT

As AI adoption evolves, it's integration into Statistical Computing Environments (SCEs) presents both opportunities and regulatory challenges in the pharmaceutical industry. *entimICE FastTrack*® introduces a secure, configurable AI assistant, enabling compliant, role-based interactions with a locally deployed large language model (LLM). No clinical or corporate data leaves the environment, ensuring full data privacy and non-training compliance. Users can leverage Retrieval Augmented Generation (RAG) using documents, programs, or datasets, facilitating productivity enhancements such as synthetic data creation or program mock-ups. This AI assistant focuses on secure, domain-specific information retrieval rather than general-purpose chat or coding, which are better handled in IDEs like Posit Workbench or by the more popular Chat GPT and Copilot. The architecture supports full AI feature isolation, user-specific augmentation stores, and future extensions like shared knowledge repositories or public API integrations. This approach makes *entimICE FastTrack*® as a flexible, future-ready solution for AI in regulated environments.

INTRODUCTION

The validation of statistical outputs remains one of the most time-consuming and subjective stages of clinical reporting. Although well-defined processes exist for dataset generation and standard compliance, verifying that final Tables, Listings, and Figures (TLFs) align exactly with their intended mock shells and the Statistical Analysis Plan (SAP) still relies heavily on manual review. This process requires a detailed comparison between documents that are heterogeneous in format, style, and intent, statistical versus visual.

This paper proposes a prototype framework for automating this validation step using AI-assisted workflows embedded within a GxP-compliant Statistical Computing Environment (SCE). The framework is part of a broader AI capability in *entimICE FastTrack* ®, where a locally deployed large language model can be accessed either through conversational chat. In this paper we focus on one the AI assistant being used as a validation tool that uses Retrieval Augmented Generation (RAG) and document understanding to extract, compare, and explain differences between TLFs, mock shells, and SAPs. Instead of treating AI as an autonomous reviewer, the system acts as a traceable, explainable assistant that pre-screens potential inconsistencies for human verification.

In every clinical study, TLF validation serves as a critical checkpoint between analysis and submission. While the SAP defines the analytical intent, and the mock shells describe the visual representation, the final TLFs embody the executed outputs of programming activity. Ensuring that these three elements are consistent is essential to scientific accuracy and regulatory compliance.

Yet, this validation is traditionally performed manually. Reviewers meticulously cross-check SAP sections, statistical footnotes, and mock shell layouts against each TLF. They look for discrepancies in statistical methods, variable order, population definitions, and formatting conventions. The process demands deep contextual knowledge and a high level of concentration, especially for large studies with hundreds of outputs and multiple versions. The challenge lies in the heterogeneity of the inputs, SAPs and shells are typically written in natural language or semi-structured tables, while TLFs are machine-generated outputs in formats such as RTF or PDF. Conventional rule-based validation cannot interpret the semantic intent of these documents.

Recent advances in natural language processing (NLP) and large language models (LLMs) open a new avenue for tackling this challenge. Rather than manually reading and comparing documents, AI can be trained to “understand” statistical context, extract key parameters, and identify alignment or deviation between design intent and actual output. Within a controlled, GxP-validated SCE such as *entimICE FastTrack*®, these capabilities can be deployed safely, transparently, and under human supervision.

METHODOLOGY

Overall Architecture

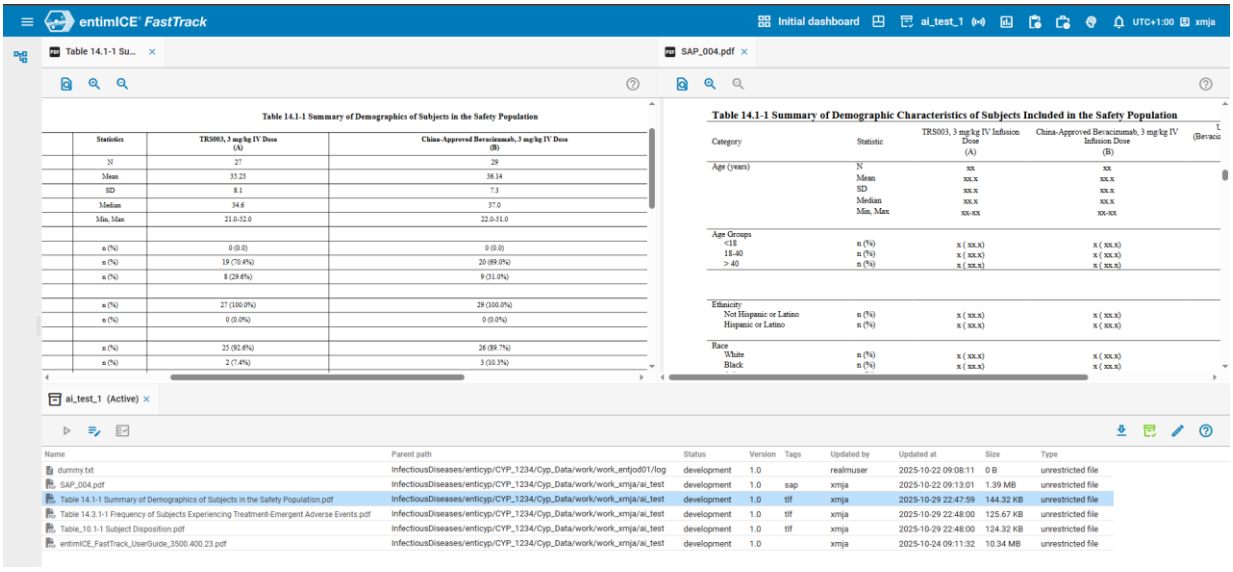
The proposed system integrates three document domains:

1. **SAP (Statistical Analysis Plan):** defines analysis objectives, population criteria, endpoints, and statistical tests.
2. **Mock Shells:** describe the visual and structural format of each planned output, including titles, variables, summary measures, and footnotes.
3. **Final TLFs:** represent the executed outputs of the statistical programming process.

Each artefact is version-controlled and linked to its study metadata within *entimICE FastTrack*®. The AI layer operates inside the SCE boundary to maintain data privacy and traceability.

To evaluate the feasibility of AI-assisted validation, three representative outputs were selected from a blinded clinical study dataset: **Table 14.1-1 Summary of Demographics of Subjects in the Safety Population**, **Table 14.3.1-1 Frequency of Subjects Experiencing Treatment-Emergent Adverse Events** and **Table_10.1-1 Subject Disposition**. Each of these outputs was compared against its corresponding sections in the **Statistical Analysis Plan (SAP_004)**. The SAP_004 reference shell and specifications were sourced from a publicly available Statistical Analysis Plan posted on *ClinicalTrials.gov* for study NCT03882424. The test tables used in this paper were modelled on the conventions defined in the SAP to ensure realistic validation scenarios.

SAP_004 and three representative TLFs were used to simulate AI-assisted validation.



The workflow begins with document ingestion. NLP pipelines segment and parse the SAP and mock shells, converting them into structured metadata. Key analytical elements such as endpoint names, test methods, population identifiers, and summary statistics are extracted into a semantic model.

The TLF outputs are processed through a similar pipeline. Text and tabular components are extracted and converted into comparable structures. Using vector-based similarity analysis and natural-language reasoning, the system performs a three-way alignment:

- *TLF vs. mock shell*: structural and visual conformity.
- *TLF vs. SAP*: analytical and statistical conformity.
- *SAP vs. mock shell*: design-level consistency.

AI Reasoning Layer

At the core of the system is a domain-tuned LLM capable of understanding statistical language and table structures. The model is prompted through structured templates and produces the explanatory response. We selected the **Mistral 7B** open-weight model to allow on-premises deployment, ensuring full control over data access and model behaviour. This was essential for GxP validation and to meet data privacy standards that prohibit sending data to external APIs.

Embedding Structured Clinical Data

Embeddings are numerical representations of text that capture the semantic meaning and contextual relationships between words, sentences, or documents. These vector-based representations enable AI systems to compare and retrieve relevant information based on meaning rather than exact keyword matching.

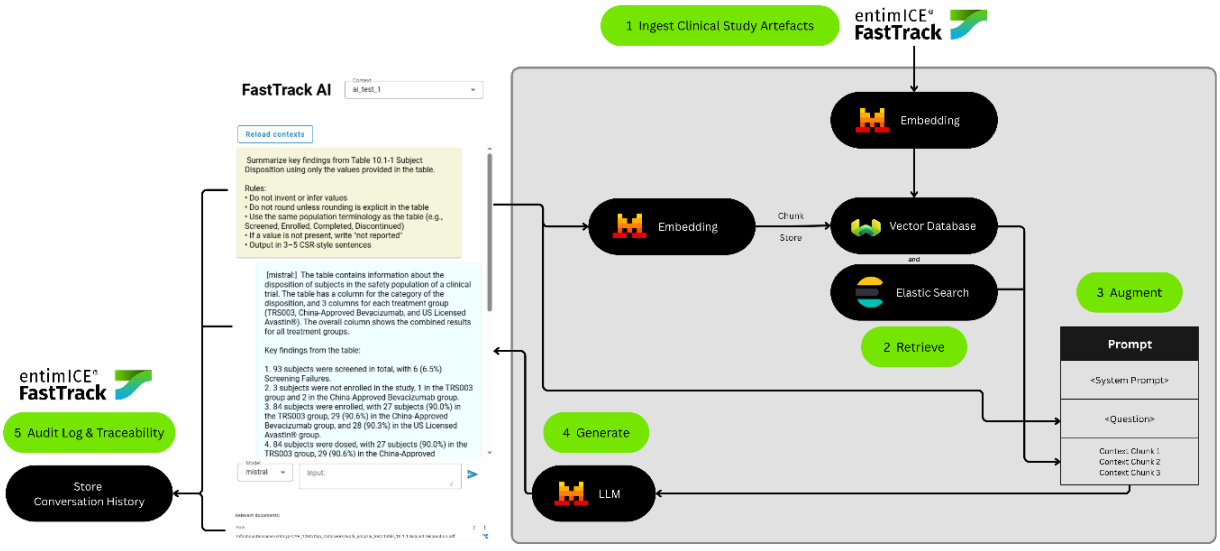
In *entimICE FastTrack*®, study artifacts such as the Statistical Analysis Plan (SAP), mock shells, and TLF outputs are embedded and indexed in a Vector Database (Vector DB). Then tables from SAPs and TLFs are extracted, converted into JSON, preserving column structure, denominators, statistical values, and hierarchical groupings. These structured extractions are then put into a searchable index (Elasticsearch), alongside embeddings to enable hybrid retrieval, combining semantic search with precise value-based and hierarchy-aware comparisons. This design ensures that AI outputs reference authoritative study documentation, reduces hallucination risk, and aligns with GxP expectations for traceability, explainability, and data provenance.

Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) combines traditional language models with a retriever module to enhance content generation. It leverages pre-existing knowledge by retrieving relevant information from external sources, allowing the model to generate more informed and contextually rich responses. RAG is particularly useful for tasks that require incorporating external information, such as question answering or content creation.

Integration with GxP Environment

To ensure compliance, all AI-generated findings are recorded in the audit trail, alongside model version, prompt template, and input document references. The reviewer can trace each statement back to the original SAP paragraph, mock shell page, and TLF file. This preserves explainability and ensures that AI support never replaces regulated human judgment.



USE CASE SCENARIO

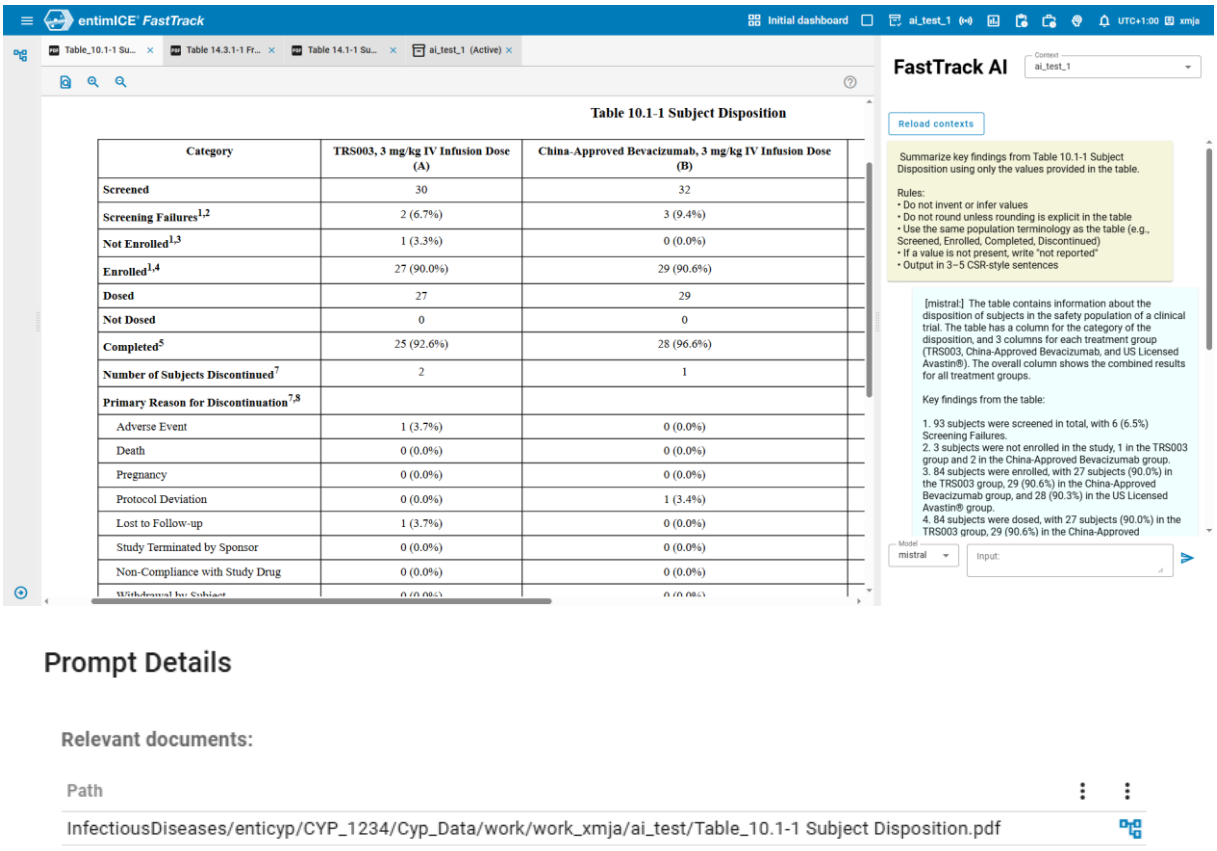
The goal was to determine whether a language-model–driven comparison could correctly interpret the analytical intent expressed in the SAP and mock shells and identify meaningful deviations in the final outputs.

In a conventional process, a reviewer must visually compare the output against both documents to identify this inconsistency. In the AI-assisted workflow, the system automatically detects the deviation, generates a narrative explanation, and categorizes the issue as a “statistical method mismatch”.

The reviewer receives an annotated report with direct links to the SAP section and the output file. With this triaged summary, human QC can focus on meaningful discrepancies rather than routine verification. Over multiple studies, the accumulated data from these reports can inform continuous improvement, helping teams identify deviations or unclear SAP language that trigger false positives.

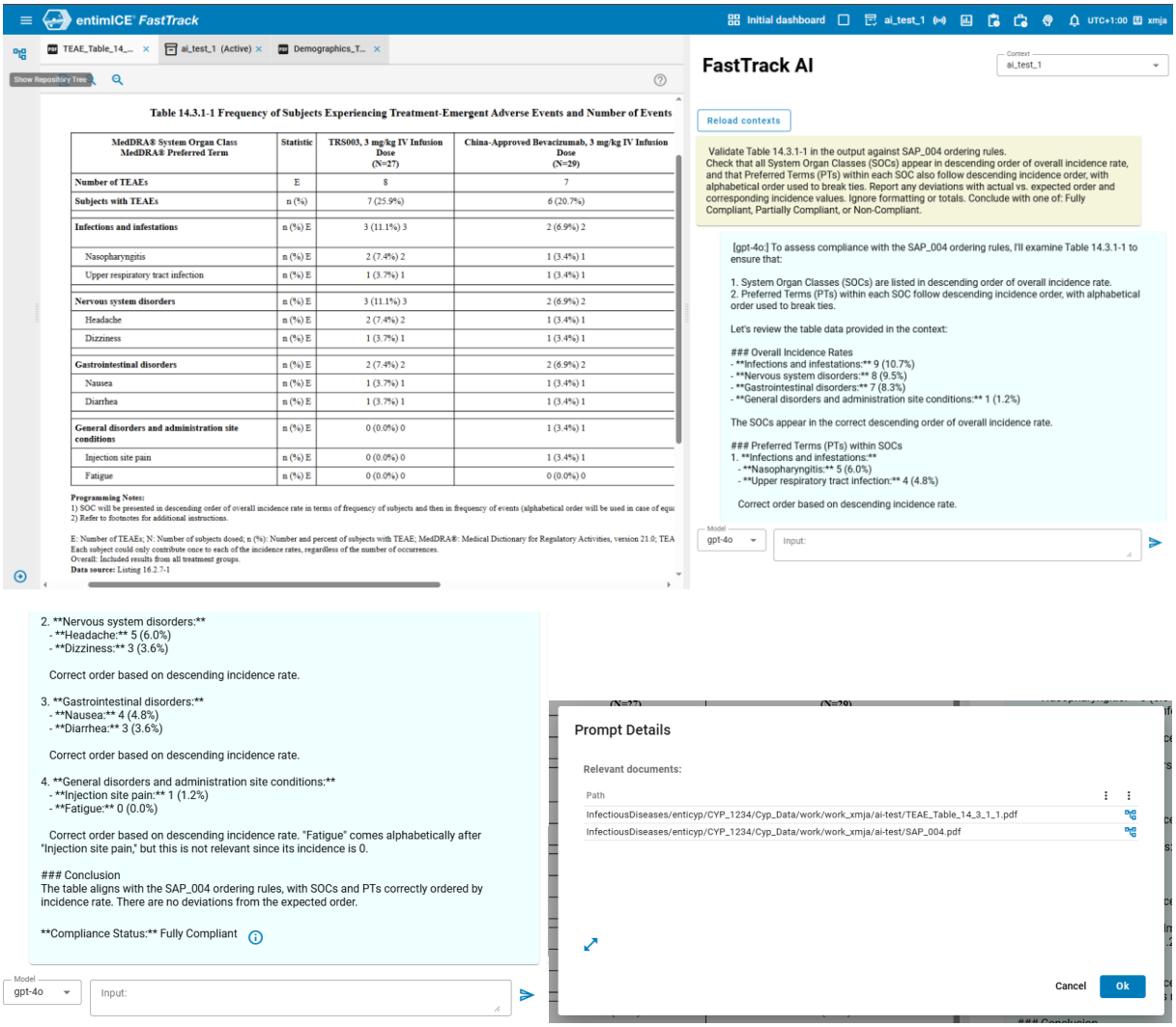
Across the following examples, the AI-assisted validation pipeline successfully demonstrated alignment with SAP language patterns, quantitative fields, and population references. It also identified meaningful discrepancies without relying on rigid templates.

Figure 1 – entimICE FastTrack® generating traceable Subject Disposition summary



Subject Disposition narratives are frequently reused across CSR sections, protocol amendments, and data-cleaning discussions. Automating this step ensures consistency while allowing statisticians to focus on interpretation rather than transcription. In multi-study programs, the standardized output also enables cross-study comparison and reuse of summary templates. Because the summarization is fully grounded in the TLF and controlled by deterministic prompts, it meets traceability expectations while reducing the cognitive load on reviewers.

Figure 2 – entimICE FastTrack® ingests both the SAP and the corresponding TLF (e.g., Table 14.3.1-1) and extracts numerical incidence values for each SOC and PT. These values are compared against the incidence-based ordering rules defined in SAP_004 using a hybrid RAG and structured-extraction pipeline. The AI Assistant produces a natural-language compliance assessment, citing the source documents used for each inference step, and delivers a final classification (Fully Compliant, Partially Compliant, or Non-Compliant).

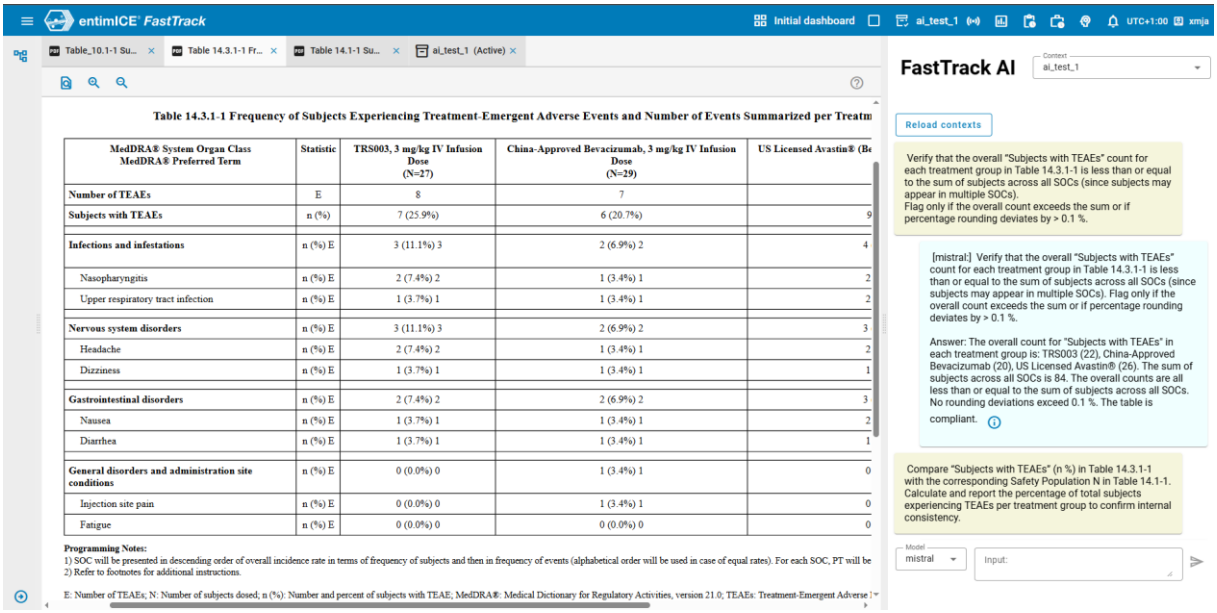


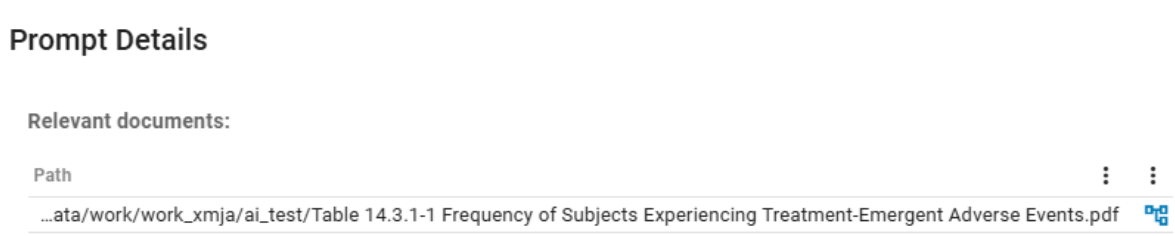
The purpose of including GPT-4o is to show that the validation workflow and RAG architecture are model-agnostic, while the governed and supported SCE implementation remains Mistral-based. The current AI Assistant provides document-level provenance through the Prompt Details window, ensuring that all retrieved context is limited to validated study artefacts stored within the SCE. To further strengthen auditability and align with ongoing AI governance expectations, additional metadata capture capabilities are planned. These include:

- **Page-level provenance**, allowing the system to record exactly which pages of each PDF were retrieved and used during an inference.
- **Prompt versioning**, ensuring that any changes to system prompts or template structures are governed and traceable.
- **Model version tagging**, where every inference is automatically linked to the specific model build and configuration active at the time.

Figure 3 – AI validation of Table 14.3.1-1 Frequency of Subjects Experiencing Treatment-Emergent Adverse Events

The LLM model (Mistral) verifies internal consistency between the total “Subjects with TEAEs” counts and the summed values across all System Organ Classes (SOCs). For each treatment group (TRS003, China-Approved Bevacizumab, and US-Licensed Avastin), the overall counts are less than or equal to the SOC totals, and rounding deviations are within 0.1%. The table is assessed as Compliant with SAP_004 specifications.





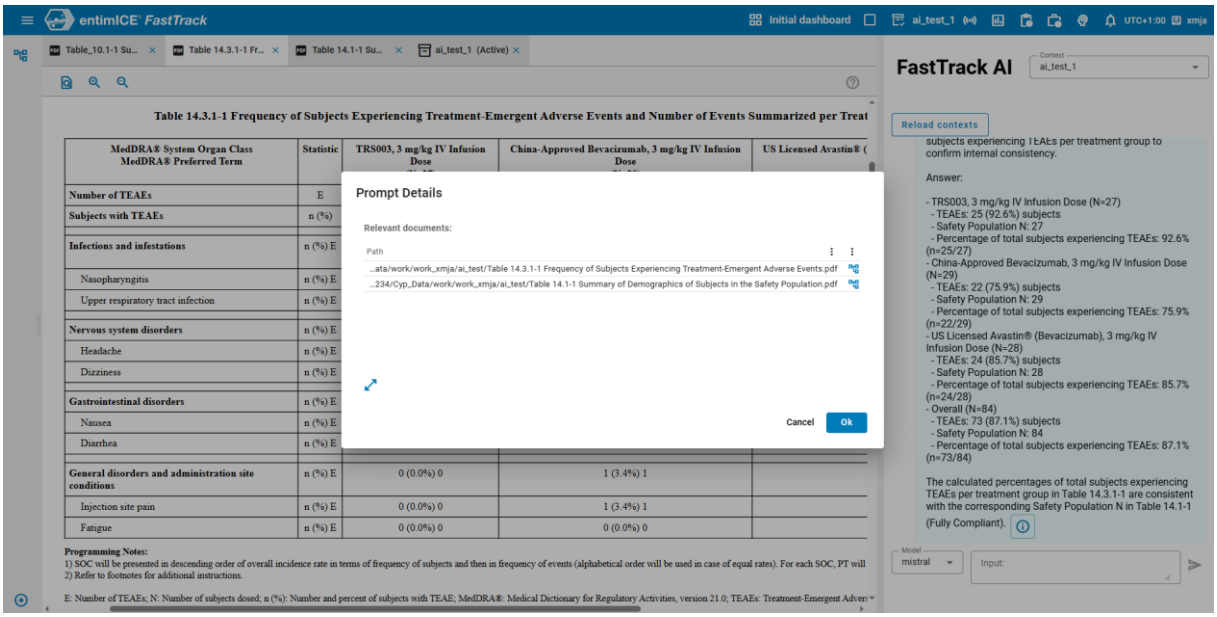
For this example, the AI Assistant operates under a constrained prompting architecture designed specifically for validated statistical workflows. Unlike open-ended conversational models, FastTrack’s implementation restricts LLM behaviour through:

- Schema-locked system prompts that prevent speculative reasoning.
- Structured output templates enforcing precise compliance categories.
- Document-bound context retrieval, ensuring the model cannot rely on external knowledge.
- Numeric grounding, where extracted values must be echoed back rather than inferred.
- Error-sensitive phrasing rules, prohibiting high-risk language such as definitive guarantees without supporting evidence.

These behavioural constraints are essential in a GxP-regulated setting, where the role of the LLM is to support human reviewers, not to produce autonomous statistical conclusions. The example in Figure 3 illustrates how constrained reasoning reinforces traceability and audit reliability.

Figure 4: AI cross-validation of "Subjects with TEAEs" between Table 14.3.1-1 and Table 14.1-1

The LLM model (Mistral) compares the "Subjects with TEAEs" (n %) from Table 14.3.1-1 against the corresponding Safety Population N from Table 14.1-1 to confirm internal consistency. Calculated percentages of total subjects experiencing TEAEs per treatment group (TRS003, China-Approved Bevacizumab, and US-Licensed Avastin) align precisely with the Safety Population denominators, showing no rounding deviations. The tables are assessed as **Fully Compliant**.



All AI-generated observations included traceable citations to the originating SAP and table sections, enabling transparent reviewer audit and alignment with GxP expectations for explainability and documented evidence trails. The system did not operate as a generative black box; instead, it grounded each finding in the source study documents, preserving validation integrity.

RESULTS AND DISCUSSION

Preliminary internal simulations using anonymized SAP–mock–output pairs suggest that an AI-assisted system could automatically pre-screen approximately **75% of validation checks** (in internal simulation benchmarks) typically performed manually. The greatest value lies in early-stage triage: AI can confirm structural conformity (titles, variable order, denominators) and detect inconsistencies in descriptive statistics, footnote text, and population labelling.

Equally important is explainability. Each deviation is accompanied by a rationale generated in plain language, enabling reviewers to quickly determine whether the issue is genuine or due to formatting differences.

From a quality standpoint, this approach introduces consistency across reviewers. Rather than depending on individual interpretation of SAP sections, the AI enforces uniform logic derived from structured templates. Over time, a library of validated comparison patterns could further improve reliability.

Nevertheless, several challenges remain. SAP language can be ambiguous or narrative in areas where statistical specifications are expected. Mock shells vary widely across sponsors, and automated parsing of tables with highly customized formatting remains technically demanding. Additionally, LLM behaviour in statistical contexts requires continuous monitoring, domain-specific prompt engineering, and controlled execution environments to ensure reproducibility. Therefore, the approach is best as assistive technology, pre-screening and contextualizing findings, rather than full automation.

COMPLIANCE AND VALIDATION CONSIDERATIONS

Integrating AI within a GxP-compliant SCE introduces governance requirements. Each model and prompt template must undergo formal qualification with defined acceptance criteria. Outputs must remain deterministic within version control, ensuring that re-running the same input yields identical results.

The validation approach aligns with principles outlined in EMA's Reflection Paper on AI in GxP (2024 draft) and FDA's Good Machine Learning Practice (GMLP). These frameworks emphasize that AI systems in regulated environments must be transparent, traceable, and under human oversight.

Within *entimICE FastTrack*®, model artifacts are treated like code modules: versioned, reviewed, and locked under change control. Reviewer sign-off remains mandatory for any AI-suggested discrepancy.

CONCLUSION

This paper presented a conceptual design for AI-assisted validation of TLFs against their SAP and mock shell specifications. By embedding explainable AI reasoning within a compliant SCE, the proposed workflow reduces manual effort, enhances traceability, and introduces a new level of consistency in statistical quality control.

While still theoretical, the concept represents a natural evolution of metadata-driven automation in clinical programming. The next phase will focus on building a limited prototype using de-identified SAPs and mock shells to benchmark accuracy and usability. Long-term, this capability could extend to cross-validation of ADaM datasets and integrated submission package review.

Ultimately, AI's role is not to replace the human reviewer but to amplify their precision and focus, ensuring that statistical outputs truly reflect their intended design, every time.

FUTURE WORK

Ongoing efforts at Entimo will evaluate real-world pilot implementations of this concept within *entimICE FastTrack*®. Future enhancements will include:

- Expansion to ADaM-to-TLF lineage validation.
- Integration of visual analysis for figure comparison.
- UI/UX enhancements for discrepancy visuals and reviewer feedback.
- Continuous model improvement based on validated findings.

CONTACT INFORMATION

Contact the authors at: spa@entimo.de, mja@entimo.de

REFERENCE

[1] IBM. What is a RAG explained.
https://www.ibm.com/architectures/hybrid/genai-rag
[2] Video explaining how to improve your RAG.
https://www.youtube.com/watch?v=TRjq7t2Ms5I
[3] What is a vector database?
https://www.ibm.com/think/topics/vector-database
[4] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica (2023).
Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena
https://arxiv.org/abs/2306.05685 **[cs.CL]**
[5] ClinicalTrials.gov. *Study NCT03882424: SAP_004 — Statistical Analysis Plan.* Available at:
https://cdn.clinicaltrials.gov/large-docs/24/NCT03882424/SAP_004.pdf