Paper ML11

Beyond the Hype:
Practical guidance for LLM adoption in Statistical Programming

Eyal Wultz, Bioforum, Tel Aviv, Israel
Bremer Louw, Bioforum, Bloemfontein, South Africa

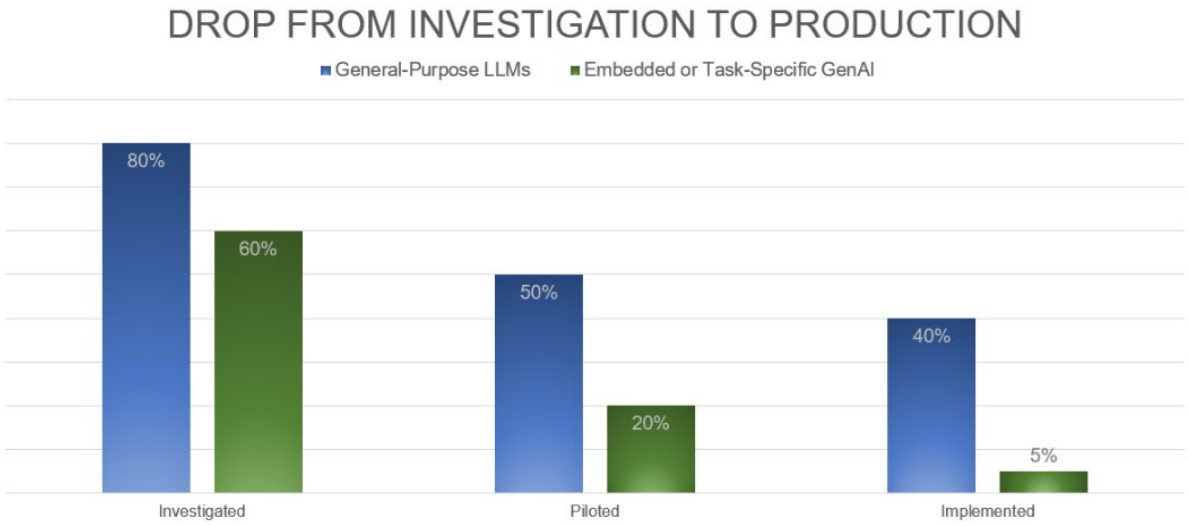
ABSTRACT

As organizations begin integrating Large Language Models (LLMs) into their data services, many struggle to move beyond experimentation and achieve reliable production use. This paper provides practical guidance for adopting LLMs in statistical programming environments, with a focus on embedded, task-specific systems rather than general-purpose tools. Drawing on real-world implementation experience, we discuss key challenges related to context management, performance, quality assurance, and data privacy. We further identify areas within the statistical programming lifecycle – such as controlled terminology mapping and SDTM variable mapping – where LLMs have demonstrated measurable and repeatable value. The goal of this paper is to support teams in adopting LLMs in a manner that balances innovation with operational rigor.

INTRODUCTION

According to *The GenAI Divide: State of AI in Business 2025 by MIT*[1], 95% of organizations report no measurable return on their GenAI investments. Although more than 80% of companies have evaluated general-purpose tools such as ChatGPT, Claude, and Gemini, only a small fraction have progressed embedded or task-specific GenAI systems beyond pilot stage, and just 5% have reached production.

According to the report, general-purpose LLMs that boost individual productivity – such as ChatGPT, Claude, and Gemini – have been explored by more than 80% of companies, and nearly 40% report deploying them. In contrast, adoption of embedded or task-specific GenAI systems lags significantly: only 60% of organizations have evaluated such tools, about 20% have reached the pilot stage, and just 5% have reached production.



The GenAI Divide: State of AI in Business 2025[1]

Healthcare and life sciences are among the industries where GenAI has not yet led to structural disruption. This gap does not appear to be driven by regulation or model capability, but rather by implementation approach. In this paper, we outline practical considerations for moving from experimentation to production with task-specific LLM systems, informed by our experience designing, piloting, and deploying such solutions in statistical programming workflows.

LIMITATIONS OF LARGE LANGUAGE MODELS

General-purpose LLMs are widely adopted for low-risk productivity tasks but present several limitations that restrict their use in mission-critical statistical programming.

QUALITY CONCERNS

LM output quality can vary significantly, particularly for tasks requiring complex internal logic or deep domain expertise. While acceptable for exploratory or assistive use, this lack of repeatability limits adoption in regulated environments where accuracy, traceability, and auditability are essential. As a result, human-in-the-loop validation remains unavoidable.

MEMORY AND CONTINUOUS LEARNING

Although some LLM platforms offer memory features, these mechanisms cannot be relied upon to consistently improve accuracy over time for high-stakes workflows. Mission-critical applications require explicit mechanisms to capture, validate, and reuse approved outputs in a controlled manner.

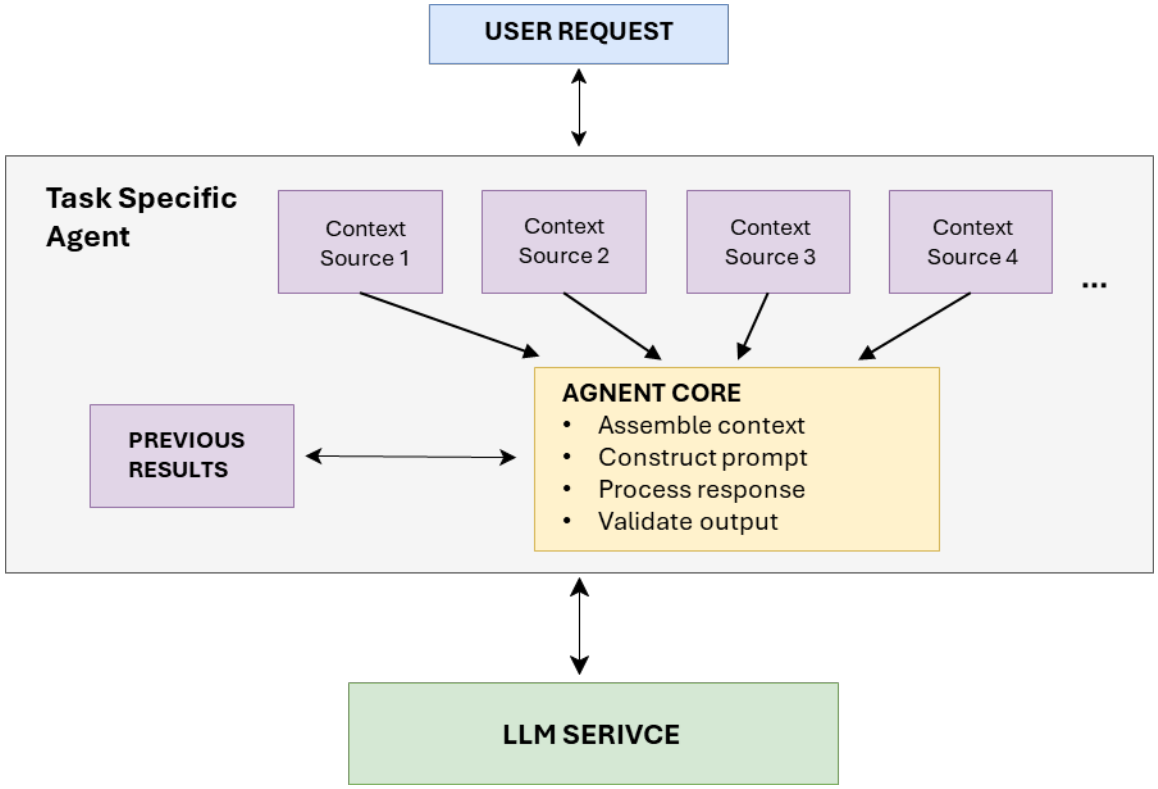
ADAPT TO CONTEXT

General-purpose LLMs rely heavily on users to manually provide detailed context for each task. This approach is inefficient and error-prone. In contrast, task-specific systems can automatically derive context from data, metadata, standards, and workflow state.

SOLUTION APPROACH – TASK-SPECIFIC SYSTEMS

Task-specific systems embed LLM capabilities directly into specialized applications. In this architecture, an AI Agent orchestrates interactions with the LLM by assembling relevant context, constructing optimized prompts, validating responses, and presenting recommendations within the existing workflow.

The following diagram illustrates this general architecture:



Users interact only with the application interface; the LLM operates transparently in the background.

As discussed in the MIT paper, To deliver meaningful value in mission-critical environments, such systems should:

- Focus on narrow, high-value use cases
- Integrate deeply into existing workflows
- Scale through explicit continuous learning mechanisms
- Minimize configuration overhead
- Demonstrate fast, measurable impact

SDTM CONVERSION USE CASES

To illustrate the concepts discussed above, we examine two common use cases encountered during the conversion of clinical data into SDTM-compliant datasets. Bioforum has developed an in-house SDTM conversion platform, JetConvert™, designed to increase automation while reducing conversion timelines, improving quality, and providing full transparency across the conversion process.

The two selected use cases employ task-specific LLMs embedded directly within JetConvert™ to generate conversion recommendations, enabling users to make faster and more accurate decisions. These use cases are evaluated against the adoption criteria outlined above. The first criterion is discussed jointly for both use cases, while the remaining four are assessed separately for each. It is important to note that in all cases the LLMs provide recommendations only; no conversion actions are performed autonomously, and final decisions remain under Human-in-the-Loop control.

FOCUS ON NARROW HIGH-VALUE USE CASES

Converting clinical source data from multiple inputs into a coherent, SDTM-compliant dataset is a complex and continuously evolving task. It requires intricate internal logic and deep domain expertise, and therefore constitutes mission-critical work. As such, it is poorly suited to general-purpose LLMs and instead requires a custom, embedded, task-specific solution.

Even if a general-purpose LLM could generate accurate SDTM mappings in isolation, using it for full study conversion would remain impractical. A typical study involves hundreds of variables and values, each requiring extensive context—such as source data characteristics, CRF annotations, SDTM version-specific guidance, sponsor conventions, and industry best practices. Manually assembling this context and reintegrating results into conversion tools would be

inefficient and error-prone. Moreover, general-purpose LLM workflows offer no systematic mechanism to reuse approved mappings or guidelines across studies, limiting scalability and long-term quality improvement.

USE CASE 1: CONTROLLED TERMINOLOGY RECOMMENDATIONS

BACKGROUND

Controlled terminology mapping involves converting raw source values into SDTM-compliant terms. Manually performing this task across thousands of values per study is repetitive, time-consuming, and prone to inconsistency.

WORKFLOW INTEGRATION

JetConvert™ addresses this challenge through a specialized Controlled Terminology Agent that applies a three-tier strategy:

1. Reuse of previously approved mappings
2. Deterministic string matching against codelists
3. LLM-based inference for genuinely ambiguous cases

This approach prioritizes speed and validated results while reserving LLM usage for scenarios where it adds clear value.

In JetConvert™, all source data is uploaded, and mappings are created to convert the source data to SDTM. As part of the mapping, the user can perform various textual modifications of which one is Controlled Terminology Mapping. This module of the software interacts with a specialized Controlled Terminology LLM Agent to provide mapping recommendations, supplemented with a confidence ranking, and present it back to the user for consideration.

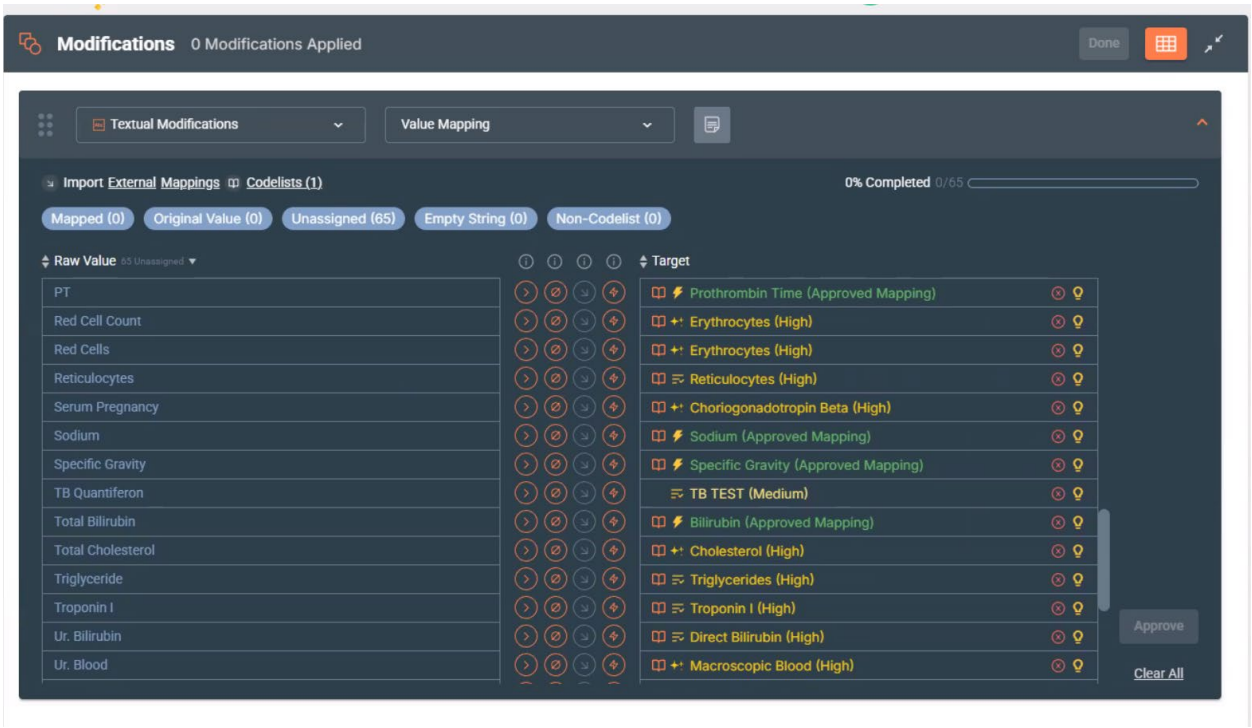
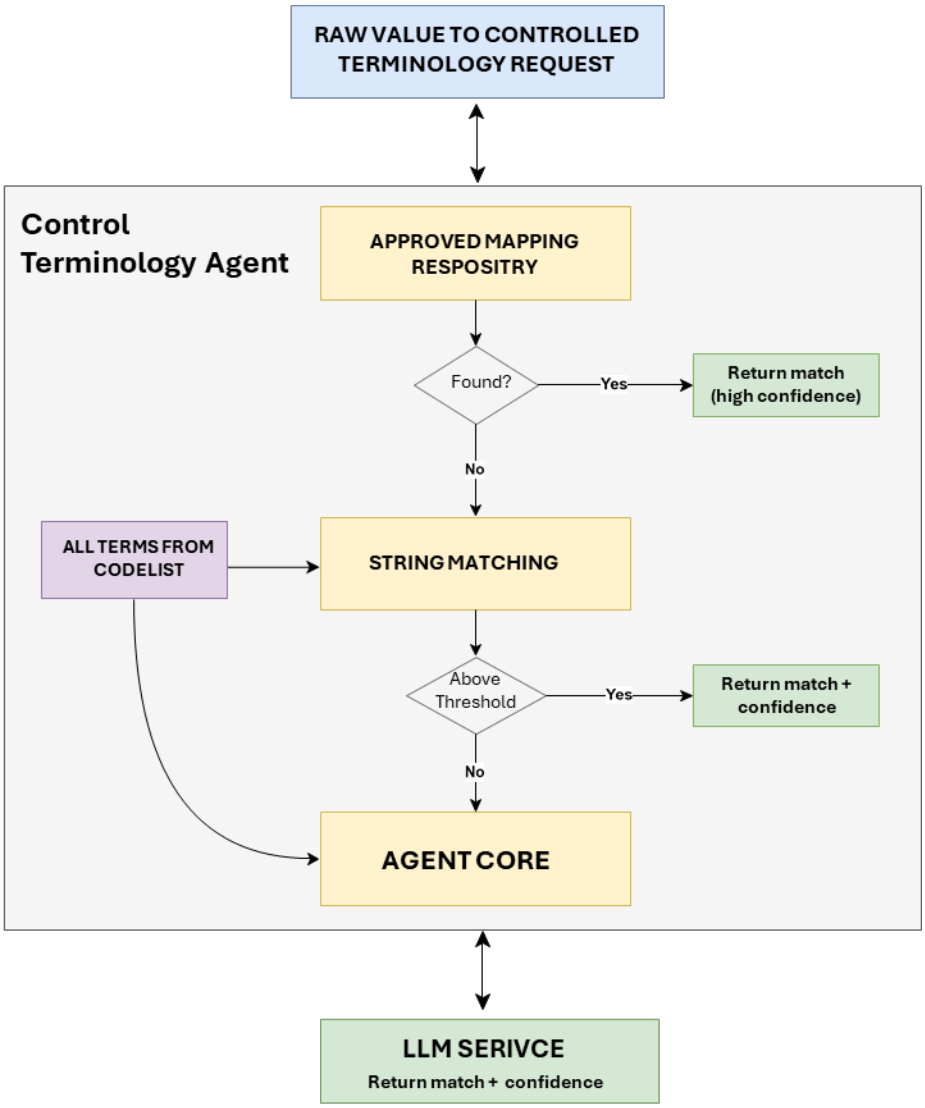


Image of controlled terminology recommendations in JetConvert™

On the left you can see the raw values received for laboratory test names (LBTEST) and on the right the controlled terminology recommendations with a confidence ranking.

With this approach, as the LLM is seamlessly integrated into JetConvert™. The users are not interacting with the LLM directly and nothing changes in the way they interact with the system compared the pre-LLM version of the system. The only noticeable difference to the users is the additional recommendations provided for their consideration.
Below is an illustration of the Controlled Terminology Agent schema –



CONTINUOUS LEARNING

When designing a system, the "do it only once" concept is one of the biggest contributors to efficiency and quality gains. This methodology is very applicable to statistical programming, where we tend to do very similar tasks repetitively across multiple studies.

This is where the Approved Mapping Repository adds significant value, for this use case we integrated a repository that stores previous mappings, each consisting of the raw source value, the SDTM target (variable and code list), and the converted SDTM value, along with their approval flag. Before sending any term to the LLM, the system checks the repository. If a mapping has been previously reviewed and approved, it is reused automatically, allowing both the Mapper and the Reviewer to skip that term entirely. Additionally, exact matches to the code list are auto-mapped and marked as reviewed. This ensures that controlled terminology mapping and review is truly done only once.

This approach addresses a key constraint of using LLMs in regulated environments, each mapping would still require verification. The general purpose LLM has no context about which mappings have previously been reviewed and approved, resulting in redundant work for reviewers. By reusing approved mappings, verification happens only once, and those mappings can be applied consistently across studies without repeated review.

CONFIGURATION OVERHEAD

Focusing on Controlled Terminology mapping, which is a small part of the complex SDTM conversion process, ensured that the required context and configuration remain minimal. The agent and LLM only need to receive the following information:

- Source value – Retrieved from source data
- Controlled terminology version – Configured at study creation

- Code list identifier – Obtained from SDTM standards already available in the system

Due to the seamless integration between the Controlled Terminology Agent and JetConvert™, all needed context is already available in the system, hence no additional configuration is needed. Configuration in JetConvertTM is done only once on a study leve. For example, the version of SDTMIG that should be used is specified only once per study and all integrations and modules of the system re-use this information as needed.

MEASURABLE VALUE

The best way to show visible value is to have metrics that can inform decisions. For this use-case we compared the Controlled Terminology Agent results with previously-mapped studies where terms were manually mapped by an SDTM expert and then independently validated. The Agent was able to correctly map over 91% of the terms. Of the remaining items, 6% were returned as unknown, and 2% were returned with low confidence or ambivalent results. Here is how this translates to actual time saving –

Task	Performance	Time Saving	Total per 1,000 terms
Auto mapped	40%	2 minutes	13.3 hours
Correct results	52%	1 minute	8.6 hours
No results	6%	-	-
Incorrect results	2%	-	-
Total	22 hours / 33.3 hours = 66%		

By ‘Auto mapped’ we include both previously-mapped terms as well as one-to-one match with Controlled Terminology. The ‘Correct results’ task then include both string matching and LLM.

These are only initial results and we expect the CT Agent to reach significantly better results within merely few months as it handles more studies.

USE CASE 2: VARIABLE-LEVEL MAPPING RECOMMENDATIONS

BACKGROUND

Variable-level mapping aligns source variables with SDTM targets based on source data characteristics, CRF annotations, SDTM standards, and sponsor-specific conventions. This task requires significant expertise and manual effort.

WORKFLOW INTEGRATION

The Variable Mapping Agent embedded in JetConvert™ automatically assembles the relevant context and presents variable-level mapping recommendations directly within the workflow. Source data are uploaded into the system, after which users map datasets and variables to SDTM with the agent providing recommendations for consideration. Users can either accept these recommendations or override them and perform the mapping manually, without leaving the system –

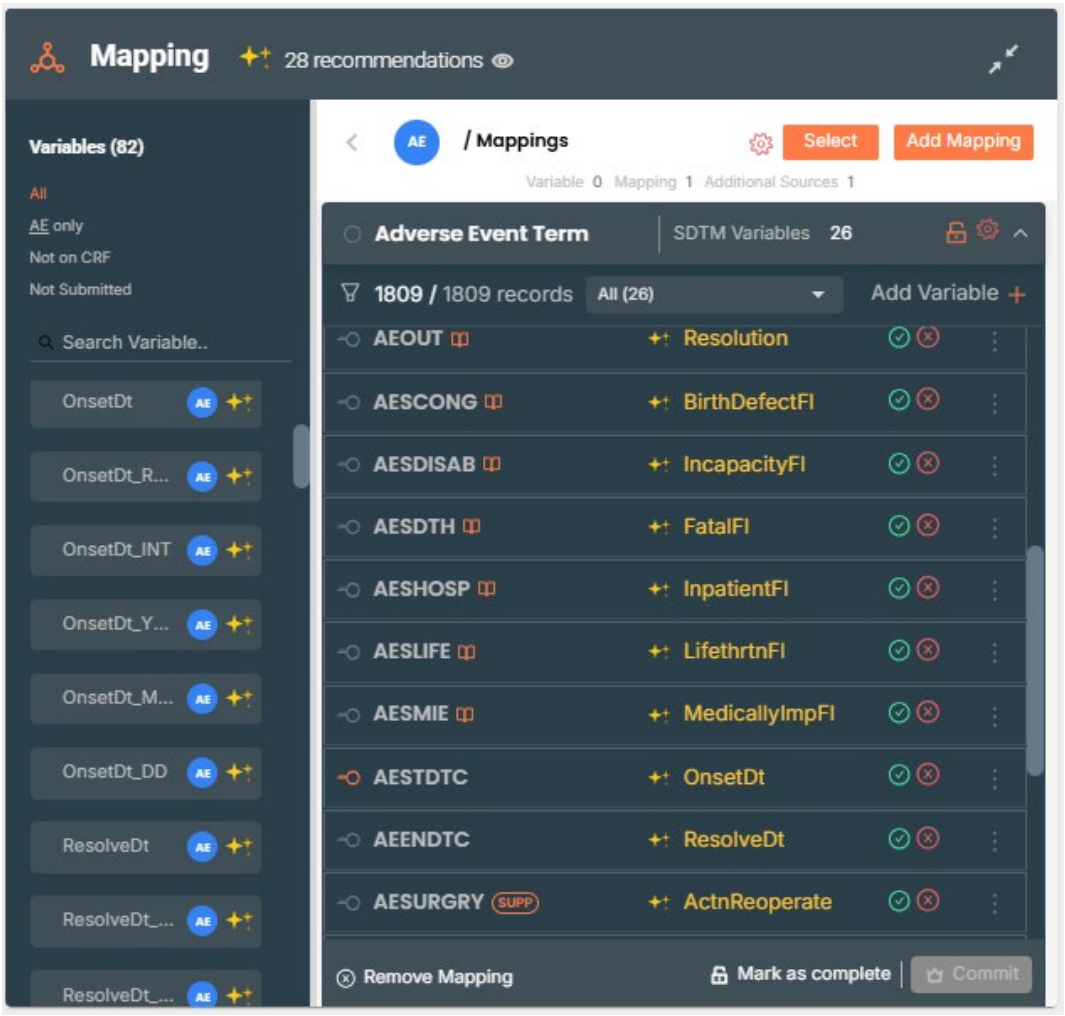
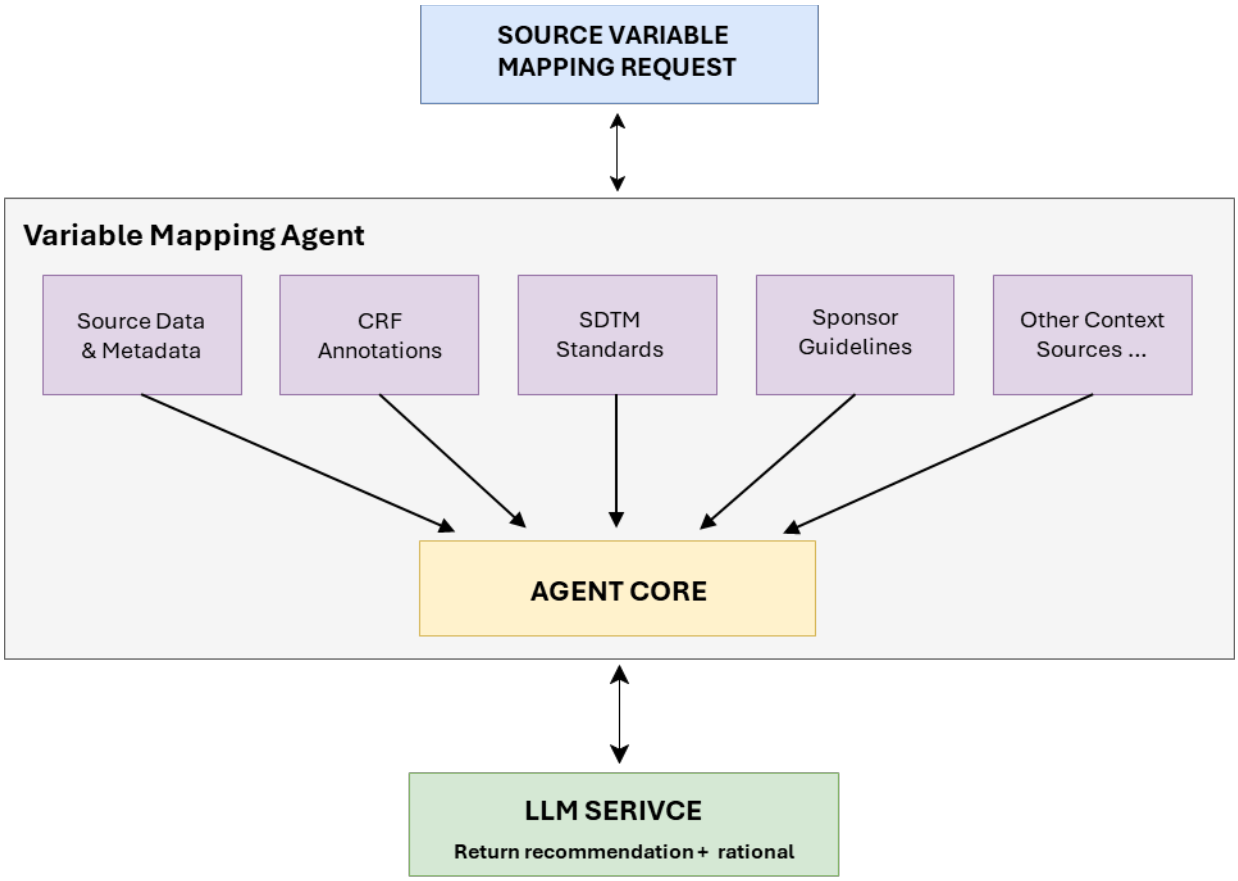


Image of variable level mapping recommendations in JetConvert™

On the left side, you will see a list of source (raw) variables available for mapping. On the right is the mapping panel where each row represents an SDTM variable (target). In the middle of each row, the system displays the suggested source variable(s) to map. To the right of each row are controls to approve or dismiss the recommendation.



Similarly to the previous use case, the users do not interact with the LLM directly, they only interact with the recommendations provided by the Variable Mapping Agent.

CONTINUOUS LEARNING

Variable level mapping is a complex process, having many factors potentially affecting where each source variable should be mapped to and if further processing, like splitting, concatenation, or merging of data, is needed.

For this use case, continuous learning is achieved through evolving guidelines rather than historical lookups. The Variable Mapping Agent supplements the base LLM query with guidelines drawn from sources such as the SDTM Implementation Guide, FDA Technical Conformance Guide, PHUSE papers, industry best practices, and sponsor-specific instructions for handling ambivalent cases. These guidelines are updated as standards, regulations, and best practices evolve. Furthermore, as new edge cases are encountered, learnings are encoded into the guidelines, ensuring the system benefits from accumulated domain expertise without requiring exact historical matches.

There are often similarities between source data and where they need to be mapped in SDTM. A great example is Sponsors using standard CRFs or Data Transfer Agreements with data vendors which significantly increase consistency across studies and re-usability of mapping logic. So, once again we must implement the “do it only once” rule, allowing the system to re-use mapping logic.

CONFIGURATION OVERHEAD

As with the previous use case, variable level mapping requires diverse context to determine the correct target variable. An SDTM expert needs to consider source data characteristics, CRF annotations, SDTM standards, industry best practices and sponsor specific guidelines. Providing this context to an LLM for each mapping would be impractical.
Due to the complete integration into JetConvert™, all needed context is already available in the system. Source data and metadata are uploaded, CRFs are linked, SDTM standards are built in, and sponsor guidelines are configured at study setup. No additional configuration is needed per mapping, the Variable Mapping Agent assembles the relevant context automatically.

MEASURABLE VALUE

We compared the Variables Mapping Agent results to previously-mapped studies where source variables were manually mapped by an SDTM expert and then were independently validated. The Agent was able to correctly auto-map more than 80% of the raw variables to SDTM variables. As with the previous use case, we expect significant improvement in the upcoming months as guidelines are refined and edge cases encountered across more studies are incorporated into the Variable Mapping Agent.

CONCLUSION

Moving from experimentation to production with task-specific GenAI systems is challenging, whether you are building these systems yourself or buying from a vendor. Following the five principles, focusing on narrow high-value use cases, seamless integration with existing workflows, scaling through continuous learning, keeping low configuration burden and showing fast and measurable value, are essential for a successful production implementation. In this paper, we provided practical examples of how we interpret and implement these principles in our system.

With the improvement in technology and LLM models, there are many potential use cases to increase efficiency and quality in the statistical programming arena. Here are a few –

- Sharing with the SDTM SME data insights that are essential for mapping considerations in order to minimize the time required to research the data manually. These will include information on the data structure collected, values, discrepancies, multiple sources handling and more.
- "Talk to your data" analytics – perform ad-hoc statistical exploration using natural-language queries to slice and dice data, cross-reference domains, and quickly surface outliers or hidden patterns.
- Generating polished drafts of Trial Design domains ready for human inspection and finalization.
- Checking data against CDISC standards, FDA requirements, and sponsor guidelines using everyday language.
- Generating or enhancing Define.xml, reviewers guide and other similar documentation, including variable derivations, data issues explanations and similar content in plain language.
- Conversion copilot - answering CDISC standards related questions in context (acting as an on-demand SME).
- TLFs generation

By sharing our emerging insights and implementation strategies, we hope that you have gained knowledge to equip you to effectively implement your next LLM adoption and balance innovation with operational rigor.

REFERENCES

[1] MIT NANDA: Aditya Challapally, Chris Pease, Ramesh Raskar, Pradyumna Chari, "The GenAI Divide: State of AI in Business 2025", July 2025

CONTACT INFORMATION

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

Author Name: Eyal Wultz
Company: Bioforum
Address: 5 Ilan Ramon, Ness Zionna, Israel
Work Phone: +97289313070
Email: eyal.wultz@bioforumgroup.com
Website: www.bioforumgroup.com