Using Large Language Models to Streamline and Enhance R&D Surveys

Thorsteinn Siglaugsson

Statistics Iceland

thorsteinn.siglaugsson@hagstofa.is

Abstract.

Large language models (LLMs) can support official statistics most effectively when they are embedded in concrete production tasks rather than treated as autonomous systems. This paper discusses three practical use cases from the production of R&D statistics at Statistics Iceland. First, ChatGPT and DeepSeek were used to help design a Microsoft SQL Server pipeline that transforms raw R&D survey outputs into a full SDMX template for Eurostat. The resulting architecture used unpivoted survey data, mapping tables, dimension and fact tables, and Cartesian joins implemented through stored procedures. Second, the paper outlines a planned approach for using LLMs to help translate Eurostat's often unstructured validation requirements into a structured rule framework suitable for implementation in SQL. Third, it proposes a framework for classifying R&D projects into Frascati Manual categories and related research areas in order to support richer analysis and more targeted R&D policy. The three cases illustrate that the main value of LLMs lies in formalising logic, accelerating design work and improving documentation, while leaving methodological judgement and final responsibility with subject-matter experts.

Keywords: Large language models; R&D statistics; SDMX; data validation; Frascati Manual; statistical production

1. Introduction

National statistical offices are under pressure to modernise production systems while preserving the core values of official statistics: quality, comparability, confidentiality and reproducibility. Large language models (LLMs) are increasingly relevant in this context because they have shown useful capabilities across software-engineering tasks such as code generation, code understanding and debugging, although their outputs still require careful human verification (Hou et al., 2024). At the same time, they are particularly well suited to tasks that involve translating between natural language, semi-structured logic and executable code.

This paper discusses three such use cases in the production and dissemination of R&D statistics at Statistics Iceland. The first concerns the design of a Microsoft SQL Server pipeline for transforming raw survey outputs into an SDMX structure for Eurostat transmission. The second concerns the challenge of translating Eurostat's largely narrative validation guidance into a structured and reusable rule framework. The third concerns the possible use of LLMs to enrich R&D statistics by classifying underlying projects into Frascati Manual categories and related research areas.

The paper is deliberately practice-oriented. In all three cases, the model is treated as an assistant for structuring logic and accelerating iteration rather than as a substitute for statistical expertise.

2. LLM-assisted design of an SDMX transformation pipeline

The first use case arose from a specific production bottleneck: how to convert raw outputs from the national R&D survey into a complete SDMX template suitable for Eurostat transmission. Earlier workflows involved substantial manual handling, which made production slower and increased the risk of format inconsistencies and plausibility problems.

ChatGPT and DeepSeek were used during the design phase to explore alternative ways of structuring the SQL solution. A particularly useful suggestion was to first unpivot the survey outputs from a wide survey format into a longer structure in which survey fields could be handled more systematically. This made it easier to separate the raw reporting structure from the reporting concepts required in SDMX.

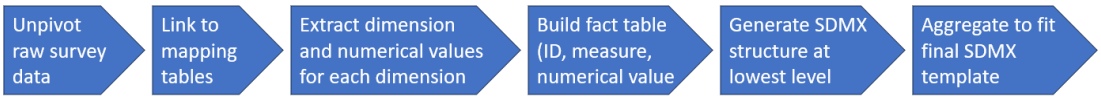

Figure 1: High-level overview of automated data-flow.

A second important design idea was to use mapping tables to connect survey fields to SDMX variables and attributes. This provided a transparent layer between source data and reporting output and made it easier to maintain the system when variables or classifications changed. A third idea was to construct the reporting structure through dimension and fact tables and then use Cartesian joins to generate the full SDMX template, including combinations that later required zero, missing or observed values.

The final implementation was built in Microsoft SQL Server entirely through a series of 33 stored procedures developed in close collaboration with the LLMs. This was important because it created a repeatable execution sequence, supported documentation and version control, and reduced dependence on ad hoc manual processing. The result was a more consistent production pipeline and faster preparation of Eurostat outputs. The main contribution of the LLMs was therefore both architectural and operational: they helped identify and build a workable transformation logic that could then be refined and implemented by staff within the statistical production environment.

3. From unstructured Eurostat requirements to structured validation rules

The second use case addresses a different issue. Eurostat validation requirements are often expressed in semi-structured or narrative form. They may combine domain restrictions, population filters, accounting identities, exceptions and reporting conventions in ways that are understandable to experts but difficult to translate directly into an operational validation system.

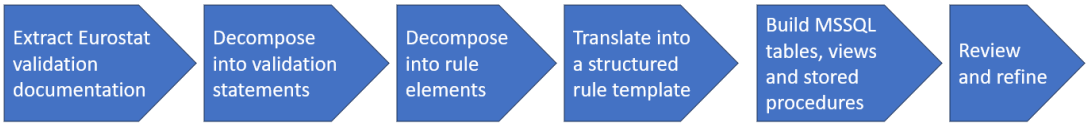

Figure 2: LLM assisted structuring of validation template.

This is a promising area for LLM support because the task is essentially one of structuring meaning. The challenge is to decompose a textual requirement into rule elements such as the relevant variables, the logical condition, the applicable population, the expected message and the severity level. Recent work on structured entity extraction shows that language models can help extract structured information from unstructured text, which makes them relevant to this problem even though statistical validation rules remain a specialised domain requiring expert review (Wu et al., 2024; Dagdelen et al., 2024).

The process of structuring unstructured validation information requires co-operation between the expert and the LLM as Eurostat documentation often contains ambiguities that require statistical judgement. For that reason, the intended use is iterative: experts provide the source guidance, the model proposes a structured interpretation, and subject-matter experts then review, revise and approve the rule before implementation. The value lies in accelerating the transition from documentation to executable logic and in improving the documentation of the validation system itself. Once the validation information has been structured in this manner, it can be applied directly to the populated SDMX template.

4. Enriching R&D statistics through project classification

The third use case is more exploratory but potentially more transformative. Traditional R&D surveys measure expenditure, personnel and institutional patterns well, but they often say relatively little about the substantive content of the projects behind the totals. If project descriptions could be classified more systematically, R&D statistics could become more analytically useful.

An LLM-assisted framework could support the extraction and classification of project information into Frascati Manual and subject-matter categories, including type of R&D activity and field of research. This would make project-level material more analytically useful while remaining grounded in the internationally recognised classification framework for R&D statistics (OECD/Eurostat, 2015).

The policy value of such a framework could be considerable. Better classification of projects into research and subject-matter areas can help policymakers identify where public support is concentrated, where gaps remain and whether different types of support are aligned with national priorities. If project classifications can later be linked to outcomes such as firm growth, commercialisation or wider spillovers, they may also strengthen the evidence base for more targeted R&D support. This matters because a broad literature finds positive private and social returns to R&D investment, albeit with substantial heterogeneity across contexts and methods (Ugur et al., 2016), while recent evidence also points to potentially large social returns to public R&D (Fieldhouse and Mertens, 2025).

This use case is also the most methodologically sensitive. Free-text descriptions can be incomplete or strategic, model outputs can reflect inconsistent interpretation, and project-level enrichment raises confidentiality concerns. Any implementation would therefore require strong safeguards, including controlled vocabularies, expert review, transparent coding rules and careful protection of identifiable information. Even so, the framework points toward a broader role for

LLMs in helping statistical offices move from transactional survey processing toward richer analytical use of survey material.

5. Discussion and conclusion

Across the three use cases, a common pattern emerges. LLMs are most useful when they help translate between forms of knowledge that are otherwise difficult to connect quickly: survey structures and SDMX concepts, narrative requirements and executable rules, or project descriptions and analytical categories. Their main contribution is therefore not autonomous production, but assistance in formalising logic and accelerating design work.

At the same time, each case also underlines the need for human control. Statistical production requires methodological judgement, traceability and accountability. LLMs can suggest plausible structures, but they do not remove the need for expert review. Their effective role in official statistics is therefore as a practical design assistant within governed workflows.

The lessons from these R&D examples are likely to be transferable to other surveys and other national statistical offices. Many organisations face similar challenges in automating SDMX outputs, operationalising validation rules and extracting more value from textual survey material. Used carefully, LLMs can help address each of these tasks while leaving responsibility for official outputs firmly with statistical institutions.

6. Appendix: List of references

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J. and Wang, H. (2024). Large Language Models for Software Engineering: A Systematic Literature Review. ACM Transactions on Software Engineering and Methodology, 33(8).

Wu, H., Yuan, Y., Mikaelyan, L., Meulemans, A., Liu, X., Hensman, J. and Mitra, B. (2024). Learning to Extract Structured Entities Using Language Models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A. and Jain, A. (2024). Structured information extraction from scientific text with large language models. Nature Communications, 15, 1414.

OECD/Eurostat (2015). Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development. Paris: OECD Publishing.

Ugur, M., Trushin, E., Solomon, E. and Guidi, F. (2016). R&D and productivity in OECD firms and industries: A hierarchical meta-regression analysis. Research Policy, 45(10), 2069-2086.

Fieldhouse, A. J. and Mertens, K. (2025). The Social Returns to Public R&D. NBER Working Paper 33780.