Ciclo di ricerca statistica 2024 - 2026

Scheda di progetto

* **Title: “***Reading Between the Lines*. An LLM application to generate quantitative checks from the reporting rules”
* **Authors***: [M. Orlandi e S. Giacinti? (DAES); M. de Leonardis (SSVA); Diana Breascu, Simone del Puppo (DIA); Umberto Proia e Andrea Magazzeni (DCR- Anacredit)]*
* **Deadline:** 06/2026; an interim document will be produced in the second half of 2025 Brief description of the activity:

On a regular basis the Bank of Italy collects a wide range of statistical and supervisory data from banks and financial institutions according to a set of reporting rules that are usually composed of a legal text (i.e. the regulation) and complementary technical documentation, glossary and manuals. In order to be released to the users, the data must pass a series of quantitative data quality checks (mainly consistency, accuracy, completeness and plausibility checks).

The aim of the work is to leverage LLMs to extract the relevant pieces of information contained in the reporting rules so as to automatically design the appropriate data quality checks that will be applied to the incoming data-sets. This solution essentially aims at reducing the time required in the pre-emptive analysis of the reporting rules that is a necessary to define the data quality rules that will be implemented in the IT application supporting the collection phase.

We intend to apply the LLM solution to the AnaCredit Regulation (and related documentation) since it represents rather a standard of the European reporting rules.

The study proposed here is articulated into three methodological steps:

1. Selection of a data collection for which the Set of Deterministic Controls (SDC1) has already been implemented. This would act as a control list for the checks identified by the new methodology
2. Application of the LLM to the documents comprising the reporting rules in order to produce an automatic Set of Deterministic Controls (SDC2).
3. Comparison between SIC1 and SDC2.

The analysis of the differences between SIC1 and SDC2 above defined would allow to obtain indications on how to best draft the reporting rules (in terms of wording, partitioning of the content, etc.) in order to facilitate its automatic reading with the LLM methodology. Furthermore, through the application of the LLM, some automatic controls found in SDC2 that had not been identified in SDC1 and that, instead, deserved to be implemented, could emerge. Finally, the analysis of deterministic checks in SDC1, that were not detected in SDC2 could lead to the decision to recalibrate that checks.

Research Need/Purpose:

The purpose of the work is to define a methodology that, starting from a reporting regulation and related technical documentation, both mainly written in natural language, identify a set of key-variables and a set of checking algorithms that connect such variables that can be implemented in the regular IT application used to manage the data collection.

# Data and working environment

Application to the AnaCredit Regulation and related technical manual.

# Data Classification:

Since the proposed study will involve the analysis of reporting regulation and technical manuals, and considering that the deterministic control system defined on the AnaCredit dataset does not contain nominative and/or sensitive data, there would be no need to anonymization.

# Statistical methodologies that will be used:

The study will involve the application of Large Language Models and, more precisely, it will concern the automatic analysis of texts written in natural language with legal and technical jargon and the production of a system of deterministic checks (plausibility and consistency) between the key-variables identified in the analysed texts.