Enhancing metadata with generative AI

Olivier Sirello (Bank for International Settlements, Switzerland)

UNECE Expert Meeting on Statistical Editing

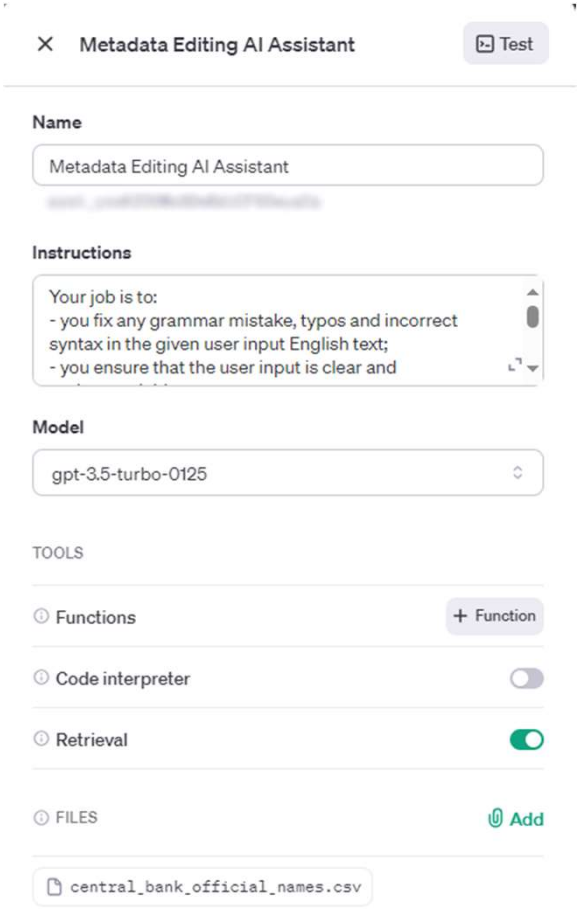
Vienna, 7 October 2024

Enhancing metadata through generative AI

- Metadata play a fundamental role in official statistics
  - Transform data into information (structural and reference metadata)
  - Enable exchange of data
  - Ultimately, secure credibility and trustworthiness
- Yet their generation and editing can be costly for compilers
  - Curation, including editing and review, is often manual
  - Typically with poor standardization, notably for reference metadata
  - Overall time-consuming and resource-intensive task
- Can generative AI help compilers to enhance metadata editing?

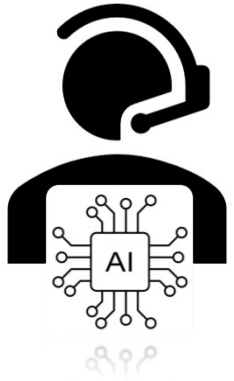
Introducing the BIS Metadata AI Editor

- A custom program for **metadata formatting** and **editing**
- Leveraging **AI-powered assistant(s)** to respond to specific sets of instructions
- An **end-to-end solution**: SDMX-compliant input and output
- **Low implementation** costs, **ease of use** for the final users



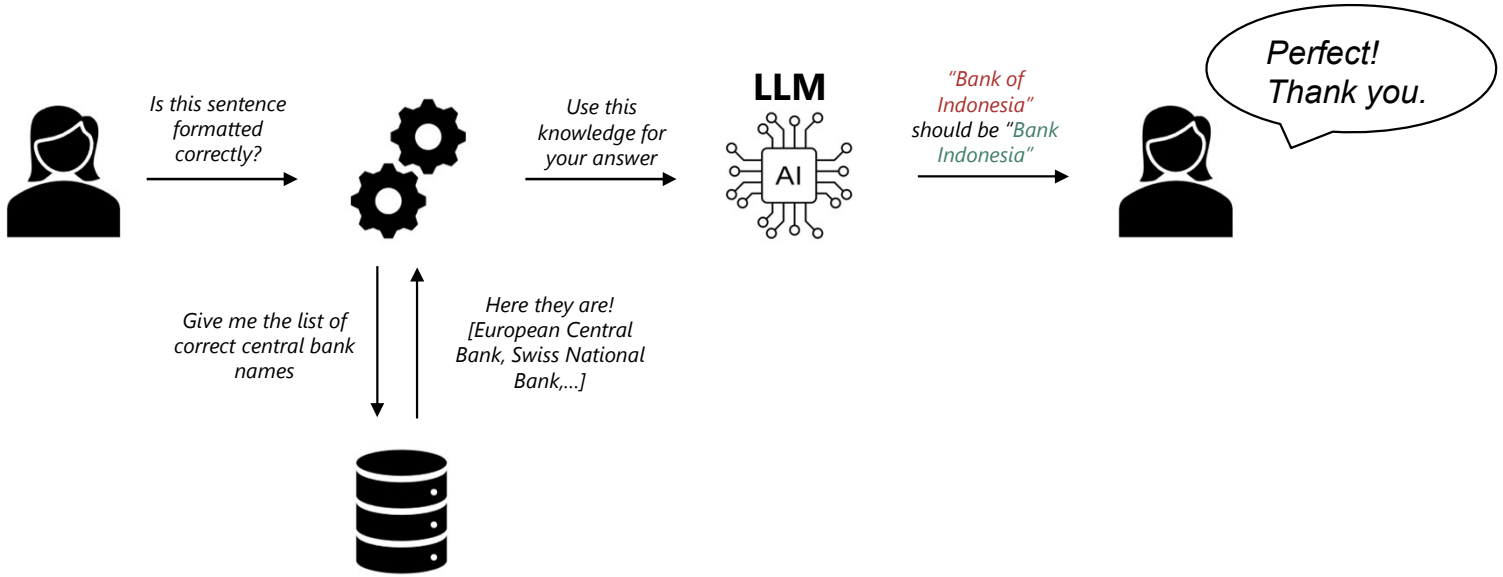
What is an assistant?

- Custom AI that uses OpenAI's models and tools
- Can call the models with specific instructions
- Can use different tools in parallel
  - Code writing – Assistant writes and runs Python code
  - Function calling – getting structured output from the model (eg JSON)
  - **Knowledge retrieval** - augments the Assistant with custom knowledge
- Can access/create files in several formats



Knowledge retrieval

- OpenAI's version of Retrieval-Augmented Generation (RAG)
  - Enables the LLM to form answers based on a custom knowledge base



  - Assistant API (version 2) available since mid-2024

Why relying on the API?

- Ability to implement an **end-to-end workflow**
- **Standardization**
  - Always the same prompt
- **Advanced analytics**: token counting, thread execution control
- **High modularity** and customization

Instructions for the Assistant

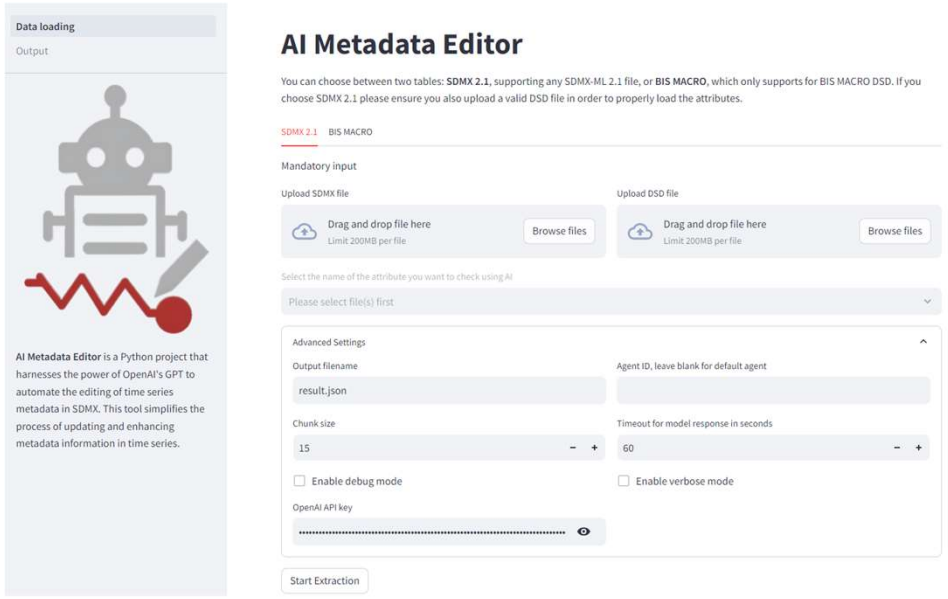
- Level of detail depends on the goal:
  - 1) Generic instructions
    - *Fix grammar mistakes, typos and incorrect syntax in the given user input*
  - 2) More "specific" requirements
    - *Abbreviate months (eg January shall be Jan) except when the month is at the end of the sentence*
  - 3) BIS specific rules
    - *Names of central banks, e.g. Magyar Nemzeti Bank and not Hungarian National Bank*
- *More ≠ better (prompt engineering)*

Tools and interface

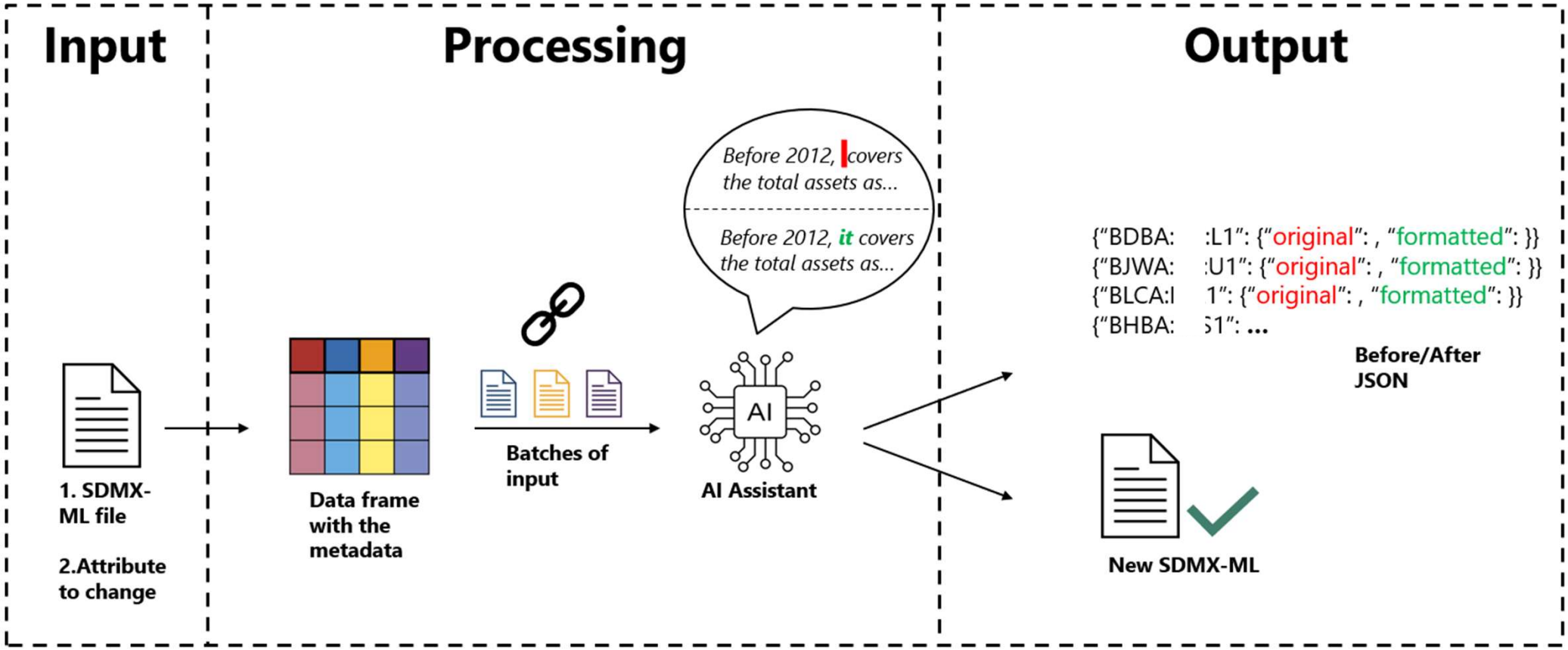
Open-source software and tools:

- Python
  - LangChain

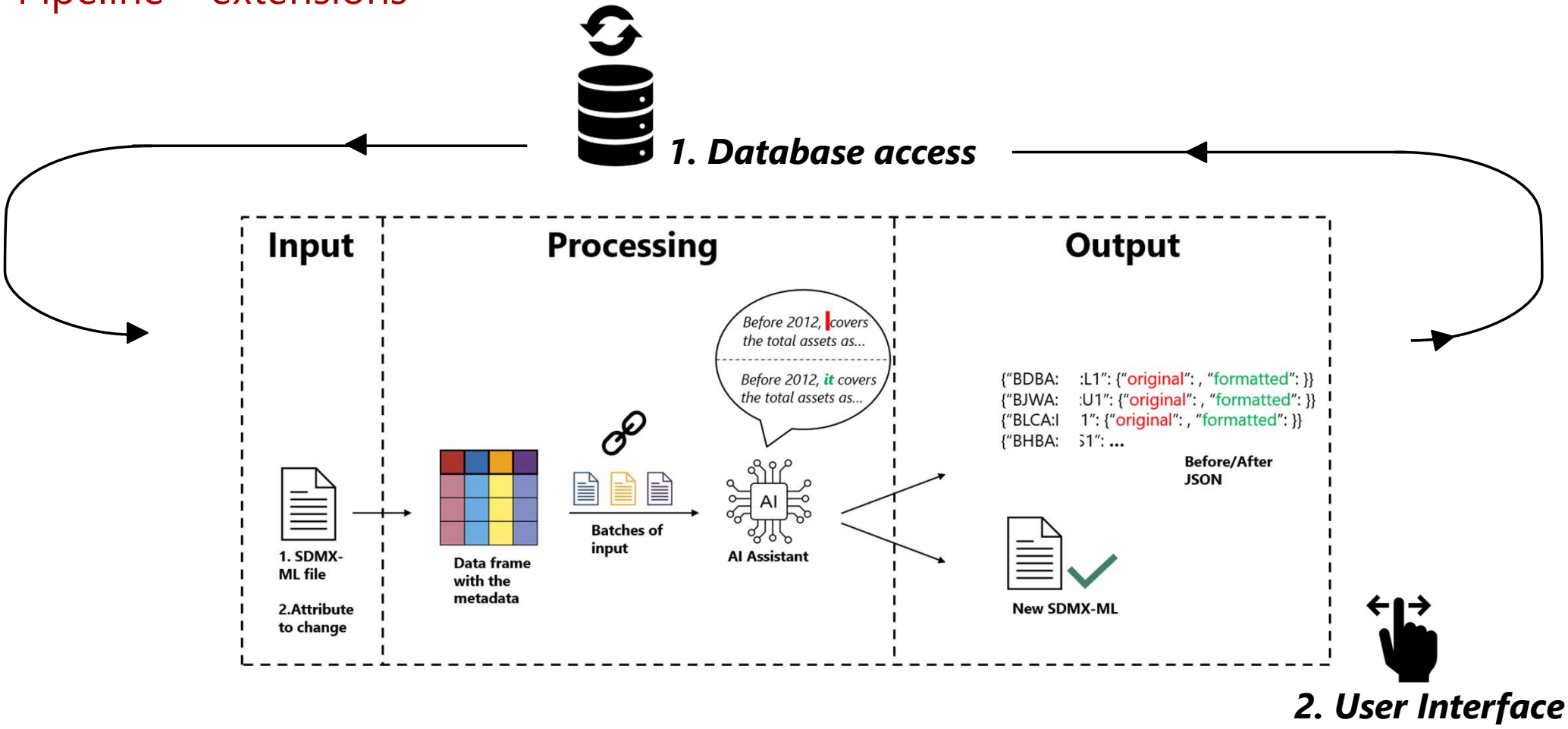
- Execution through a command-line interface
- User interface based on Streamlit



Pipeline

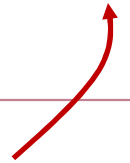

Input
Processing
Output
1. SDMX-ML file
2.Attribute to change
Data frame with the metadata
Batches of input
Before 2012, covers the total assets as...
Before 2012, it covers the total assets as...
AI
AI Assistant
{"BDBA: :L1": {"original": , "formatted": }}
{"BJWA: :U1": {"original": , "formatted": }}
{"BLCA:I 1": {"original": , "formatted": }}
{"BHBA: 51": ...
Before/After JSON
New SDMX-ML

Pipeline – extensions



Results

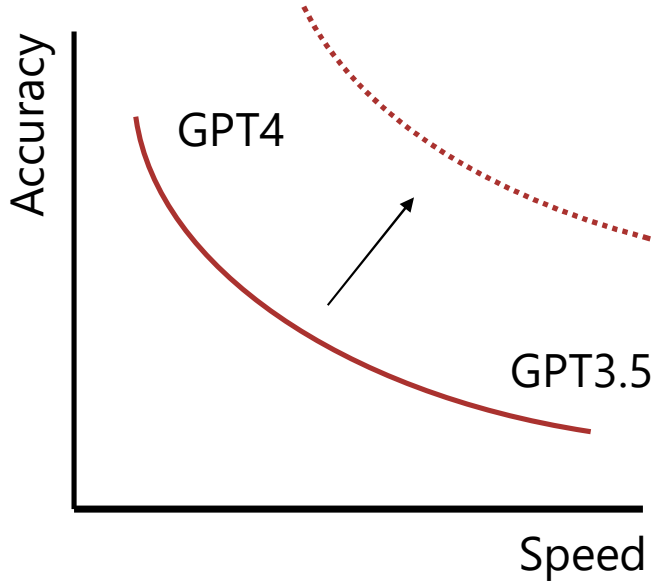
Before	After
Before 2012, ir covers the total assets…	Before 2012, it covers the total assets…
The series on commercial property prices is sourced from Central Bank of…	The series on commercial property prices is sourced from the Central Bank of…
… the source is the historical table A2 and before 1969, the table 3.6	… the source is the historical Table A1, and before 1969, Table 3.6 …
The series is sourced from the Riksbank's assets and liabilities (weekly report)	The series is sourced from the Sveriges Riksbank's assets and liabilities



As per BIS official names of member central banks

Requirements, challenges and risks

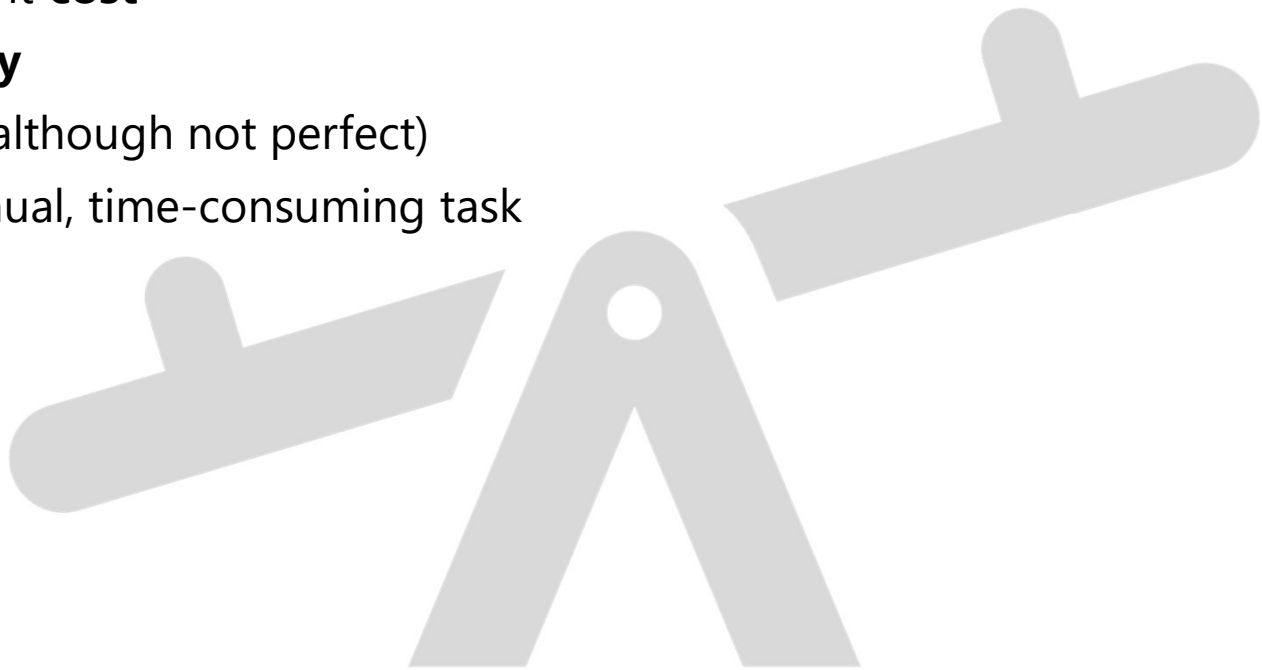
- Restricted to **public** information
- **Low reproducibility** but business case is mostly **one-off**
- Dependency on an external service
- Performance vs accuracy trade-offs
- Human-in-the-loop!
  - The only safe way of onboarding LLMs in their current form
  - Version control is key



Summary: advantages and disadvantages

+ **Low** development **cost**
+ **High modularity**
+ High **accuracy** (although not perfect)
+ **Automates** manual, time-consuming task

- Not fully **reproducible**
- **IT infrastructure** dependent
- (Requires **human supervision**)



BIS

Thank you

olivier.sirello@bis.org

BIS Data Portal

Visit BIS statistics at data.bis.org