Leveraging AI-assistants for better metadata and communication

Bilyana Bogdanova, Olivier Sirello, Krzysztof Zdanowicz - Bank for International Settlements

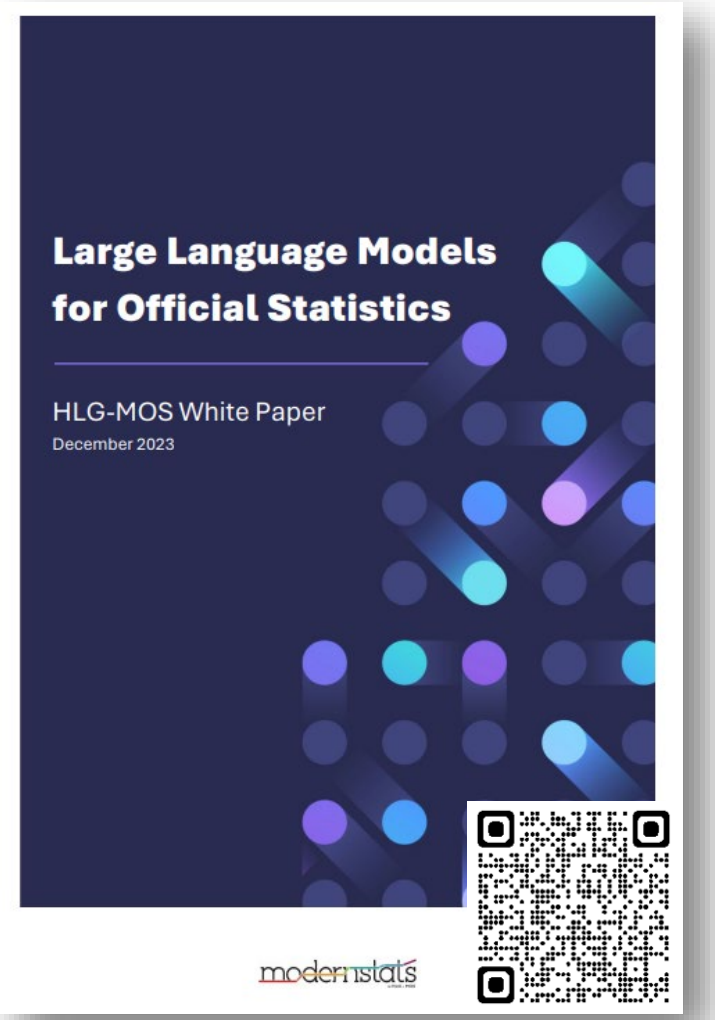
Generative AI project - High-Level Group for the Modernisation of Official Statistics

Virtual, 29th April 2024

The views expressed are those of the authors and do not necessarily reflect those of the Bank for International Settlements. All errors are our own.

Outline

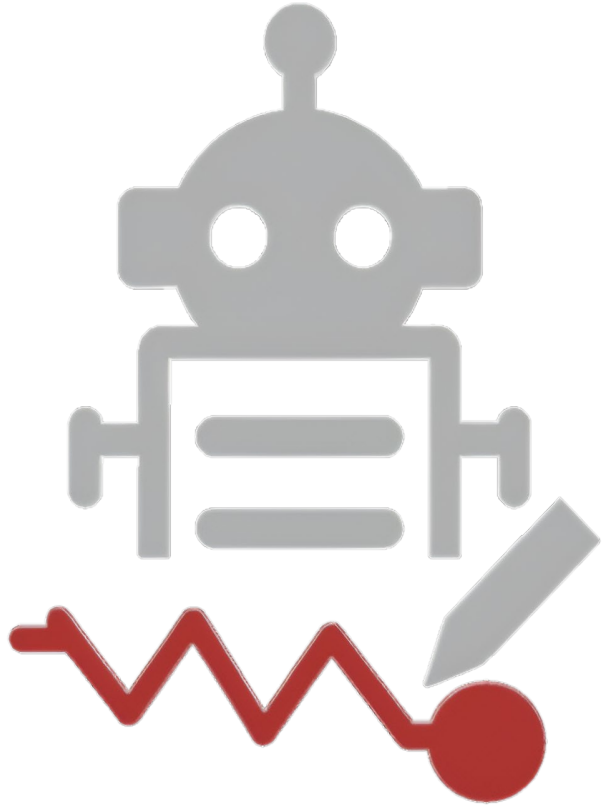
- Introduction
- Solution
- Results
- Requirements, challenges and risks



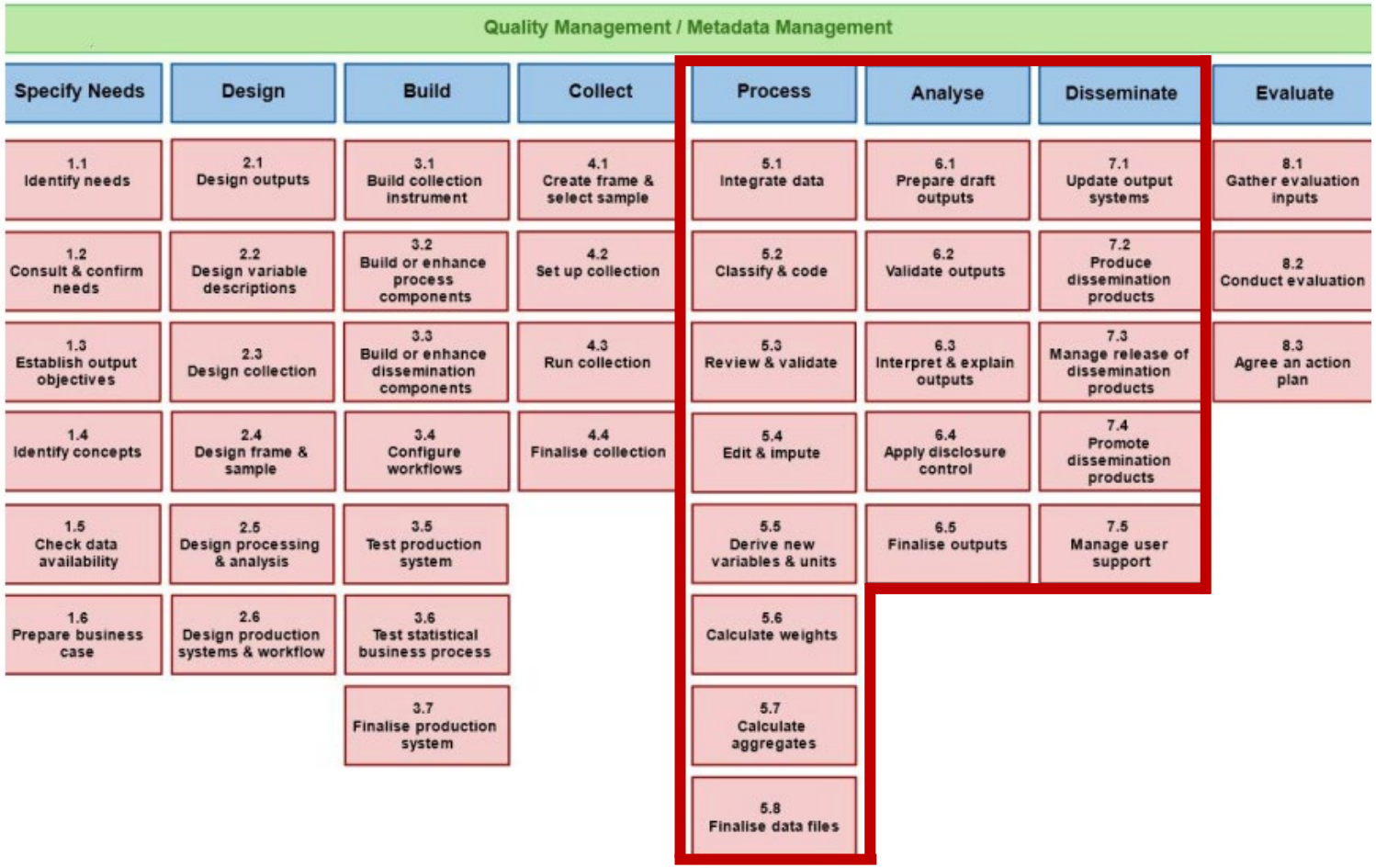
This presentation is based on UNECE (2023), Large Language Models for Official Statistics, High-Level Group for the Modernisation of Official Statistics.

At one glance

- **What**
  - Editing and checking time series metadata
- **Why**
  - Time- and resource- intensive
  - Use of knowledge retrieval is required
- **How**
  - OpenAI Assistants
  - End-to-end workflow (SDMX to SDMX)

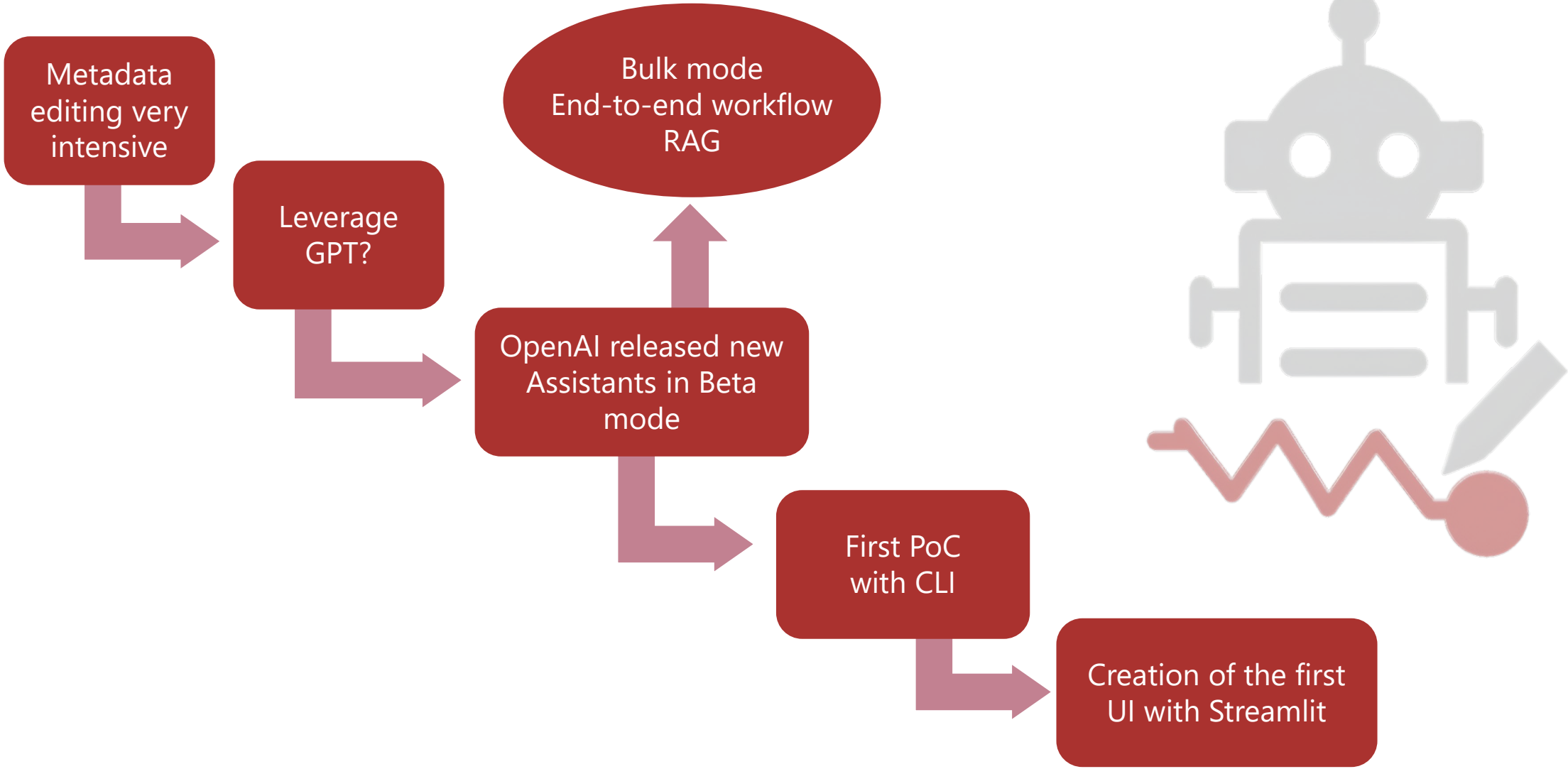


Generic Statistical Business Process Model



- Process
- Analyze
- Disseminate

The development journey: a bottom-up approach



Key principles

- **Agility**, **modularity** and **outsourcing**
  - Avoid “reinventing the wheel”
  - Leverage existing **AI-powered assistant**
  - **Generalize use cases and customize**

- **Low costs, high customization**
  - **Low implementation** and deployment costs
  - **Ease of use** for the final users
  - High **accessibility** and shallow learning curve
    - Easily accessible by anyone (eg business user) with an OpenAI account within the organisation

- An **end-to-end solution leveraging SDMX**
  - SDMX 2.1 ML file as input and output
  - Reading and parsing done through sdmxthon

The BIS Data Portal

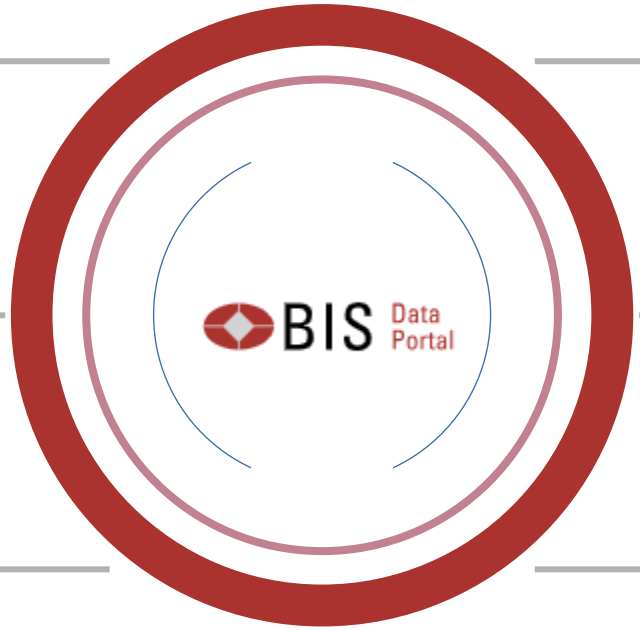
BIS statistics: why a new dissemination tool?

Consolidate heterogenous IT landscape

Improve the **user experience**

Offer a **metadata-driven** navigation

Metadata are **everywhere**



Leverage the **SDMX model**

Allow the **statistician** to manage the content

Adopt **re-usable open source** tools

The BIS Data Portal and metadata

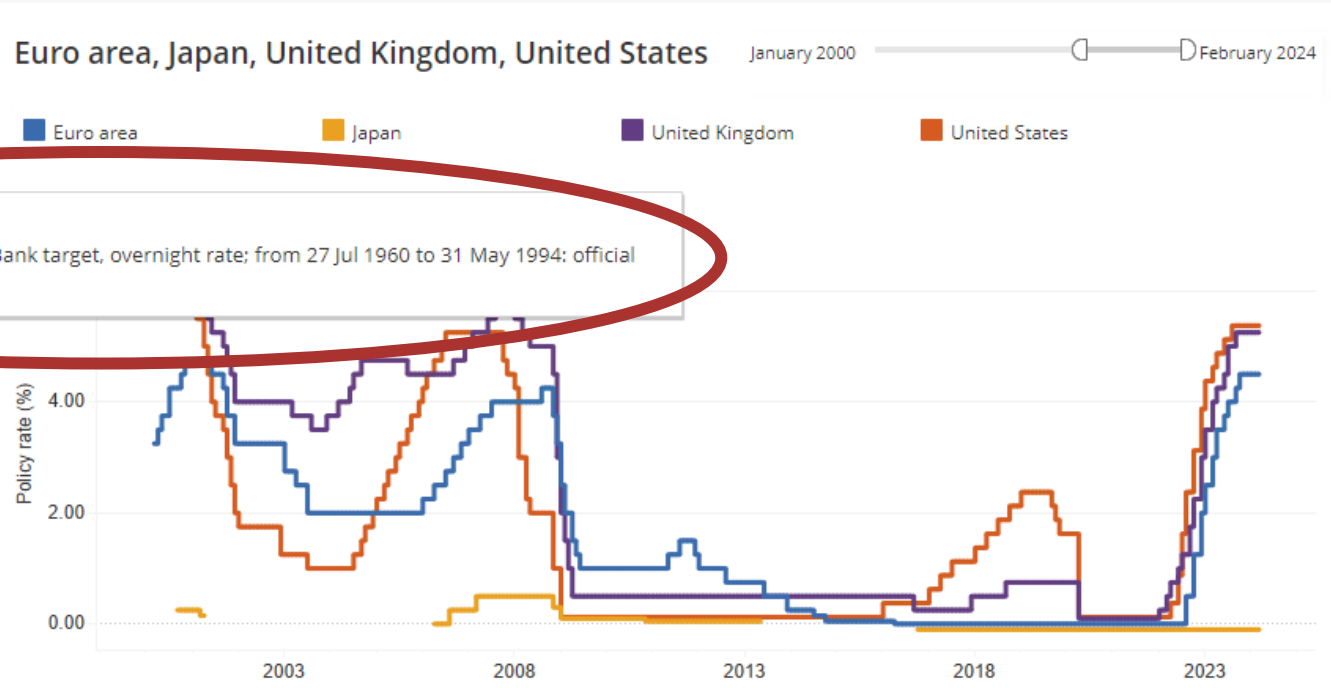
Dashboards

BIS Data Portal **Central bank policy rates**

Reference area: (Multiple values)

	December 2023	January 2024	February 2024
Argentina	100.0	100.0	
Australia	4.4	4.4	
Brazil	11.8	11.8	11.3
Canada	5.0	5.0	5.0
Chile	8.3		
China			
Colombia	3.0		
Czechia	6.6		
Denmark	3.6	3.6	3.6
Euro area	4.5	4.5	4.5
Hong Kong SAR	5.8	5.8	5.8
Hungary	10.8	10.0	
Iceland	9.3	9.3	9.3
India	6.5		
Indonesia	6.0	6.0	6.0
Israel	4.8	4.5	4.5
Japan	-0.1	-0.1	-0.1
Korea	3.5	3.5	
Malaysia	3.0	3.0	3.0
Mexico	11.3	11.3	11.3

Reference area: **Canada**
Compilation: From 1 Jun 1994 onwards: Central Bank target, overnight rate; from 27 Jul 1960 to 31 May 1994: official bank rate.



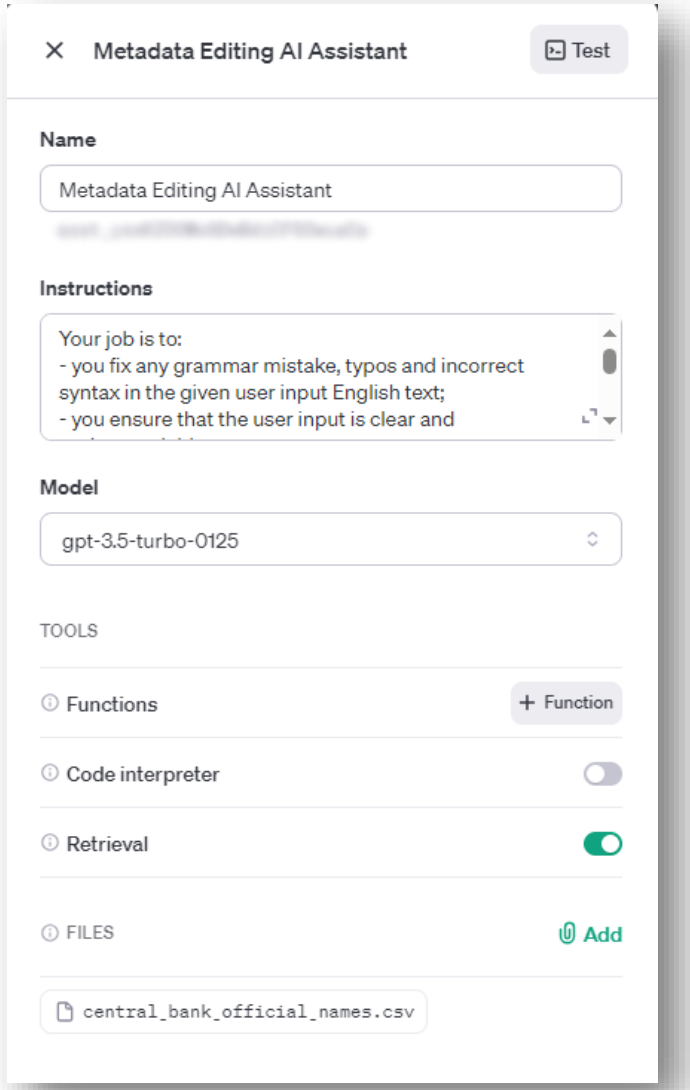
Solution

AI Metadata Editor – OpenAI Assistants

- A custom program for **metadata formatting** and **editing**

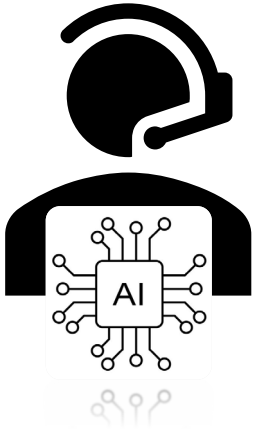
- Users may quickly:

  - Create their own assistant (eg ID)
  - Define the set of instructions
  - Select the appropriate model
  - Add other useful functions



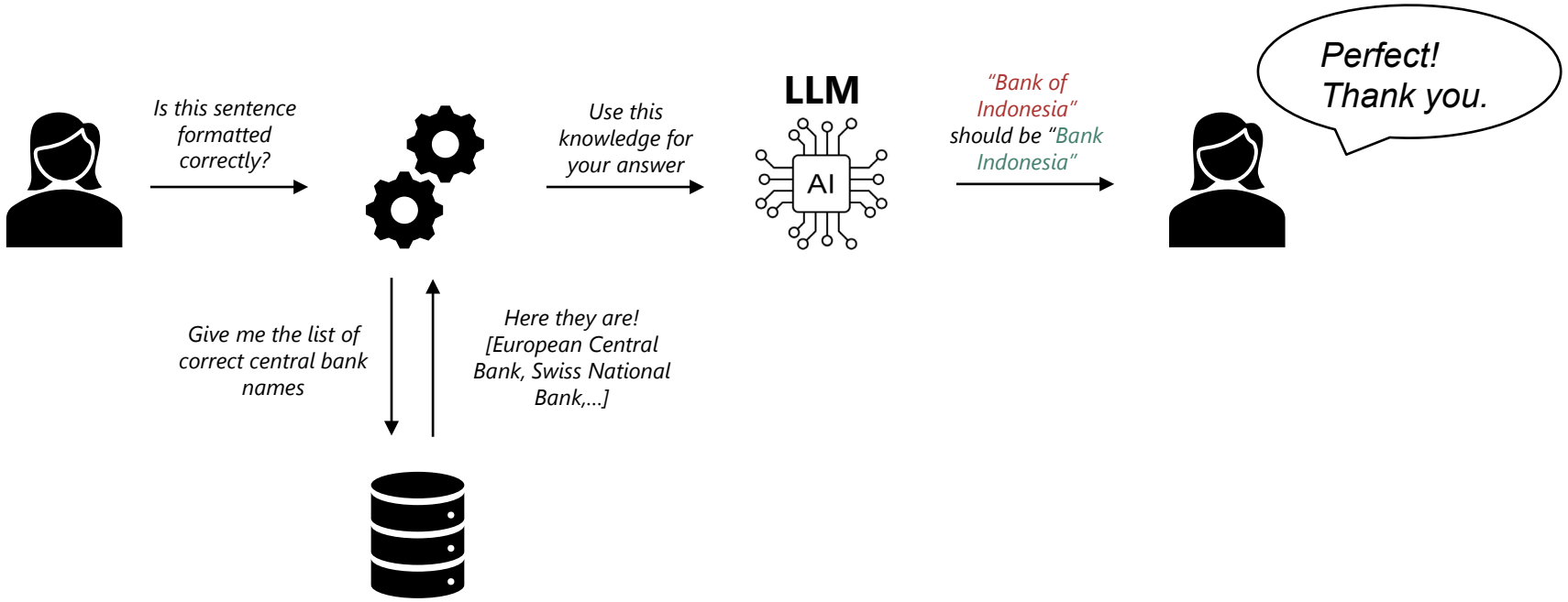
What is an assistant?

- 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

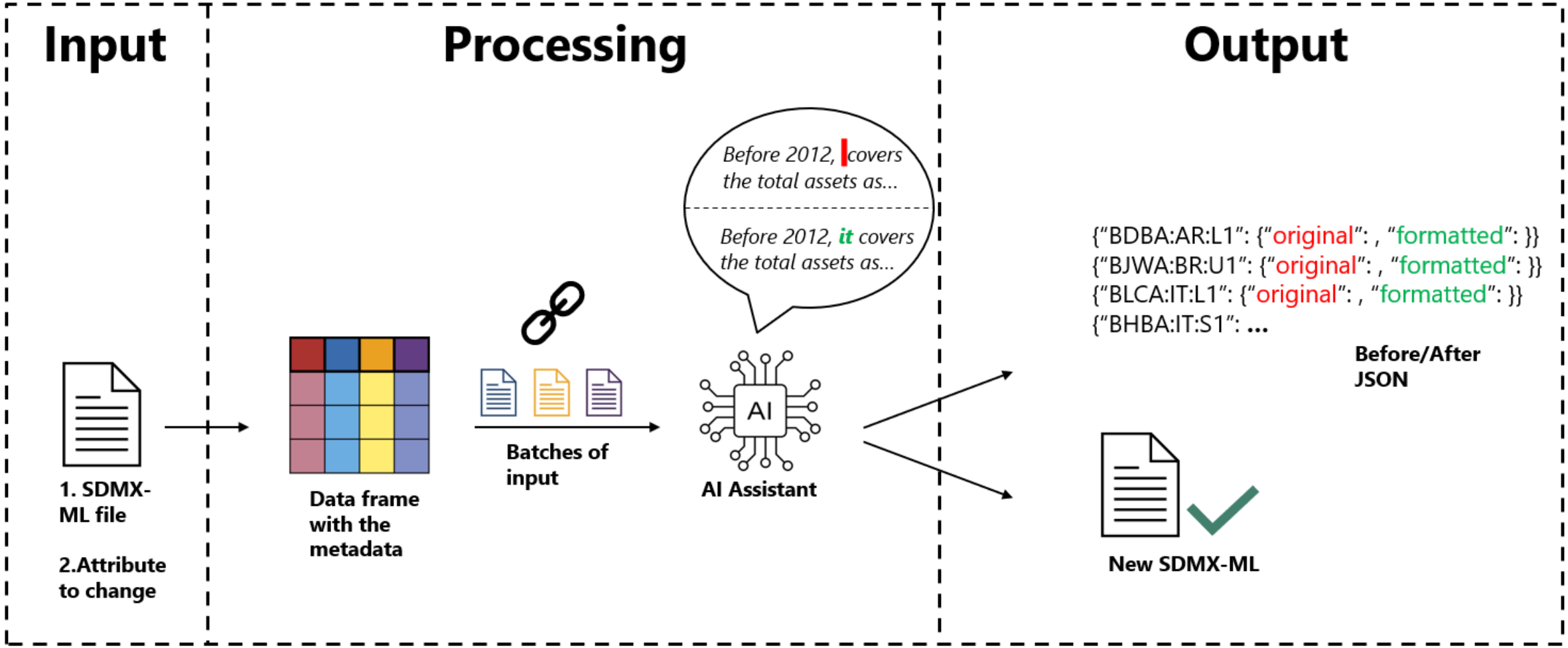


  - Only available since Nov 2023 – still beta version

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)*

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:AR:L1": {"original": , "formatted": }}
{"BJWA:BR:U1": {"original": , "formatted": }}
{"BLCA:IT:L1": {"original": , "formatted": }}
{"BHBA:IT:S1": ...
Before/After JSON
New SDMX-ML

Results

Results

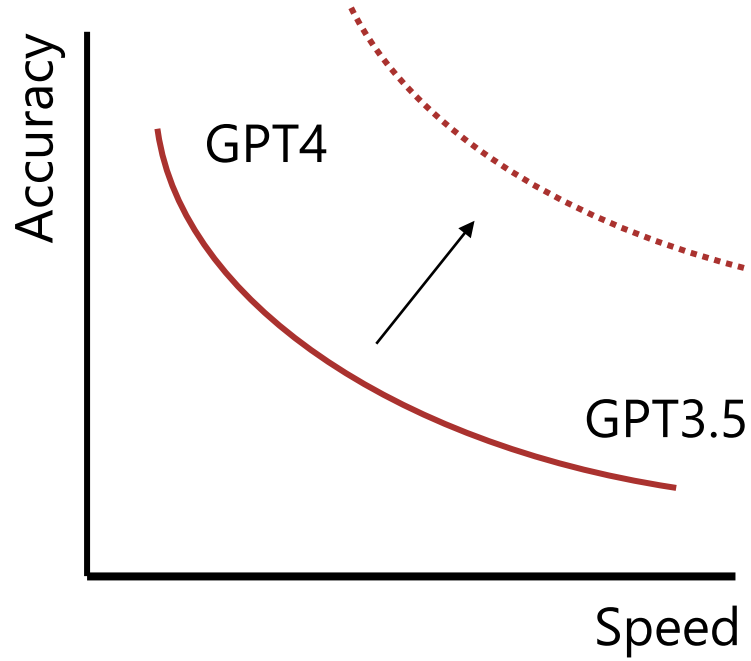
Before	After
Before 2012, ir covers the total assets…	Before 2012, it covers the total assets…
The series on commercial property prices is sourced from Central Bank of…	The series on commercial property prices is sourced from the Central Bank of…
… the source is the historical table A2 and before 1969, 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

Requirements, challenges and risks

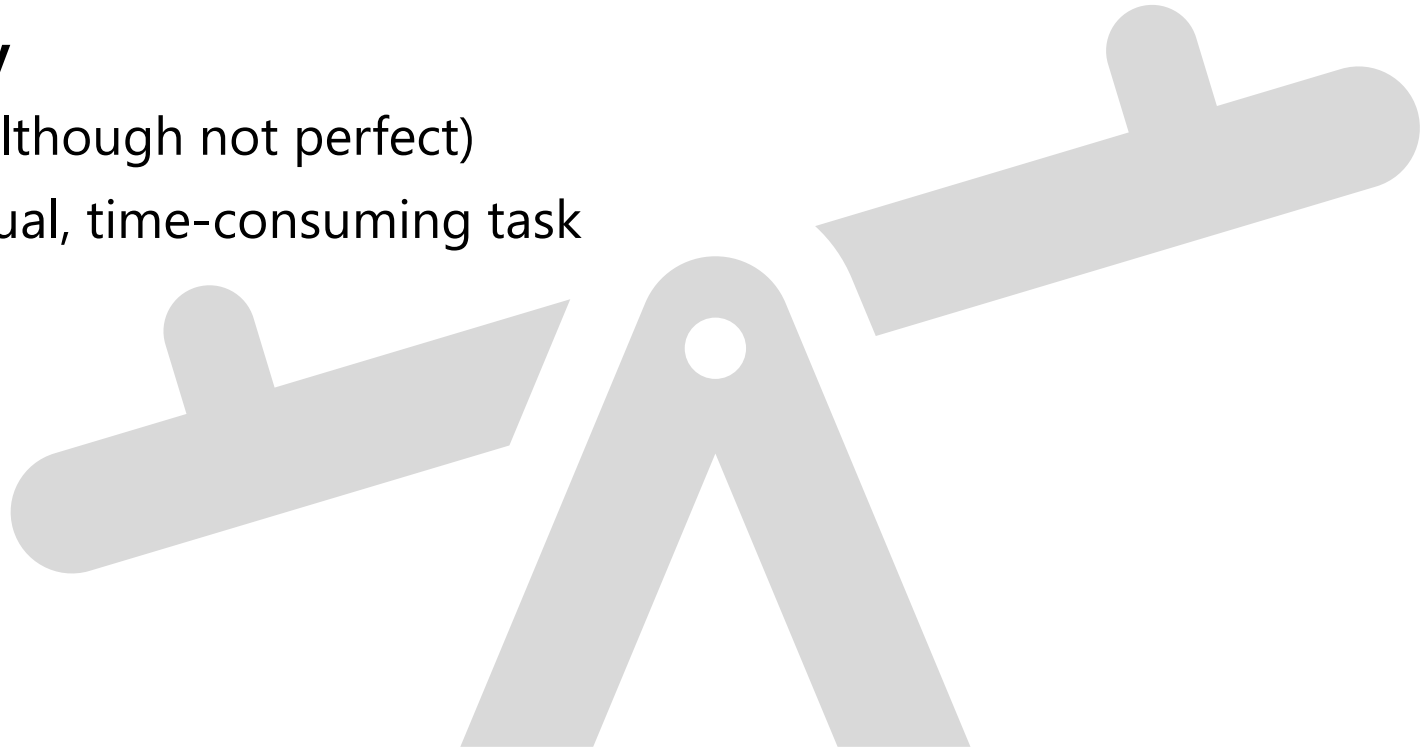
- Restricted to **public** information
- **Low reproducibility** but business case is mostly **one-off**
- Dependency on an external service
- Performance vs accuracy trade-offs
- Human-in-the-loop!
  - The only safe way of onboarding LLMs in their current form
  - Version control is key



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!