Not Just Another Black Box:
AI in the Age of Trusted Statistics

Sónia Quaresma
Instituto Nacional de Estatística

What Has Generative AI Brought to the Table for Official Statistics?

Large Language Models

These technologies offer new capabilities that can:

- transform how statistical agencies process data;
- build infrastructure;
- interact with users;

Powerful capabilities

Support innovation and efficiency

Extracting Structure from Unstructured Data:

- GenAI can analyze free-text inputs (like survey comments or administrative records) and extract relevant structured variables, enabling statistical use of previously untapped data sources

Powerful capabilities

Support innovation and efficiency

Processing Multiple Data Modalities:

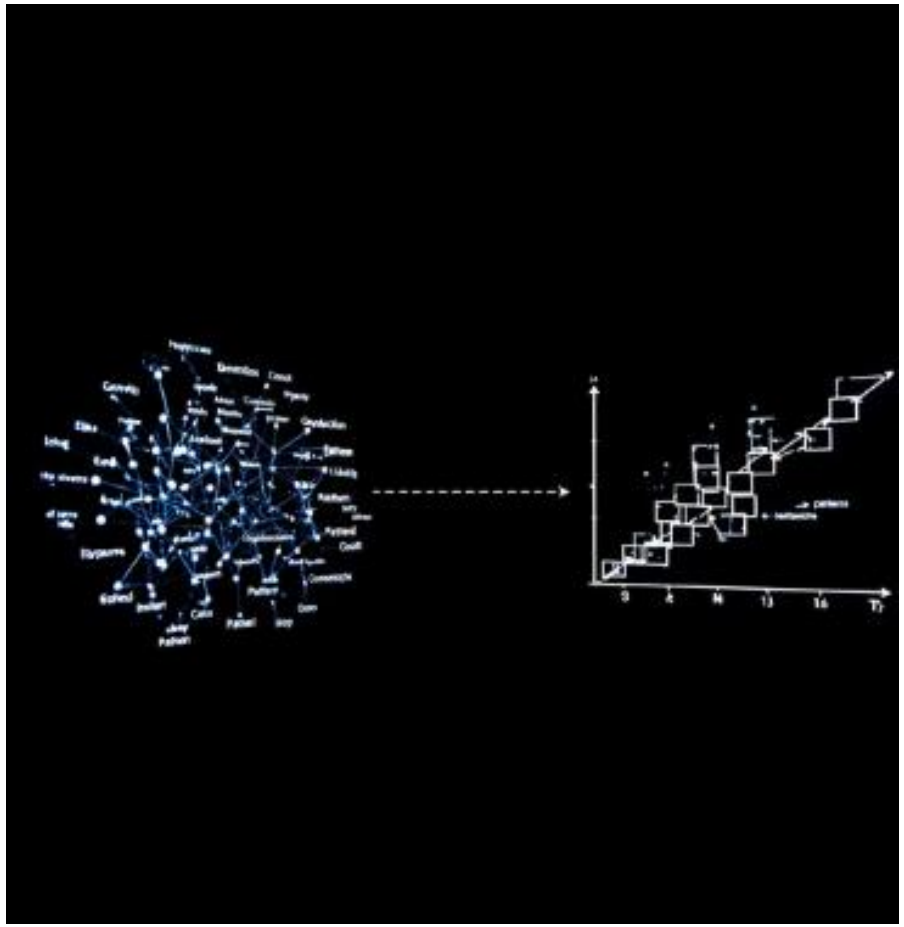
- Modern GenAI tools can interpret not just text, but also audio, images, and video, potentially expanding the scope of data collection and integration (e.g., image classification, speech-to-text processing in interviews, video metadata analysis)

Powerful capabilities

Support innovation and efficiency

Support for Causality and Pattern Detection:

- While not a substitute for formal causal modelling, LLMs can surface relationships or hypotheses from large datasets that may guide further statistical analysis or exploratory work

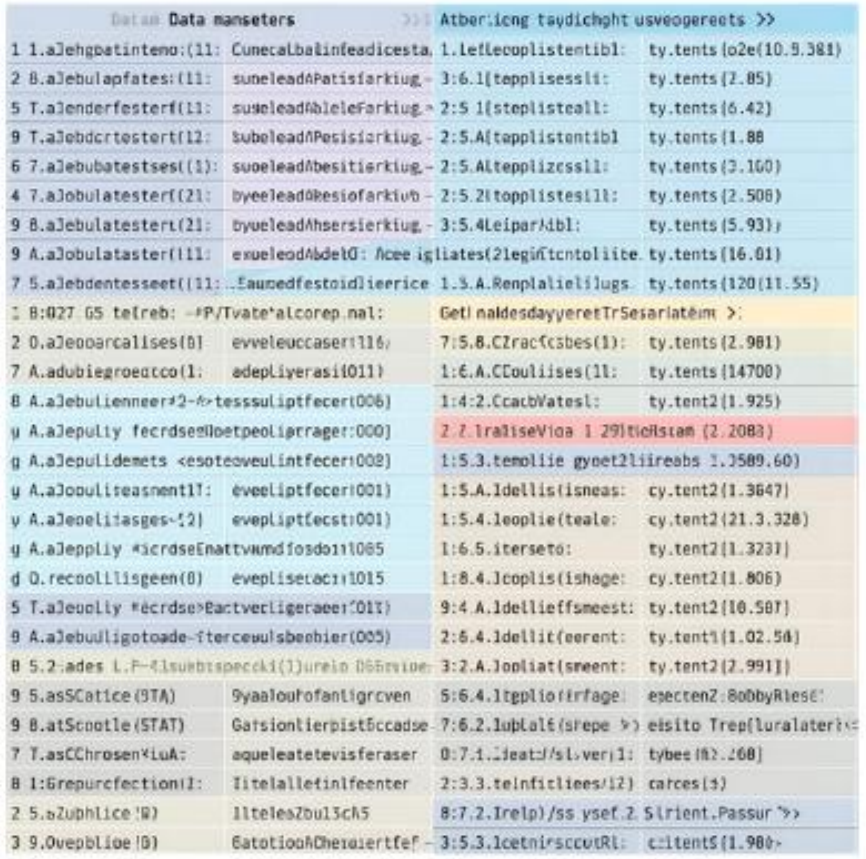


Powerful capabilities

Support innovation and efficiency

Semantic Interoperability:

- GenAI helps harmonize data from different sources by aligning schema definitions, resolving naming conflicts, and disambiguating contextual meanings, key for cross-country or multi-source statistical projects.



Powerful capabilities

Support innovation and efficiency

Contextual Aggregation and Disambiguation:

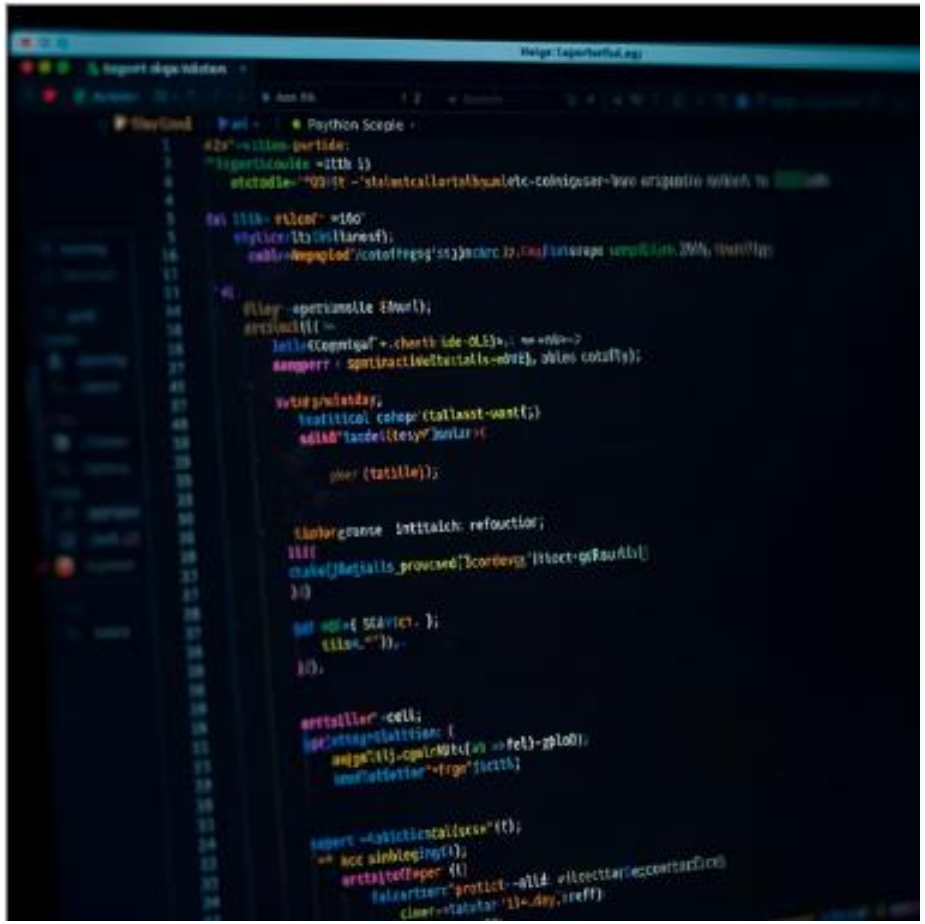
- By understanding the broader linguistic context, LLMs can aggregate values intelligently or clarify ambiguous terms.

Powerful capabilities

Support innovation and efficiency

Coding Made Easier:

- LLMs can generate statistical code (e.g., in R, Python, SQL) based on natural language prompts, translate code between languages, or suggest efficient workflows, accelerating development and prototyping



Powerful capabilities

Support innovation and efficiency

Synthetic Data Generation:

- GenAI can produce realistic synthetic datasets useful for testing methods, system development, or training purposes, while helping protect the privacy of real individuals.

Significant Risks
For Official Statistics

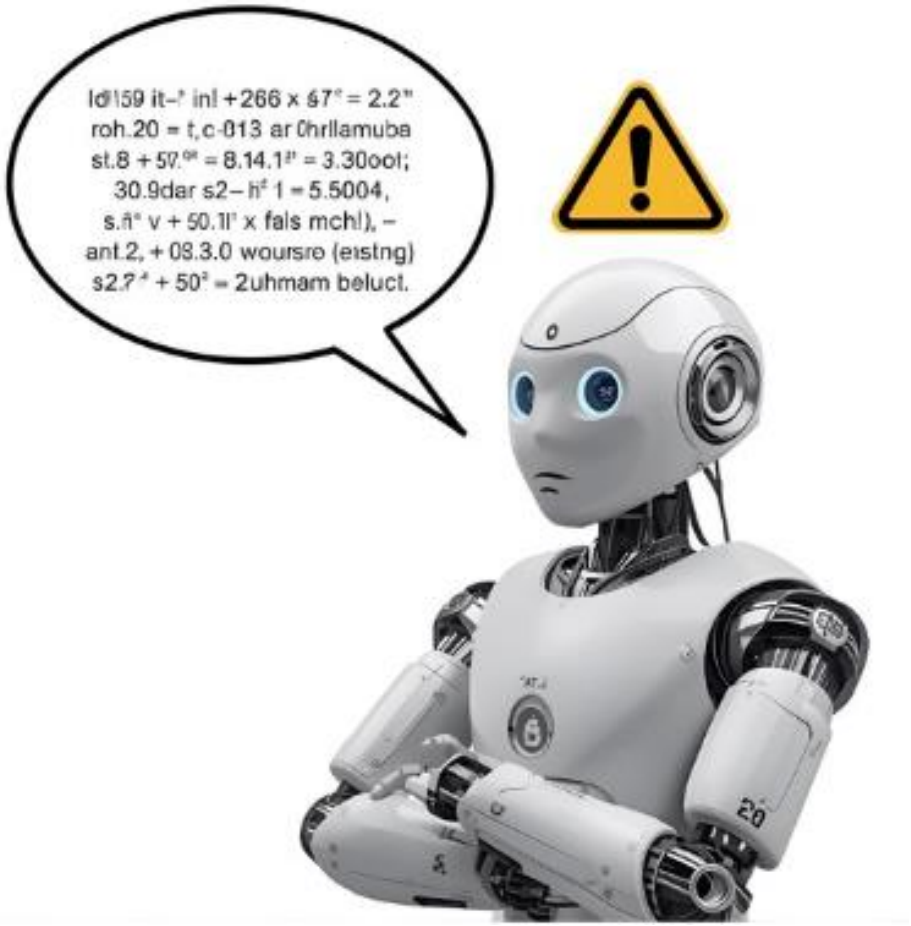
particularly sensitive in this field are:

- Trust
- Transparency
- Accuracy

Significant Risks For Official Statistics

Uncontrolled Uncertainty and Hallucinations:

- LLMs may produce plausible sounding but entirely incorrect or invented content, known as hallucinations, posing a risk of misinformation if outputs are not validated.



Significant Risks
For Official Statistics

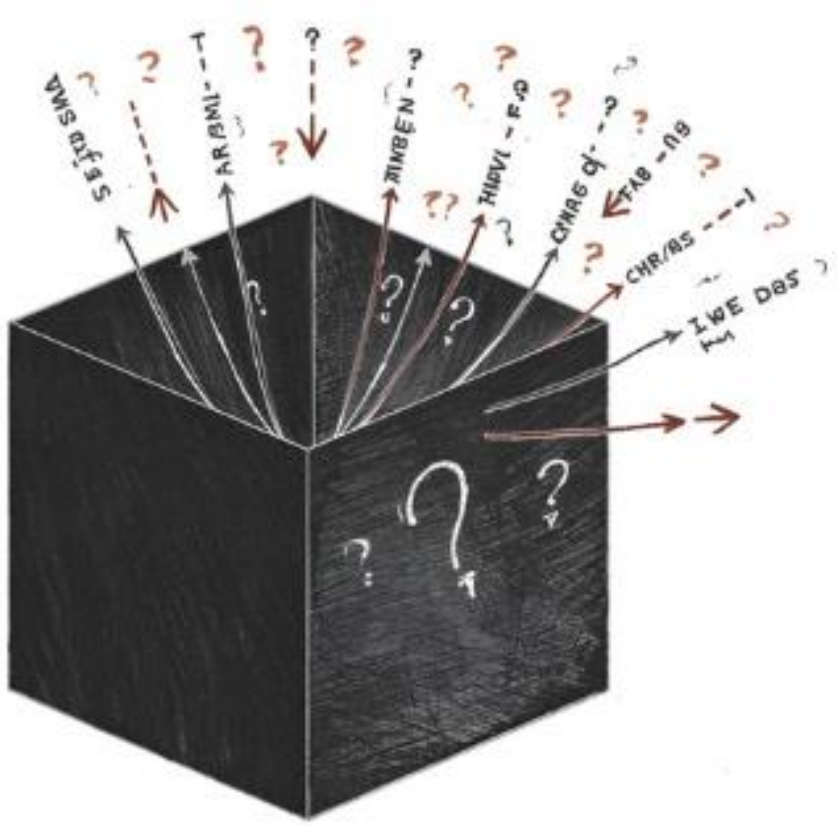
Bias:

- Inherited from their training data, LLMs may produce biased or unfair outputs, which can conflict with the principles of impartiality and fairness that underpin official statistics

Significant Risks For Official Statistics

Lack of Transparency:

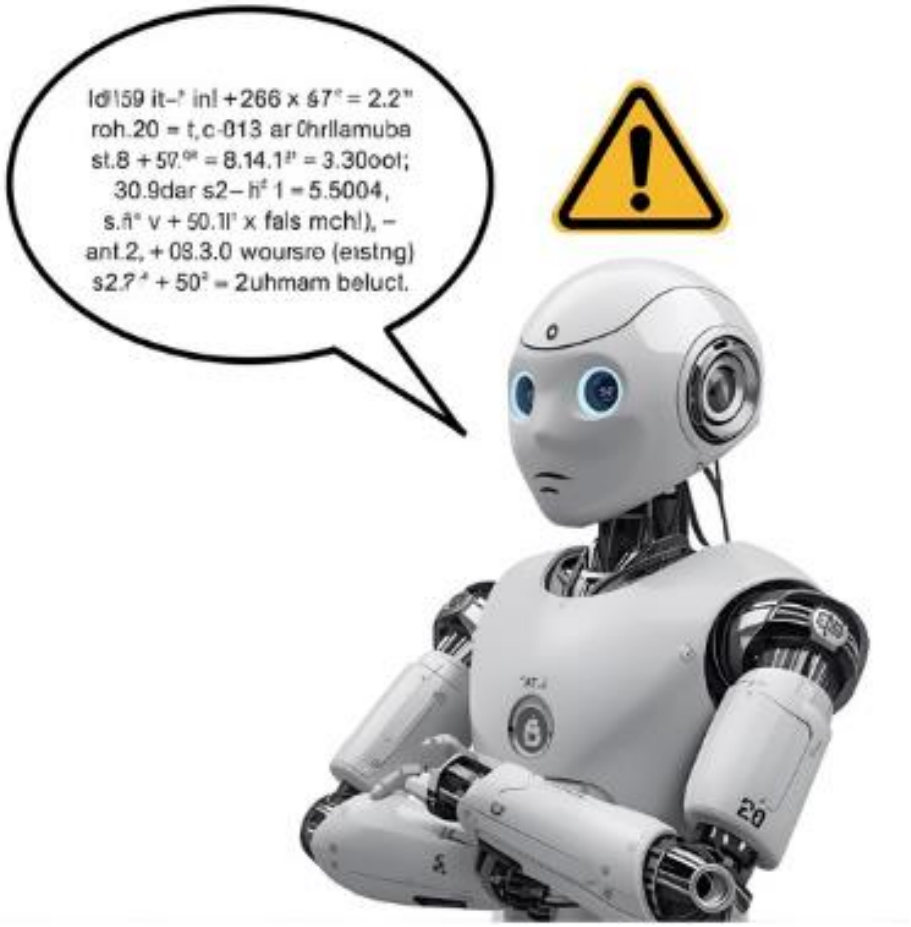
- LLMs are often considered “black boxes,” making it difficult to understand or trace the rationale behind specific outputs



Significant Risks For Official Statistics

Numerical Inaccuracy:

- GenAI models are poor at performing calculations or interpreting numeric data accurately, making them unsuitable for tasks requiring exact values or arithmetic precision.



Significant Risks
For Official Statistics

Limited Reasoning Capabilities:

- While LLMs can reproduce patterns seen in training data, they often fail at logical reasoning, deduction, and generalising to unseen situations, limiting their utility in analytical tasks requiring inference or abstraction

Significant Risks For Official Statistics

Suggestive but Not Self-Solving:

- LLMs can suggest code structures or outline problem-solving steps but cannot execute code or evaluate its correctness. They are assistive, not autonomous decision-makers.

Significant Risks For Official Statistics

High Computational Costs:

- Running GenAI models, especially at national scale, requires significant infrastructure, energy, and maintenance, raising questions of cost-efficiency and sustainability.

Significant Risks For Official Statistics

Lack of Theoretical Guarantees:

- Unlike traditional statistical models, GenAI tools typically lack the formal mathematical foundations needed for reproducibility, error bounds, or inference guarantees.

Significant Risks
For Official Statistics

Scalability Challenges with Big Data:

- GenAI models are not inherently optimized for large-scale data processing. Integrating them into big data pipelines can be costly and technically demanding.

Significant Risks For Official Statistics

Vulnerability to Prompt Hacking:

- LLMs can be manipulated through carefully crafted inputs to produce misleading or inappropriate outputs, raising concerns about reliability, especially in open-ended tools.

Significant Risks For Official Statistics

Privacy and Confidentiality Risks:

- If not properly controlled, LLMs may unintentionally reproduce sensitive information seen during training, which could violate confidentiality policies fundamental to official statistics.

Should We Still Make the Case for Using LLMs?

Aligning Use with Statistical Principles

Accuracy and Reliability:

- Outputs must be verified, contextualized, and never used as the sole source of truth for statistical indicators.

Should We Still Make the Case for Using LLMs?

Aligning Use with Statistical Principles

Relevance and Timeliness:

- LLMs can help speed up interaction and document processing but must rely on curated, up-to-date sources to avoid spreading outdated information.

Should We Still Make the Case for Using LLMs?

Aligning Use with Statistical Principles

Transparency and Impartiality:

- Clear documentation of the models, training data, and limitations is essential to maintain trust and ensure fair use.

Should We Still Make the Case for Using LLMs?

Aligning Use with Statistical Principles

Explainability:

- Mechanisms should be implemented to trace how conclusions were reached, particularly when used to assist decision-making or public communication.

Should We Still Make the Case for Using LLMs?

Aligning Use with Statistical Principles

Accessibility:

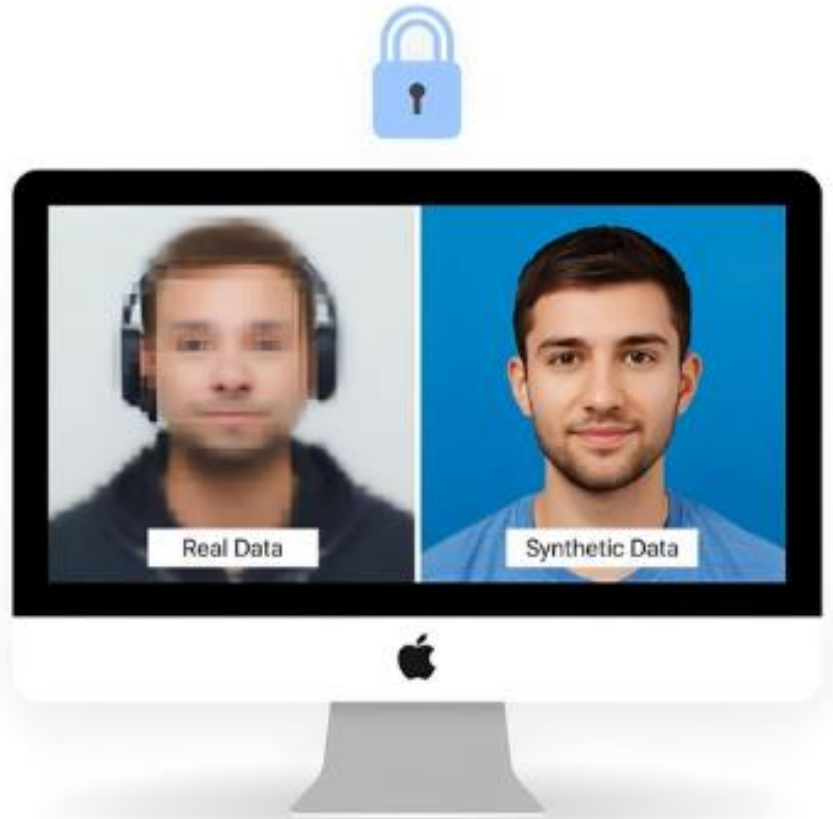
- Well-designed LLM interfaces can make statistical information more approachable, especially for non-expert users.

Should We Still Make the Case for Using LLMs?

Aligning Use with Statistical Principles

Confidentiality:

- All implementations must rigorously safeguard private data, preventing the leakage of sensitive information from training or interactions.



From Potential to Practice
Prototyping with LLMs at INE

Promising Use Cases:

- **Summarizing historical documents** - LLMs can assist in synthesizing large volumes of previously published material, such as statistical reports, metadata, and methodological notes, making it easier to extract relevant insights and provide contextual overviews.

From Potential to Practice
Prototyping with LLMs at INE

Promising Use Cases:

- **Helping users navigate published content** - With the increasing amount of information available on INE's portal, LLMs could help users find relevant tables, publications, or concepts by interpreting natural language queries and mapping them to structured resources.

From Potential to Practice
Prototyping with LLMs at INE

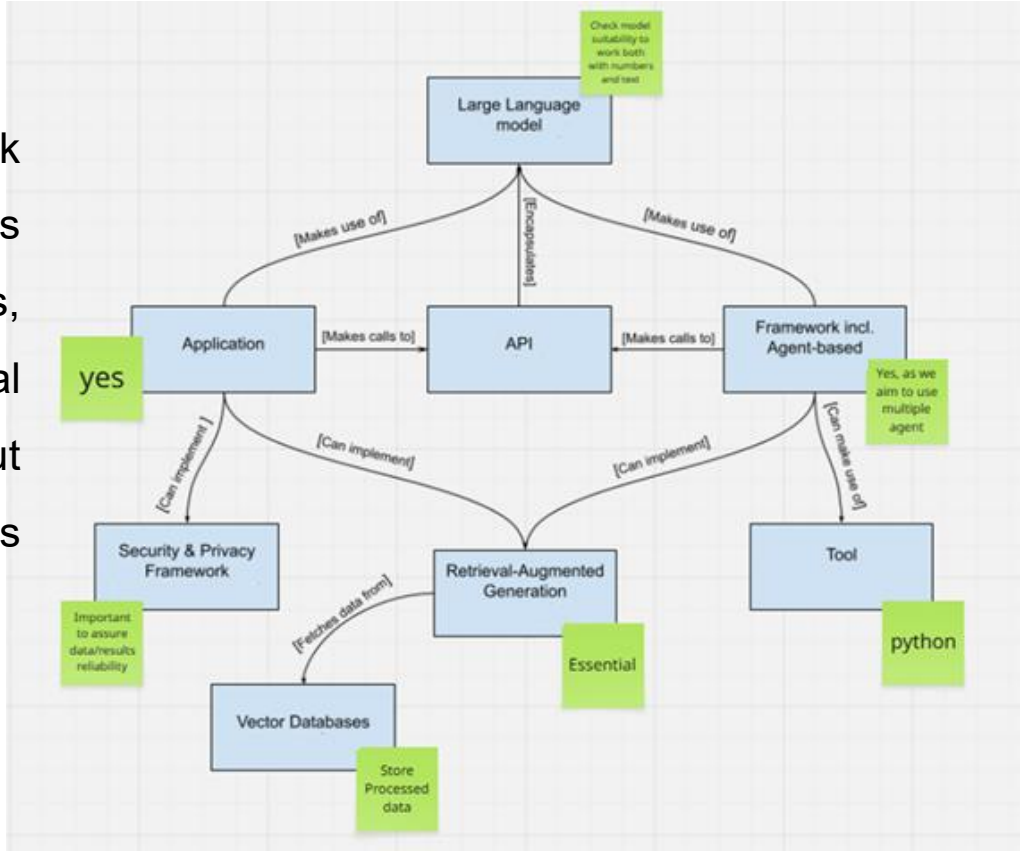
Promising Use Cases:

- **Supporting metadata management** - LLMs may support harmonization of terminology, the completion of metadata fields, or the alignment of definitions across sources, improving coherence and reducing manual work.

Bringing It All Together

A Prototype for Smarter Dissemination

At the heart of this prototype is a common yet time-consuming task faced by many statistical offices, summarizing analytical reports into clear, concise overviews for publication. These summaries, sometimes of historical data, are typically required in both the local language and English. The challenge is not just linguistic; it's about ensuring consistency, clarity, and contextual fidelity across domains.



Input

The system begins by ingesting a report, usually a PDF authored by a subject-matter expert, and extracting its content for semantic processing.

12 de junho de 2025
ATIVIDADE DOS TRANSPORTES
Estatísticas rápidas do transporte aéreo
abril 2025

MOVIMENTO DE PASSAGEIROS NOS AEROPORTOS NACIONAIS MANTEVE TENDÊNCIA DE CRESCIMENTO

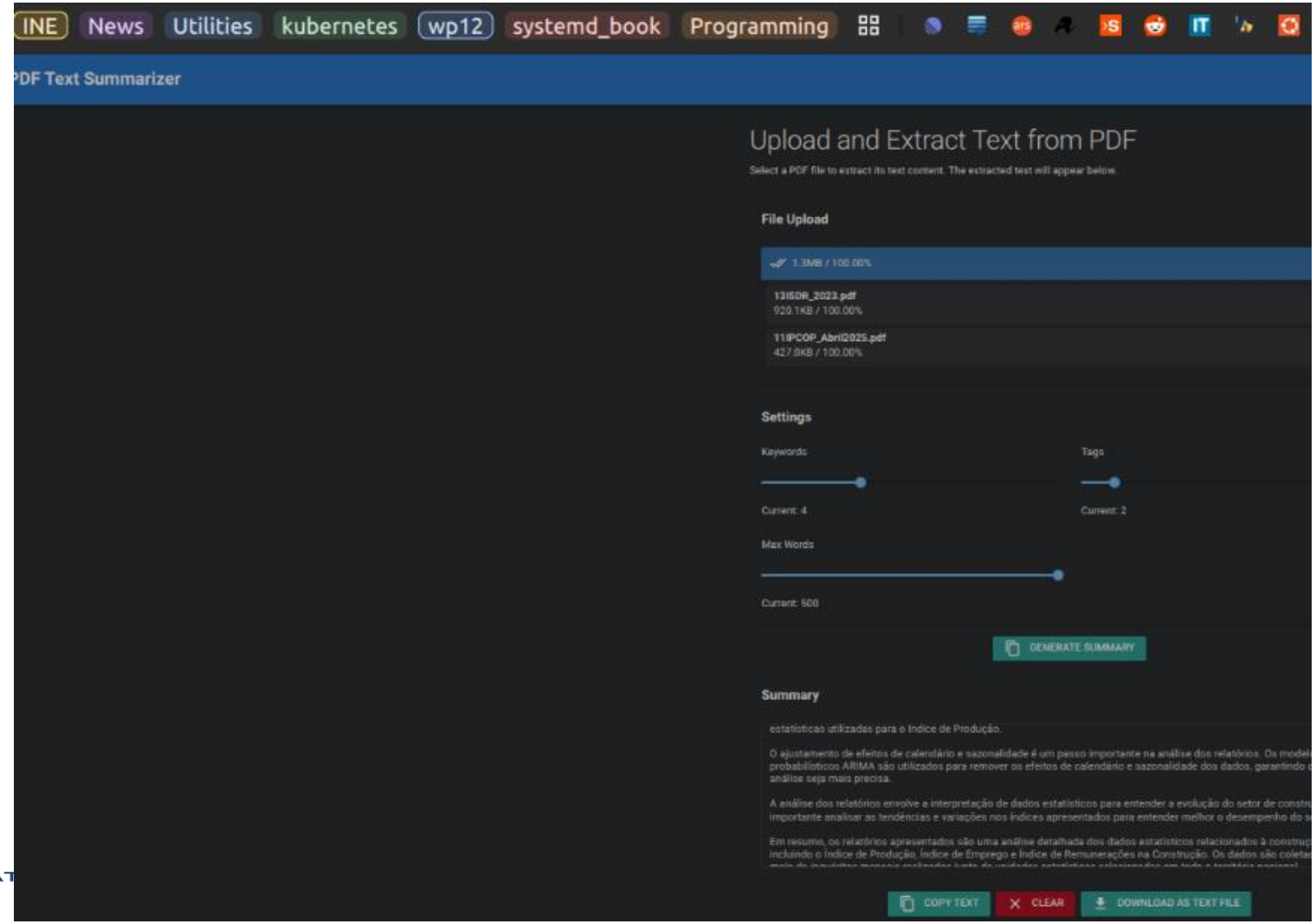
Em **abril de 2025**, nos aeroportos nacionais movimentaram-se 6,5 milhões de passageiros e 21,4 mil toneladas de carga e correio, correspondendo a variações de +8,1% e +1,7%, respetivamente, face a abril de 2024 (+2,1% e +3,9% no mês anterior, pela mesma ordem).

Em abril de 2025 registou-se o desembarque médio diário de 109,9 mil passageiros, valor superior ao registado em abril de 2024 (101,5 mil; +8,3%).

O Reino Unido passou a ser o principal país de origem e de destino dos voos, considerando os primeiros quatro

Instituto Nacional de E
Statistics Portugal

Interface


INE
News
Utilities
kubernetes
wp12
systemd_book
Programming
PDF Text Summarizer
Upload and Extract Text from PDF
Select a PDF file to extract its text content. The extracted text will appear below.
File Upload
13ISDR_2023.pdf
920.1KB / 100.00%
11IPCOP_Abril2025.pdf
427.0KB / 100.00%
Settings
Keywords
Tags
Current: 4
Current: 2
Max Words
Current: 500
GENERATE SUMMARY
Summary
COPY TEXT
CLEAR
DOWNLOAD AS TEXT FILE

Code

A Retrieval-Augmented Generation (RAG) pipeline is then deployed, enabling the LLM to work with the actual content of the report rather than relying on pretraining alone.

wp12_hackathon > dissemination_summary_prototype > summarizer_unified.py

```
129   class PDFSummarizer:
207       def process_pdf(self, pdf_path: str, use_vector_store:bool = False, document_loader:str
242           else:
252
253               # 3 - create the summary chain
254               chain = self._create_summary_chain()
255
256               result = chain.invoke({
257                   "content": context,
258                   "max_keywords": max_keywords,
259                   "max_tags": max_tags,
260                   "max_words": max_words,
261                   "out_lang": out_lang
262               })
263
264               return result
265
266
267       def _create_summary_chain(self):
268           """
269           Create a runnable sequence for direct text summarization (non-vector store mode).
270           """
271           chat_prompt, _ = self._get_shared_prompts()
272           return chat_prompt | self.llm | self._safe_json_parse
273
```

Output

```
summarizer/.venv_pdf/bin/python /home/ivo/Desktop/Trabalho/WP12/proof-of-concepts/pdf_summarizer/testing_pdf_llm.py
Com base nos relatórios apresentados, podemos observar que há uma tendência crescente no movimento de passageiros nos aer
oportos nacionais. De acordo com o Documento 1 (página 4), entre janeiro e abril de 2025, os aeroportos nacionais movimen
taram-se 6,5 milhões de passageiros, o que representa um crescimento de +8,1% em relação ao mesmo período do ano anterior
. Além disso, o desembarque médio diário de 109,9 mil passageiros em abril de 2025 é superior ao registado em abril de 20
24 (101,5 mil; +8,3%).

O Documento 2 (página 1) também confirma essa tendência, indicando que o movimento de passageiros nos aeroportos nacionai
s manteve uma tendência de crescimento. Em abril de 2025, os aeroportos nacionais registaram um desembarque médio diário
de 109,9 mil passageiros, o que é superior ao registado em abril de 2024.

Além disso, o Documento 3 (página 2) apresenta uma análise detalhada do movimento de passageiros nos aeroportos nacionais
, indicando que os passageiros desembarcados e embarcados em voos internacionais aumentaram significativamente. Em abril
de 2025, 82,5% dos passageiros desembarcados correspondiam a tráfego internacional, atingindo 2,7 milhões de passageiros
(+8,5%), e os passageiros embarcados também apresentaram um crescimento significativo.

Em resumo, a tendência recente no movimento de passageiros nos aeroportos nacionais é um crescimento significativo, com a
umento tanto no número de passageiros desembarcados quanto no número de passageiros embarcados em voos internacionais.

 Flag                                                                                    Ln 77, Col 23   Spac
```

Bringing It All Together

A Prototype for Smarter Dissemination

More than just a tech demo, this prototype offers a glimpse into how AI can be embedded thoughtfully into statistical processes, augmenting expert work, not replacing it.

It reflects a pragmatic approach to innovation at INE:

- solving concrete problems,
- testing real tools, and
- doing so in a way that's scalable across national contexts.

Bringing It All Together

A Prototype for Smarter Dissemination

As we continue to explore AI's place in official statistics, prototypes like this remind us that progress doesn't always require revolution.

Sometimes, it's about making the everyday easier, smarter, and more consistent.

Thank you for your attention!

Questions?