

Cost-Efficient and Trustworthy Exploration of Regulatory Information Extraction Using LLMs

Lisa Murch, Sr Consultant, Global Health and Life Sciences Advisory

Soundarya Palanisamy, Sr Industry Consultant, Global Health and Life Sciences Advisory



Agenda

- Introduction
- CDISC information retrieval use case
- ▣ Regulatory information use case
- ▣ Protocol generation
- Summary

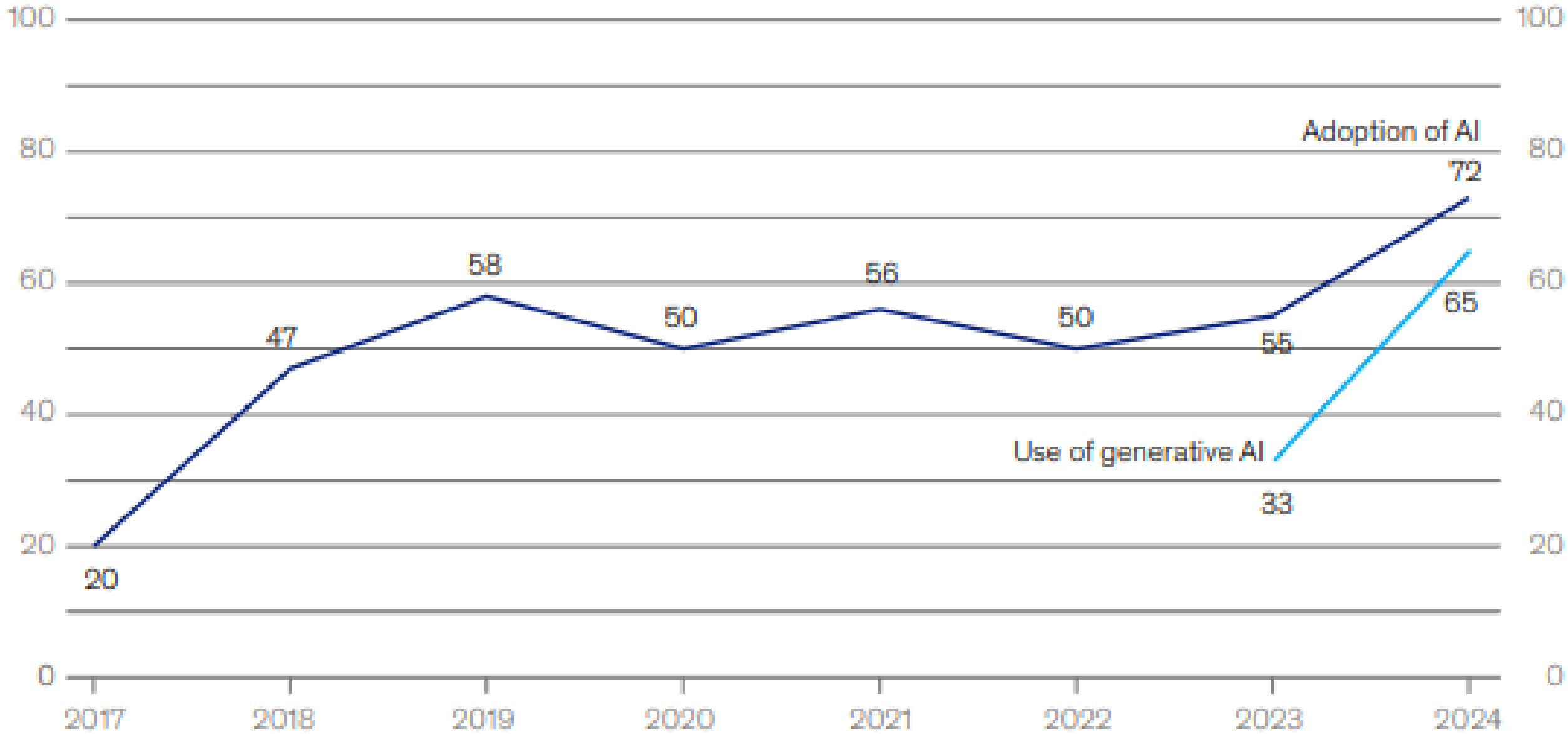
Introduction

Generative

AI

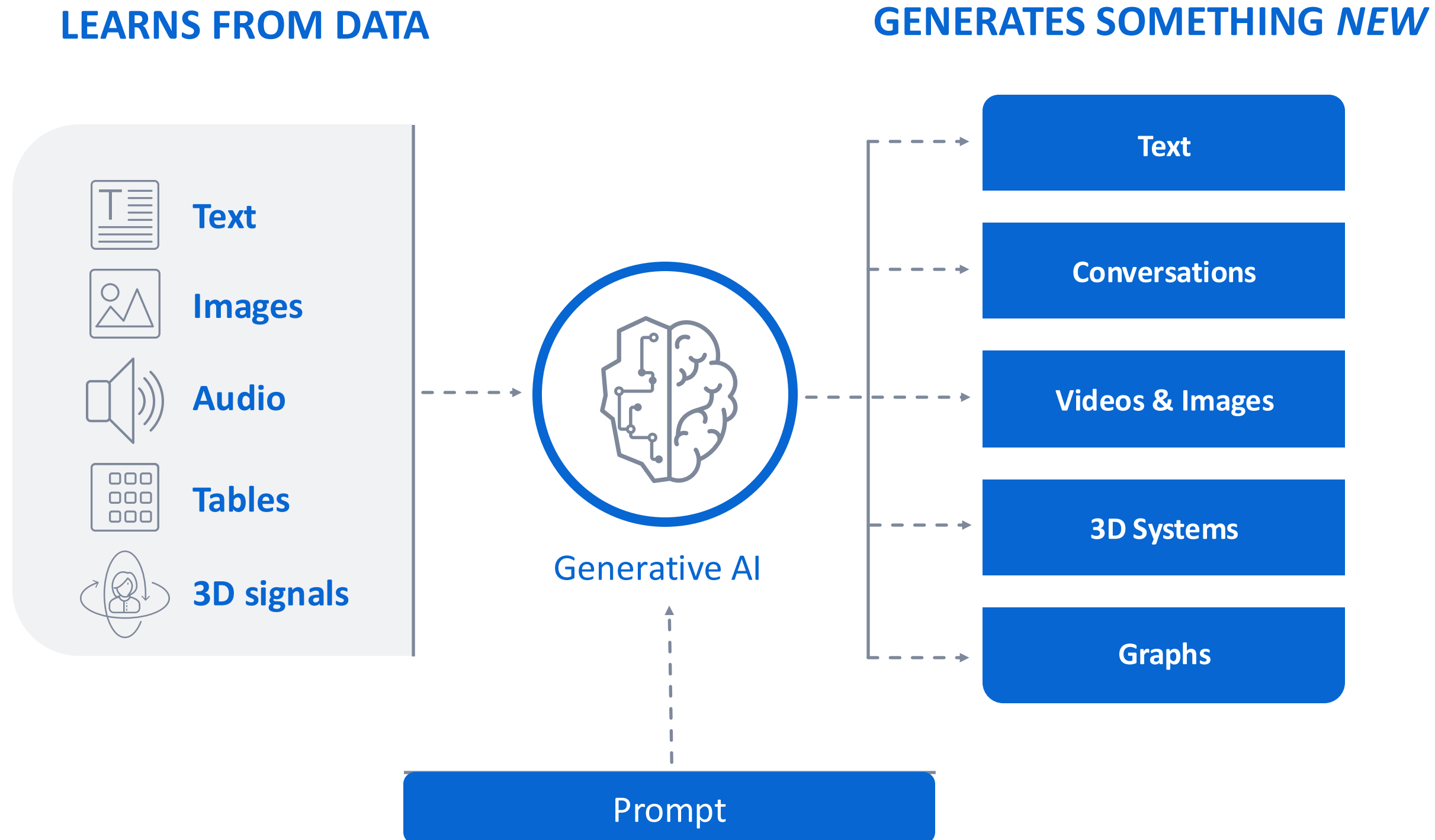
Gen AI has accelerated AI Adoption

Organizations that have adopted AI in at least 1 business function,¹ % of respondents



¹In 2017, the definition for AI adoption was using AI in a core part of the organization's business or at scale. In 2018 and 2019, the definition was embedding at least 1 AI capability in business processes or products. Since 2020, the definition has been that the organization has adopted AI in at least 1 function. Source: McKinsey Global Survey on AI, 1,363 participants at all levels of the organization, Feb 22–Mar 5, 2024

Creating New Realities



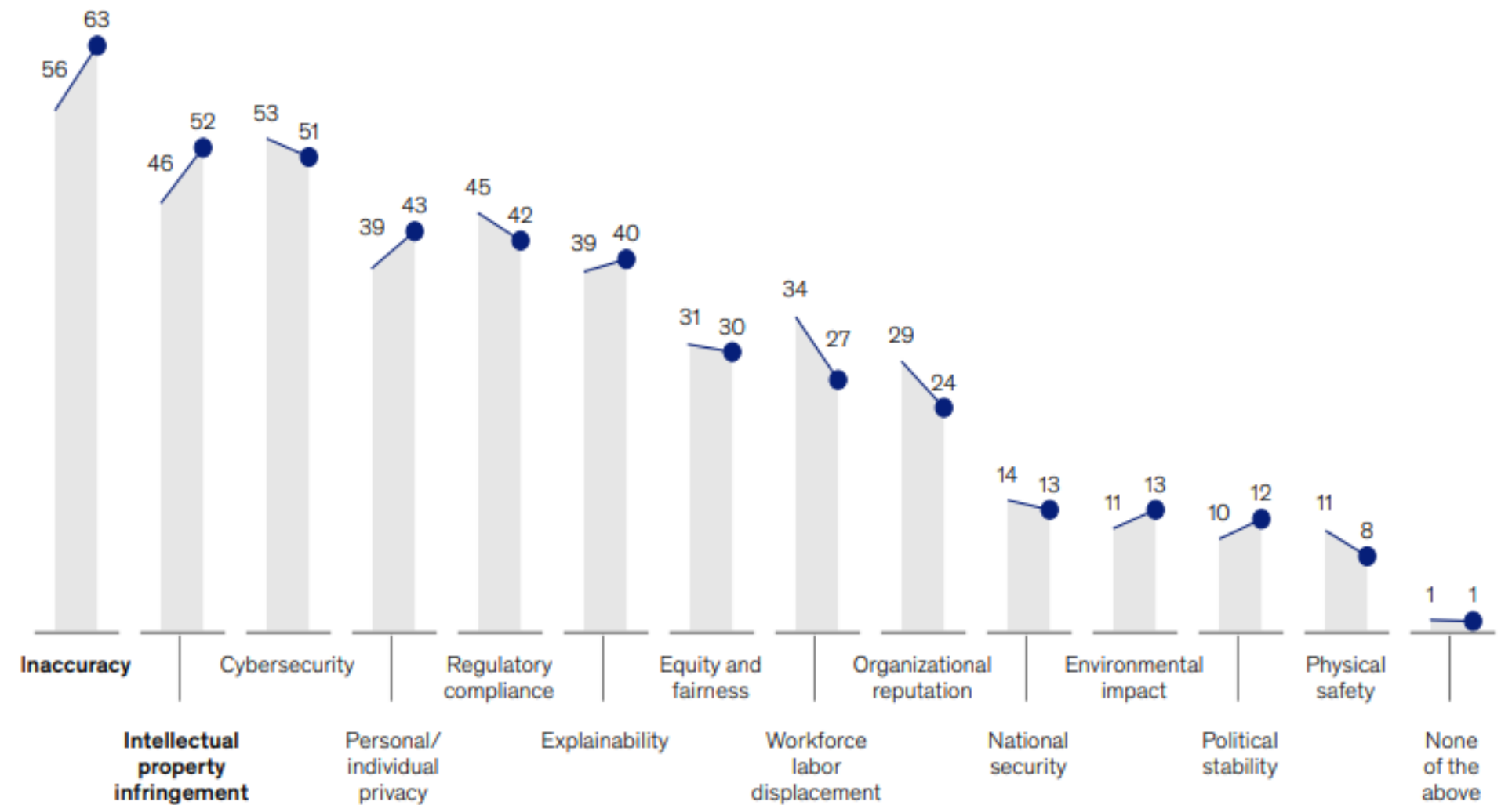
Creating New Realities



Inaccuracy and intellectual property infringement are increasingly considered relevant risks to organizations' generative AI use.

Gen AI risks that organizations consider relevant,¹ % of respondents

2023 — 2024



McKinsey & Company

CDISC information retrieval

The CDISC Case

Importance and Challenges



Complexity of the Standard

- The CDISC encompasses a set of guidelines and standards for the collection, exchange, reporting, and archiving of clinical research data.
- The breadth and complexity of these standards can be overwhelming, especially for organizations with fewer resources or less experience



Challenges in Revisions

- The continuous evolution and updates of the standard make it difficult for companies to maintain compliance without constant external support.
- Reviewing data to ensure compliance can be a lengthy and error-prone process.

Benefits

Speed of standard adoption

- LLMs can assist in interpreting and applying CDISC standards, guiding organizations through the compliance process more quickly and efficiently.
- The ability of LLMs to process large amounts of text can significantly speed up the review of documentation and the identification of non-compliance areas.

Reductions of Errors

- Through automation and the precision of LLMs, it is possible to minimize human errors in the mapping and application of standards.
- Generative AI can suggest corrections and improvements based on a vast knowledge base, enhancing data quality.

Democratization of Access

- By making LLM-based tools accessible to a broad range of organizations, barriers to entry for CDISC compliance are reduced, allowing even entities with fewer resources to meet the required standards.

Decision Support

- By providing data-based analysis and recommendations, LLMs can assist researchers and professionals in making informed decisions regarding clinical data management, improving the effectiveness of studies

Accelerating CDISC Compliance

Empowering Compliance with Regulatory Assistant

Advanced Technology Integration

- Direct access to comprehensive CDISC documentation for SDTM, ADaM.
- Utilizes a Retrieval-Augmented Generation (RAG) system for enhanced information retrieval.
- Leverages Large Language Models (LLMs) to filter and synthesize pertinent information.

Strategic Advantages of Document Assistant

- Simplifies the process of adhering to CDISC standards, making the path to compliance more straightforward.
- Enhances the users' understanding of CDISC standards, supporting informed decision-making.
- Integrates vast amounts of data to provide comprehensive insights and recommendations, enhancing decision-making processes.

Enhanced Data Quality and Compliance

- Ensures that organizations can meet CDISC requirements with ease.
- Minimizes manual efforts in navigating and applying compliance guidelines.
- Empowers users to focus on research outcomes rather than compliance processes.

CDISC Documentation Q&A

1^o Question

Compliance Recommendation App

**Recommendation
Use Case
DM dataset**

Regulatory information use case



Our goal is the **structured extraction of information** for downstream analytical tasks.

SMPCs in a nutshell

Health Products Regulatory Authority

Summary of Product Characteristics

1 NAME OF THE MEDICINAL PRODUCT

Xymel Comp 37.5mg/325mg Film-Coated Tablets

2 QUALITATIVE AND QUANTITATIVE COMPOSITION

Each film-coated tablet contains 37.5 mg tramadol hydrochloride and 325 mg paracetamol

Excipient with known effect
Each film-coated tablet contains 0.34 mg sodium.

For the full list of excipients, see section 6.1.

3 PHARMACEUTICAL FORM

Film-coated tablet
Light yellow, oblong, biconvex, film-coated tablet.

4 CLINICAL PARTICULARS

4.1 Therapeutic indications

Tramadol/paracetamol is indicated for the symptomatic treatment of moderate to severe pain.

The use of tramadol/paracetamol should be restricted to patients whose moderate to severe pain is considered to require a combination of tramadol and paracetamol (see also section 5.1).

4.2 Posology and method of administration

Posology
The use of tramadol/paracetamol should be restricted to patients whose moderate to severe pain is considered to require a combination of tramadol and paracetamol.

The dose should be adjusted to intensity of pain and the sensitivity of the individual patient. The lowest effective dose for analgesia should generally be selected. The total dose of 8 tablets (equivalent to 300 mg tramadol hydrochloride and 2,600 mg paracetamol) per day should not be exceeded. The dosing interval should not be less than six hours.

Adults and adolescents (12 years and older)
An initial dose of two tablets of tramadol/paracetamol is recommended. Additional doses can be taken as needed, not exceeding 8 tablets (equivalent to 300 mg tramadol and 2,600 mg paracetamol) per day.
The dosing interval should not be less than six hours.

Tramadol/paracetamol should under no circumstances be administered for longer than is strictly necessary (see also section 4.4). If repeated use or long term treatment with tramadol/paracetamol is required as a result of the nature and severity of the illness, then careful, regular monitoring should take place (with breaks in the treatment, where possible), to assess whether continuation of the treatment is necessary.

Paediatric population
The effective and safe use of tramadol/paracetamol has not been established in children below the age of 12 years. Treatment is therefore not recommended in this population.

Elderly
A dose adjustment is not usually necessary in patients up to 75 years without clinically manifest hepatic or renal insufficiency. In elderly patients over 75 years elimination may be prolonged. Therefore, if necessary the dosage interval is to be extended according to the patient's requirements.

02 August 2022 CRN00CX6D Page 1 of 11

“A Summary of Product Characteristics, is a comprehensive document that provides detailed information about a medicinal product. [...] includes critical information such as:

- Composition and form
- Therapeutic indications
- Side effects
- [...]”

- written by ChatGPT (GPT4)

Extraction from Regulatory Documents

Health Products Regulatory Authority

Summary of Product Characteristics

1 NAME OF THE MEDICINAL PRODUCT

Xymel Comp 37.5mg/325mg Film-Coated Tablets

Name

2 QUALITATIVE AND QUANTITATIVE COMPOSITION

Each film-coated tablet contains 37.5 mg tramadol hydrochloride and 325 mg paracetamol

Composition

Excipient with known effect

Each film-coated tablet contains 0.34 mg sodium.

For the full list of excipients, see section 6.1.

3 PHARMACEUTICAL FORM

Film-coated tablet

Light yellow, oblong, biconvex, film-coated tablet.

Form

4 CLINICAL PARTICULARS

4.1 Therapeutic indications

Tramadol/paracetamol is indicated for the symptomatic treatment of moderate to severe pain.

Indications

The use of tramadol/paracetamol should be restricted to patients whose moderate to severe pain is considered to require a combination of tramadol and paracetamol (see also section 5.1).

4.2 Posology and method of administration

Posology

The use of tramadol/paracetamol should be restricted to patients whose moderate to severe pain is considered to require a combination of tramadol and paracetamol.

The dose should be adjusted to intensity of pain and the sensitivity of the individual patient. The lowest effective dose for analgesia should generally be selected. The total dose of 8 tablets (equivalent to 300 mg tramadol hydrochloride and 2,600 mg paracetamol) per day should not be exceeded. The dosing interval should not be less than six hours.



Why do we need to bring in an LLM for that?

--

NLP-veterans?

VALID QUESTION

Let's consider the following...

4.8 Undesirable effects

Paracetamol
Adverse effects of paracetamol are rare, but hypersensitivity including skin rash may occur. There have been a few reports of blood dyscrasias including thrombocytopenia, leucopenia, pancytopenia, neutropenia and agranulocytosis, but these were not necessarily causally related to paracetamol. Acute pancreatitis after ingestion of above normal amounts.

Phenylephrine hydrochloride

High blood pressure with headache and vomiting, probably only in overdosage. Rarely palpitations. Also, rare reports of allergic reactions and occasionally urinary retention in males.

Reporting of suspected adverse reactions

Reporting suspected adverse reactions after authorisation of the medicinal product is important. It allows continued monitoring of the benefit/risk balance of the medicinal product. Healthcare professionals are asked to report any suspected adverse reactions via HPRA Pharmacovigilance, Earlsfort Terrace, IRL - Dublin 2; Tel: +353 1 6764971; Fax: +353 1 6762517. Website: www.hpra.ie; E-mail: medsafety@hpra.ie

SMPC: Lemsip Decongestant & Flu Lemon Tablets



Yes, these are all the flags, not all the languages



No standardization

There are too many ways to describe the same thing



Contextual dependencies

What active ingredient?
how rare?....



This doesn't scale...

1/9 sections of 1/10 chapters in 1/24 languages (in the EU)



**Soo... NLP and text
analytics is *dead*?**

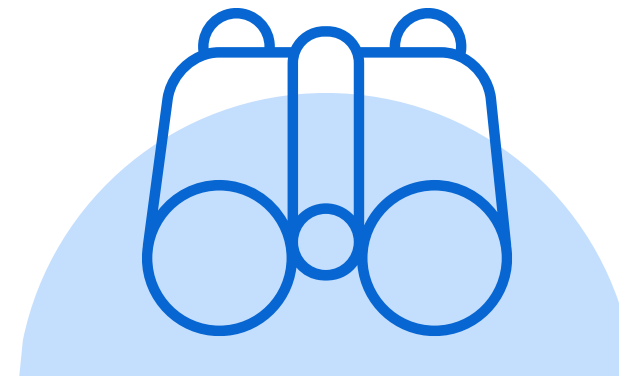
NAIVE QUESTION

No, it's not
For some good reasons



The simplest: Cost

With 3500-8000 tokens per document 1k documents cost 105-240\$, for the input only. Output is often 2-3x the price per token as the input*.



Improving Accuracy

Providing more focused “context” to an LLM, dramatically improves the output quality. This is especially true for “weaker” models.



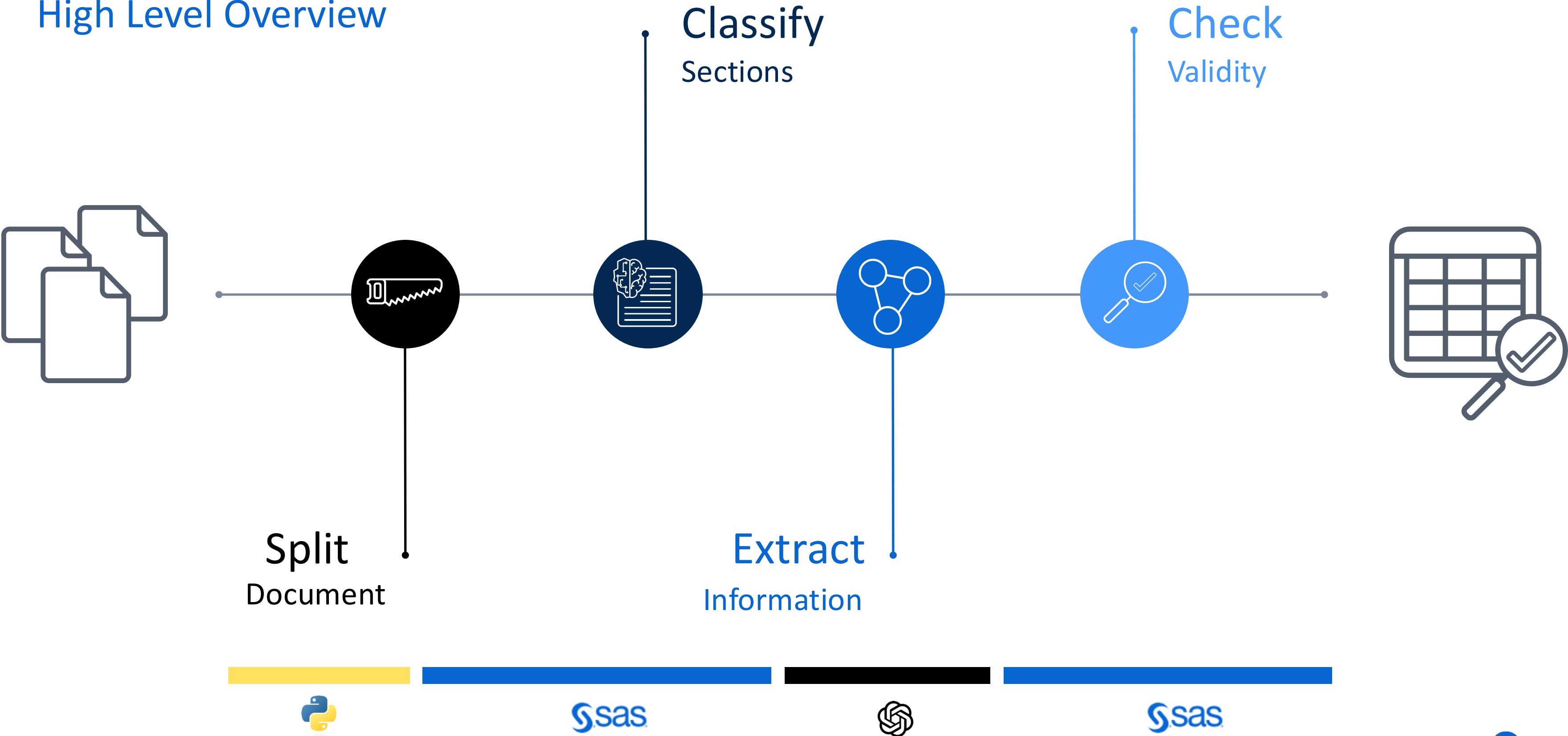
Establishing Guardrails

“*Trust but verify*” - If extractions impact downstream tasks, LLMs (and their fantasies) shouldn't work unchecked.

What we built

Processing Pipeline

High Level Overview



Steps

Type to filter list

SAS Steps : Shared

- Data (Input and Output)
- Develop
- Transform Data
- Integrate
- Statistics
- Visualize Data
- Optimization and Net...
- Data Quality
- Manage Models
- Examine Data
- Prepare Data
- Enrichment
- Statistical Process Cont...
- Machine Learning
- Econometrics
- Text Analytics

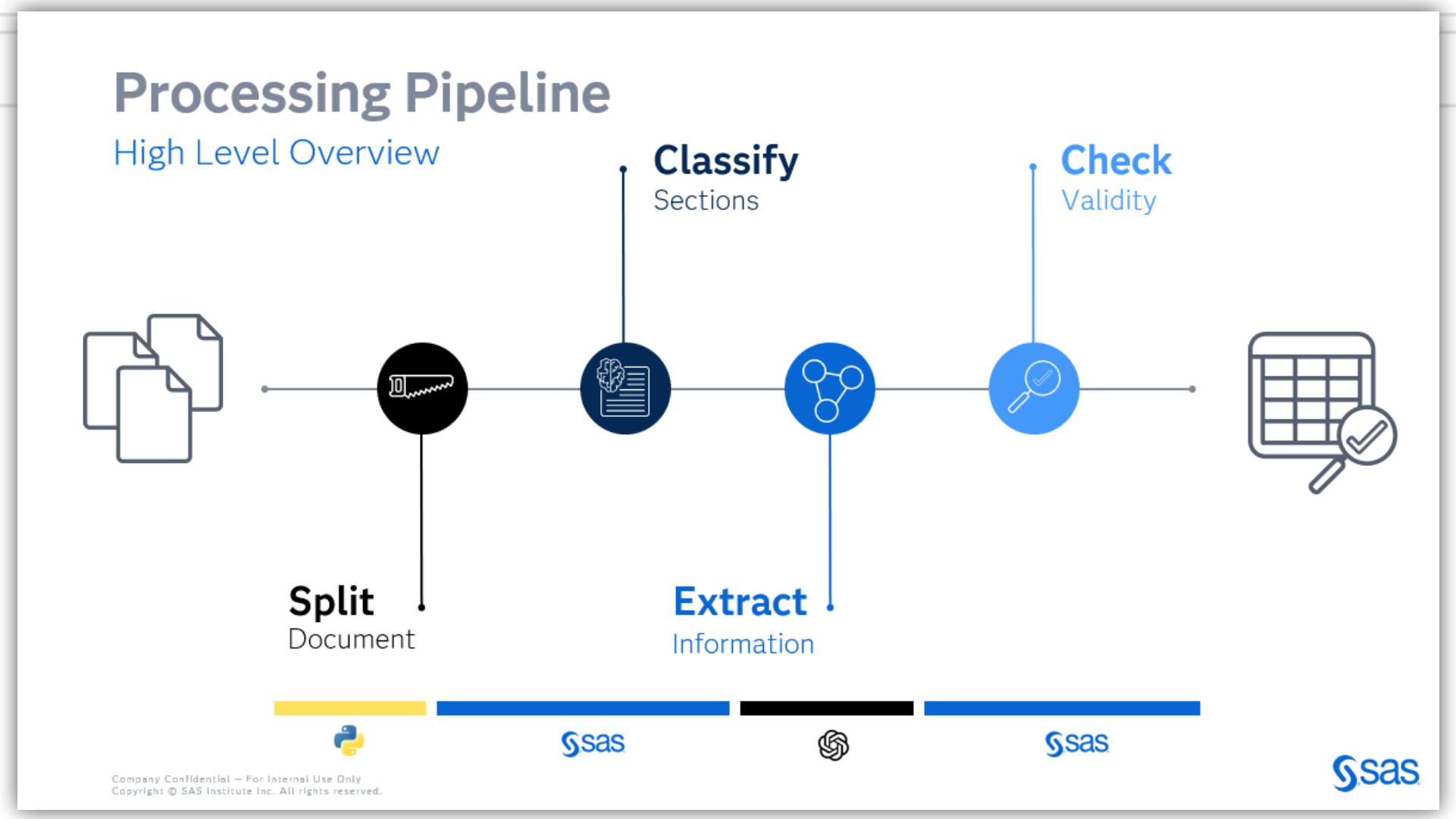
Start Page *0_Complete_Flow.flw

Run Cancel Add View

Flow Generated Code Submitted Code and Results

- SPLIT documents
- CLASSIFY sections
- EXTRACT information
- CHECK validity

A SAS Studio Flow is executing all the tasks in the processing pipeline



Steps

Type to filter list

SAS Steps : Shared

- Data (Input and Output)
- Develop
- Transform Data
- Integrate
- Statistics
- Visualize Data
- Optimization and Net...
- Data Quality
- Manage Models
- Examine Data
- Prepare Data
- Enrichment
- Statistical Process Cont...
- Machine Learning
- Econometrics
- Text Analytics

Start Page * 0_Complete_Flow.flw x +

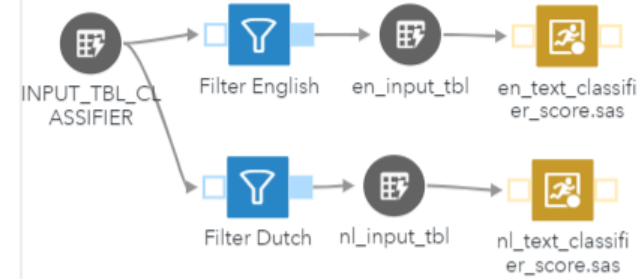
Run Cancel Add View

Flow Generated Code Submitted Code and Results

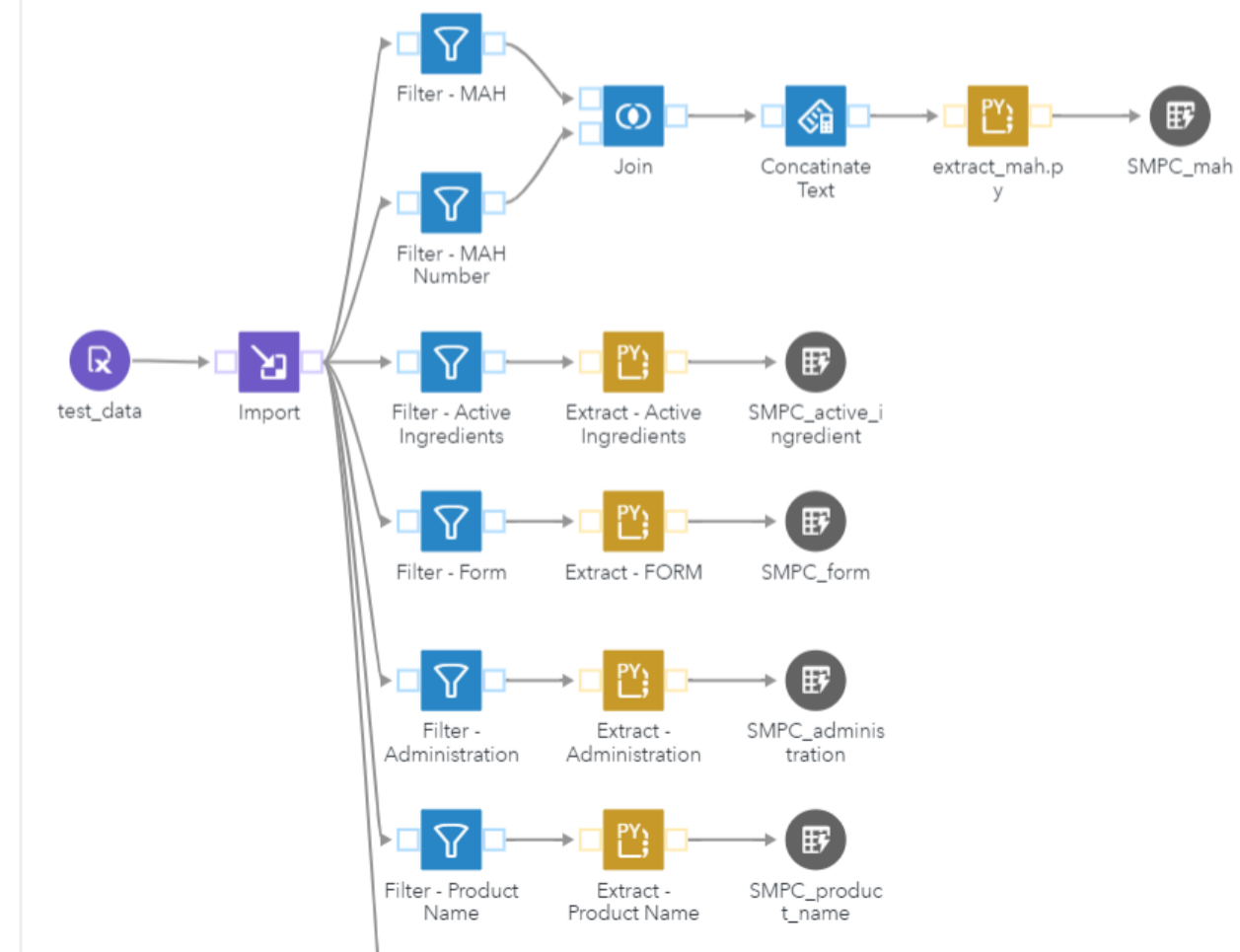
SPLIT documents



CLASSIFY sections



EXTRACT information



The SAS Studio Flow combines seamlessly:

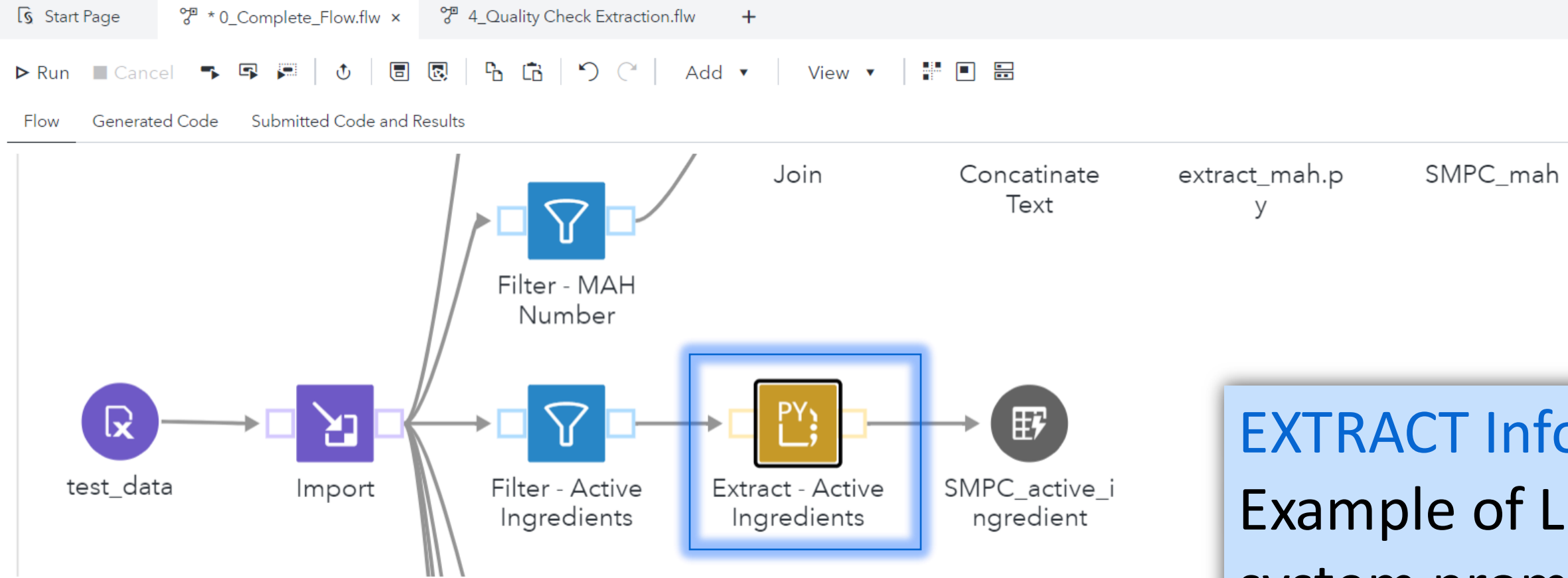
- SAS Steps
- custom steps
- SAS code
- Python code

Steps

Type to filter list

SAS Steps : Shared

- Data (Input and Output)
- Develop
- Transform Data
- Integrate
- Statistics
- Visualize Data
- Optimization and Net...
- Data Quality
- Manage Models
- Examine Data
- Prepare Data
- Enrichment
- Statistical Process Cont...
- Machine Learning
- Econometrics
- Text Analytics



EXTRACT Information:
 Example of LLM API call and
 system prompt definition

Extract - Active Ingredients

Code Node Notes

SAS Content:/Projects/SMPC/2_code/llm_extraction/extract_active_ingredients.py

```

41
42
43
44 response = client.chat.completions.create(
45     model="gpt-4-1106-preview",
46     messages=[
47         {
48             "role": "system",
49             "content": "" You are a world class software engineer and you are tasked with fixing a JSON string.
50             You are provided with the JSON string and an error message.
51             You need to fix the JSON string so that it is valid JSON.
52             Only return the fixed JSON string without any additional information.""
53         },
54         {
55             "role": "user",
56             "content": content
57         }
58     ],
59     temperature=0,
60     max_tokens=4096,
61     top_p=1
62 )

```

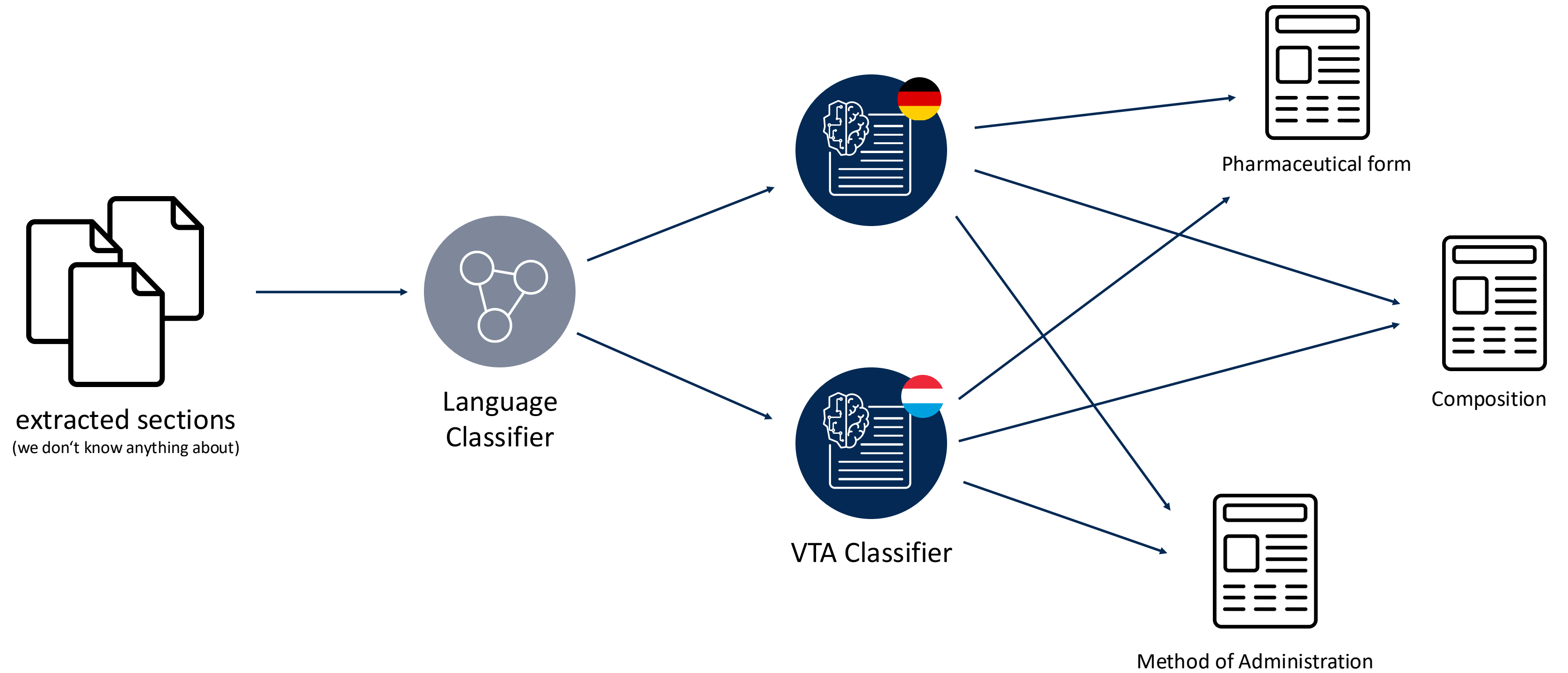
June 4, 2024, 02:22:52 PM

“Pre- Filter” – What goes into the LLM



Deciding what is relevant

Classifying sections



“Quality Checks” – output from the LLM

Trust but Verify

Applying sanity checking on extracted information



Obs	side_effects
1	Increase in liver transaminases
2	Changes of blood count such as thrombocytopenia and agranulocytosis
3	Bronchospasm (asthma induced by analgesics)
4	Hypersensitivity reactions like erythema including urticaria and anaphylactic shock
5	Serious skin reactions

1

Get extractions from LLM

Obs	rule
1	1:rule_1:(OR[NW],(AND[0.5],"increase@"),(AND[0.5],"liver transaminase@"))
2	2:rule_2:(OR[NW],(AND[0.5],"change@"),(AND[0.5],"blood count@"),(AND[0.5],"thrombocytopenia@"), (AND[0.5],"agranulocytosis@"))
3	3:rule_3:(OR[NW],(AND[0.5],"bronchospasm@"),(AND[0.5],"asthma@"),(AND[0.5],"induce@"), (AND[0.5],"analgesic@"))
4	4:rule_4:(OR[NW],(AND[0.5],"hypersensitivity reaction@"),(AND[0.5],"erythema@"),(AND[0.5],"include@"), (AND[0.5],"urticaria@"),(AND[0.5],"anaphylactic shock@"))
5	5:rule_5:(OR[NW],(AND[0.5],"serious@"),(AND[0.5],"skin reaction@"))

2

Automatically generate rules for quality check

Obs	match	_category_	confidence
1	Increase in liver transaminases	rule_1	1
2	Changes of blood count such as thrombocytopenia and agranulocytosis	rule_2	1
3	bronchospasm (asthma induced by analgesics)	rule_3	1
4	Hypersensitivity reactions like erythema including urticaria and anaphylactic shock	rule_4	1
5	serious skin reactions	rule_5	1

3

Score LTI rules, retrieve matches and calculate confidence score

Confidence Score



An Example

Source

4.8 Undesirable effects

...
Blood and lymphatic system disorders
Very rare: Changes of blood count such as thrombocytopenia and agranulocytosis
...

Possible extractions

- a) **Changes of blood count such as thrombocytopenia and agranulocytosis**
 - (OR[NW],(AND[0.5],"change@"),(AND[0.5],"blood count@"), (AND[0.5],"thrombocytopenia@"), (AND[0.5],"agranulocytosis@"))
- b) **Changes of blood count such as thrombocytopenia, leucopenia and agranulocytosis**
 - (OR[NW],(AND[0.5],"change@"),(AND[0.5],"blood count@"), (AND[0.5],"thrombocytopenia@"), (AND[0.5],"leucopenia@"),(AND[0.5],"agranulocytosis@"))
- c) **Changes of blood count such as thrombocytopenia, leucopenia, pancytopenia and agranulocytosis**
 - (OR[NW],(AND[0.5],"change@"),(AND[0.5],"blood count@"), (AND[0.5],"thrombocytopenia@"), (AND[0.5],"leucopenia@"),(AND[0.5],"pancytopenia@"), (AND[0.5],"agranulocytosis@"))

Score

1

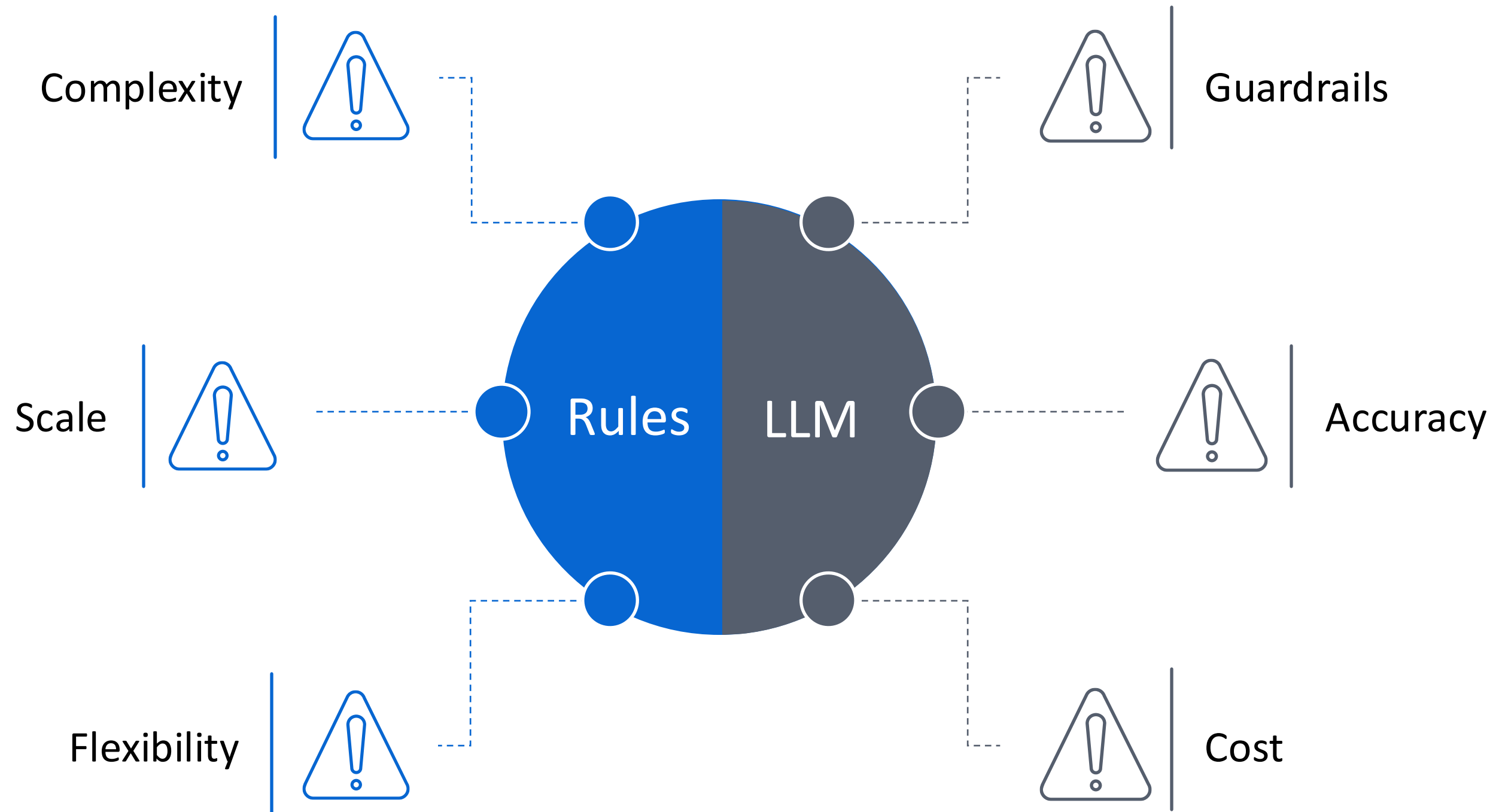
0.80

0.66



There are two sides to this story

VTA and LLMs – “Better Together”



Protocol generation



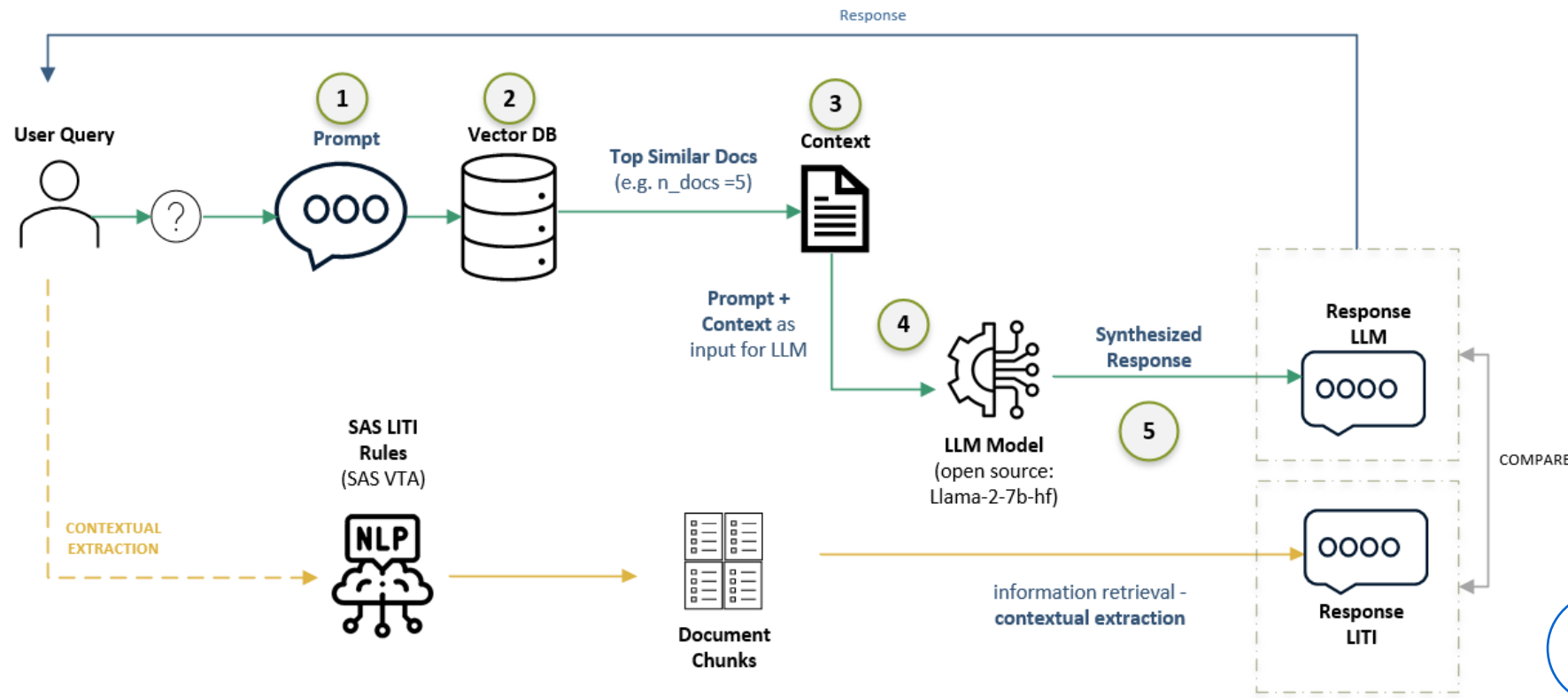
Development of USDM through translation of human-readable protocols

Jasmine Kestemont (Innovion)
Stijn Rogiers (argenx)

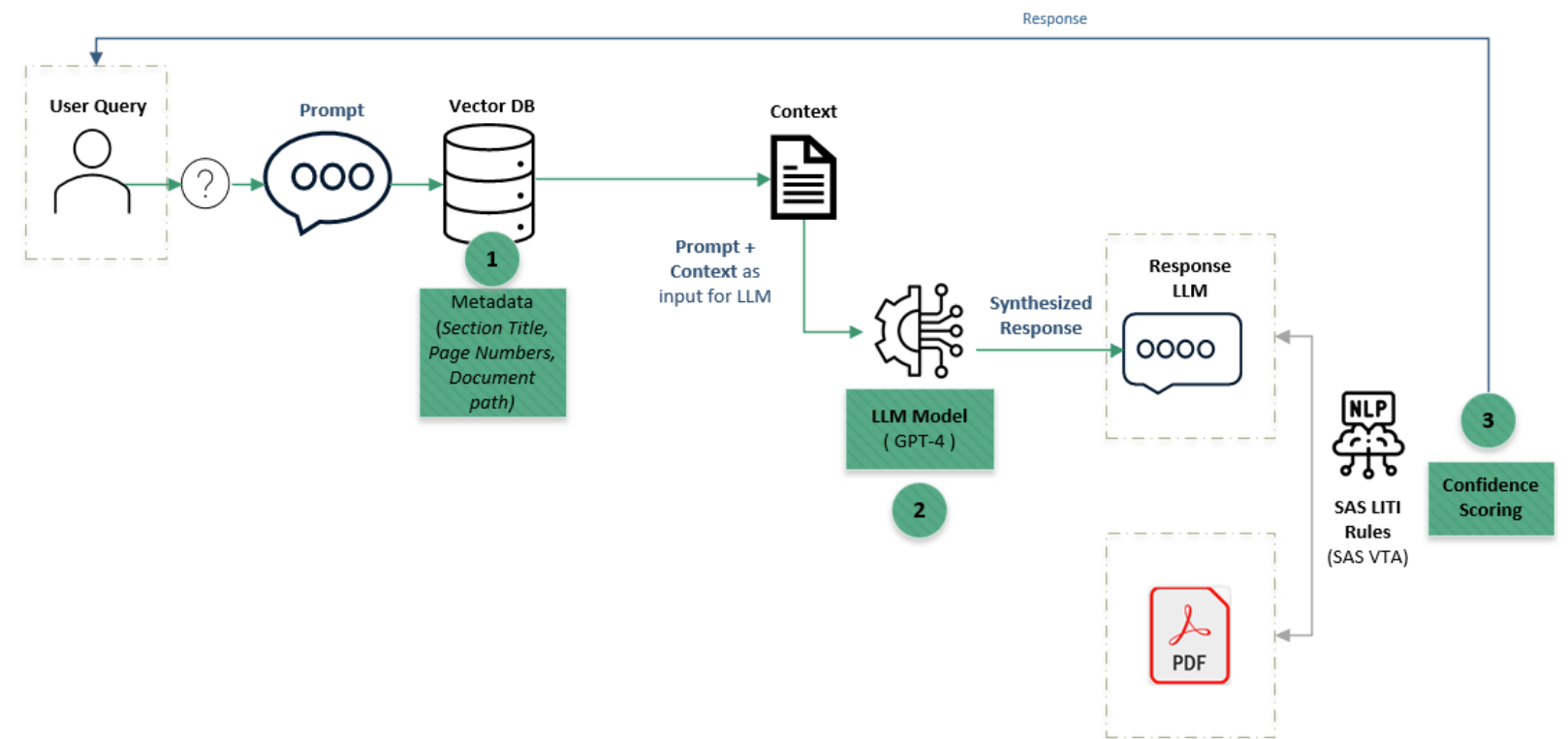
POC approach (Cont'd)

- 1) **Usage of LITI rules** for contextual information extraction
LITI includes concept rule types as well as fact rule types.
LITI is proprietary syntax from SAS.
LITI = **L**anguage **I**nterpretation for **T**extual **I**nformation
- 2) **Usage of LLM (RAG-for-LLMs)** for contextual information extraction
RAG = **R**etrieval-**A**ugmented **G**eneration
- 3) **PDF Table Extractor(s)**
- 4) **PDF Image Extractor(s)**
- 5) **Load from SAS tables into Excel and Word in an automated way**
*while retaining the link i.e. if the info in the SAS table changes, then you can just refresh the *.xlsx or *.docx file to see that new info reflected.*

1 SAS LITI (NLP) + LLM for contextual extraction



2 LLM (Open AI) + SAS LITI (NLP + Confidence Scoring)



3 Evaluating Gen AI and Text Analytics for creating a USDM

Some initial results after pre processing, highlights the extractions of inclusion and exclusion criteria by LITI rules and LLMs.

Legend :

perfect
> 75 % hit rate
between 60 % and 75 % hit rate
< 60% hit rate

Select any image to see a larger version.

Mobile users: To view the images, select the "Full" version at the bottom of the page.

	document partitioning (training / validation / test)	INCLUSION CRITERIA			EXCLUSION CRITERIA		
		Number	Number Captured With LITI	Number Captured with LLM	Number	Number Captured With LITI	Number Captured with LLM
Documents (clinical trial protocols)							
Protocol - 13 Feb 2020.pdf	TRAINING	15	15	12*	22	22	19
Protocol_22Jul2020.pdf	TEST	7	7	6	20	20	34*
ELI Lilly_NCT03421379_Diabetes.pdf		10	10	10	26	26	26
Roche_NCT04320615_COVID.pdf		7	7	7	12	12	12
CDISC_Pilot_Study.pdf		8	8	8	23	23	23

There are several challenges to overcome, such as overfitting in LITI rules and the need for extensive pre-processing.



Summary

Benefits of Leveraging Governance Layer

E.g. Using Visual Text Analytics as a pre-filter into LLMs and post- quality checks



Accuracy / Avoid Hallucinations



Time to Value



Privacy & Security



Cost Savings



Traceability



Establish Guard rails - Quality checks

Thank you