

Enhancing Pharmaceutical Quality Assurance through Large Language Models

-

A Novel Approach to Intelligent Regulatory Monitoring (ML05)

Presenter: Jake Gagnon for PHUSE US Connect 2025
Monday March 17

Authors: Hangyu Liu, Yuka Moroishi, James Melican,
Matthew Ryals, Jake Gagnon, Haleh Valian

Company: Biogen, Cambridge, MA, USA

Agenda

- ▶ Project Context, Scope, and Goal
- ▶ Overview of Existing Process and AI Improvement
- ▶ Step 1 of AI Approach
- ▶ Step 2 of AI Approach
- ▶ Summary and Next Steps

Project Context

► **Problem Statement:** Biogen needs to follow guidelines from regulatory agencies like FDA and EMA. These regulations evolve, so our internal documents must be assessed and adjusted to ensure compliance and avoid inspection findings.

► **Current Challenge:** The process of monitoring and implementing these regulatory changes is often manual, labor-intensive, and fraught with the risk of overlooking impacted Standard Operating Procedures (SOPs) due to human error.

Project Scope and Goal

- ▶ **Project Goal:** Use AI to streamline regulatory impact assessment. The AI will analyze changes, compare them with internal documents, and identify affected documents and discrepancies to ensure compliance.
- ▶ **Project Scope:** This project excludes gathering updates from agencies. A vendor will provide updates. The AI will evaluate the impact on hundreds of internal documents upon receiving updates.

Regulatory Monitoring Process without GenAI

► Human-driven Regulatory Monitoring and Impact Assessment

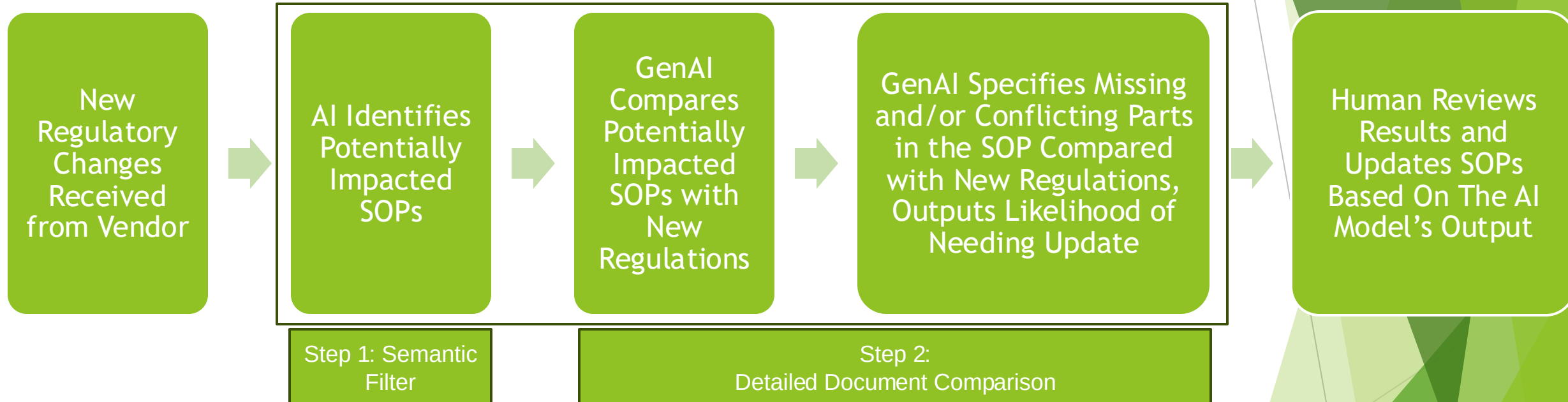
Iterate for many SOPs and other controlled documents



- Iterations are human-driven and may take days or weeks
- The quality varies due to the subjectivity of different individuals

Regulatory Intelligence with GenAI

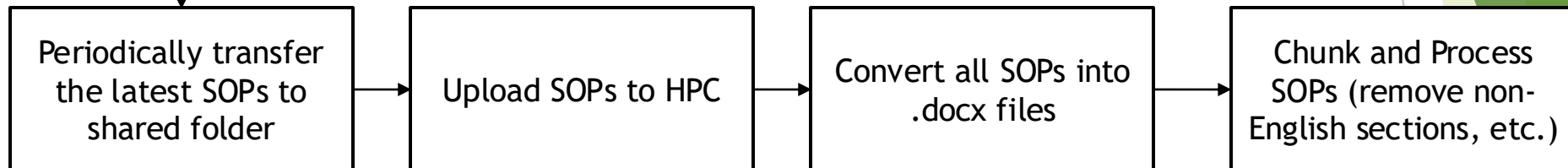
Iterate for many SOPs and other controlled documents



- The GenAI workflow completes the impact assessment automatically, and SME's role transferred from a task executor to a process supervisor
- More efficient and scalable, less subjective

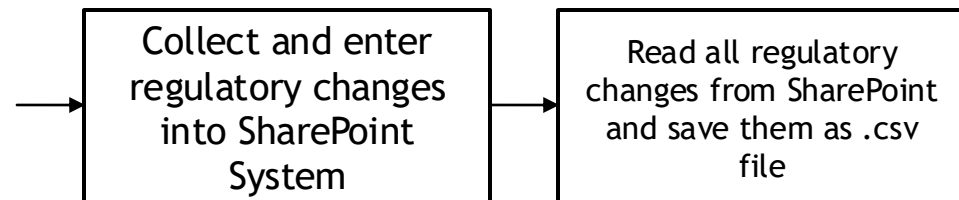
Input Data Pipeline Architecture

Document Management Team



{.json}

Vendor

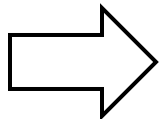


Step 1 Architecture: Identify Impacted SOPs

Convert both regulatory changes and SOP chunks into embedding vectors

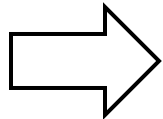
INPUT: Chunked SOPs + Summary of Regulation

`{.json}`



0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3
-0.8	-0.4	-0.5	0.1	-0.9	0.3	0.8

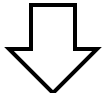
Generate similarity score at SOP chunk level



	reg_id	doc_id	doc_id_chunk_index	cosine_similarity
0	260	109712645	0	0.774672
1	260	109712645	1	0.778733
2	260	109712645	2	0.811734
3	260	109712645	3	0.812504
4	260	109712645	4	0.806150
...	...	...	...	...
216001	1	118306467	1	0.658431
216002	1	122665726	0	0.680289
216003	1	122665726	1	0.658431
216004	1	118691552	0	0.680289
216005	1	118691552	1	0.658431

Save out results as a JSON file, where each key is a RegID and the value is a list of impacted SOP numbers.

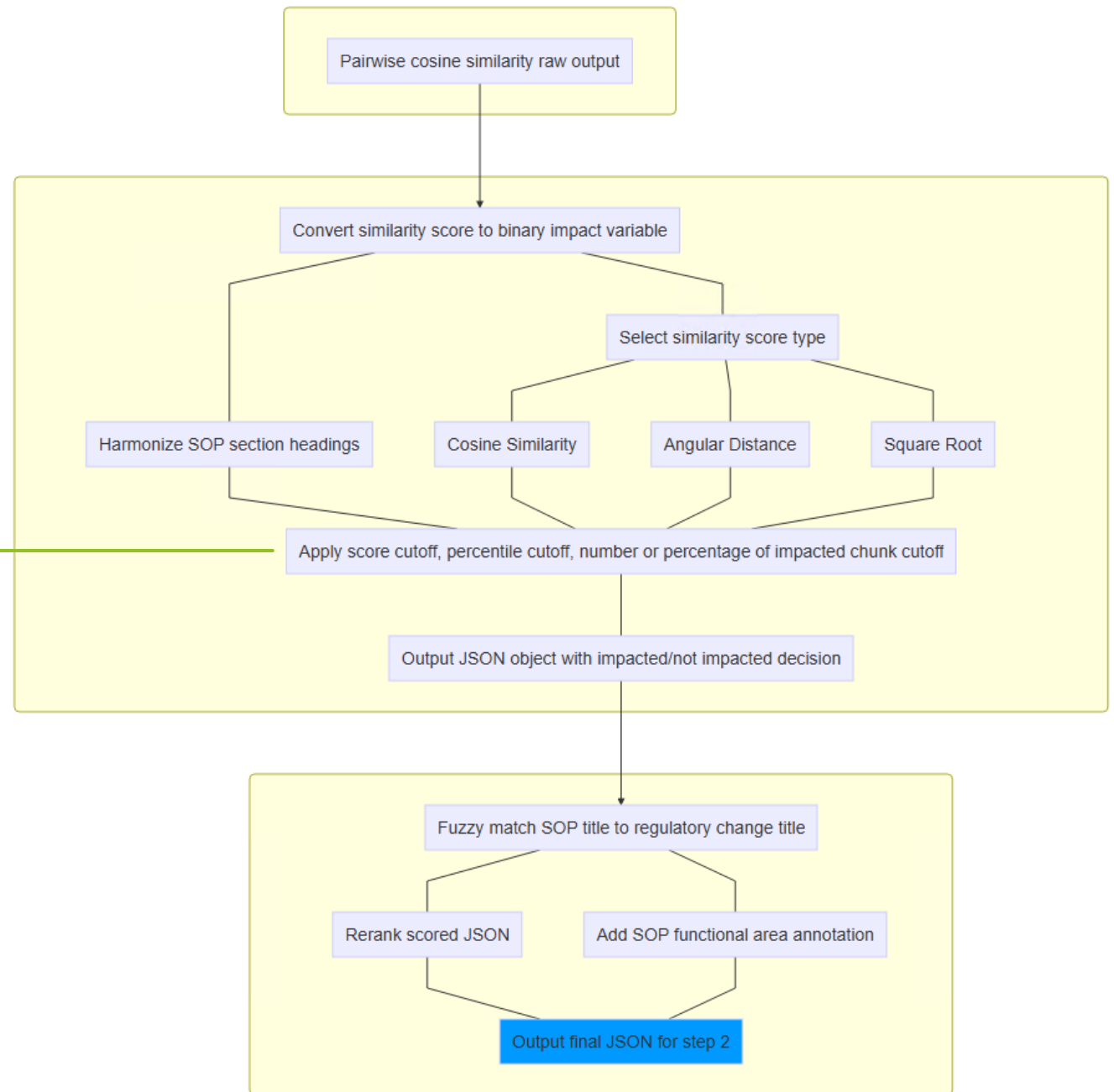
```
{ "Reg 119": ["DEV-SOP-1001", "DEV-SOP-1002", .....],  
  "Reg 120": ["DEV-SOP-1003", "DEV-SOP-1004", .....],  
  "Reg 121": ["DEV-SOP-1005", "DEV-SOP-1006", .....],  
  "Reg 122": ["DEV-SOP-1007", "DEV-SOP-1008", .....],  
  .....}
```



Aggregate similarity scoring results at SOP-Reg pair level, and apply proper thresholding strategy

Step 1 Architecture: Apply Similarity Threshold

Performed grid search based on subset of ground truth defined by SME to optimize these cutoff values

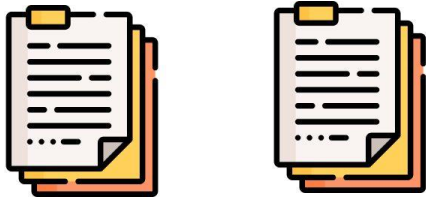


Step 2 Architecture: AI-generated Interpretation

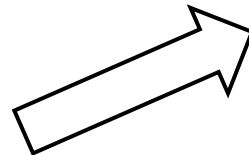
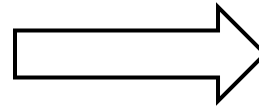
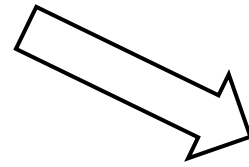
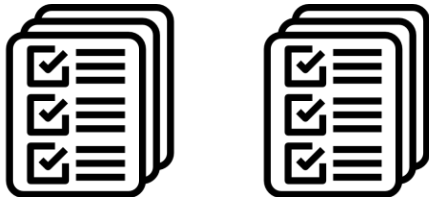
INPUT 1: JSON file from Step 1, where each key is a RegID and the value is a list of potentially impacted SOP numbers

```
{ "Reg 119": ["DEV-SOP-1001", "DEV-SOP-1002", .....],  
  "Reg 120": ["DEV-SOP-1003", "DEV-SOP-1004", .....],  
  "Reg 121": ["DEV-SOP-1005", "DEV-SOP-1006", .....],  
  "Reg 122": ["DEV-SOP-1007", "DEV-SOP-1008", .....],  
  .....}
```

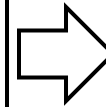
INPUT 2: SOP Documents



INPUT 3: Regulation Documents as .pdf file



Iterate through INPUT 1.
For each Reg ID, utilize LLM to compare regulation document with all potentially impacted SOPs. Generate a human-readable paragraph explaining any missing elements or conflicts in the SOP compared to the regulations and an impact determination (High/Medium/Low)



Save out results as a JSON file, where each key is a tuple (RegID, SOP_ID) and the value will be a list, first element is the determination, second value is the AI generated explanation

{.json}

Step 2 Model and Infrastructure

Methods Considered

- Sub-question Query Engine
- Custom Queries
- ReAct Agent
- Composable Graph

Method Chosen

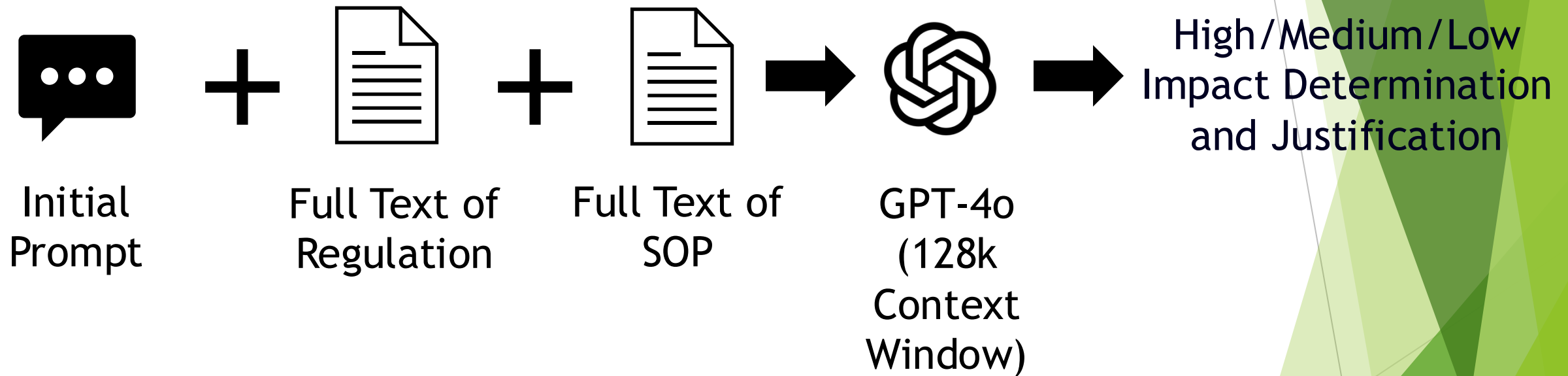
Long Context + Prompt Engineering

Infrastructure

- Azure OpenAI API
- GPT-4o
- HPC

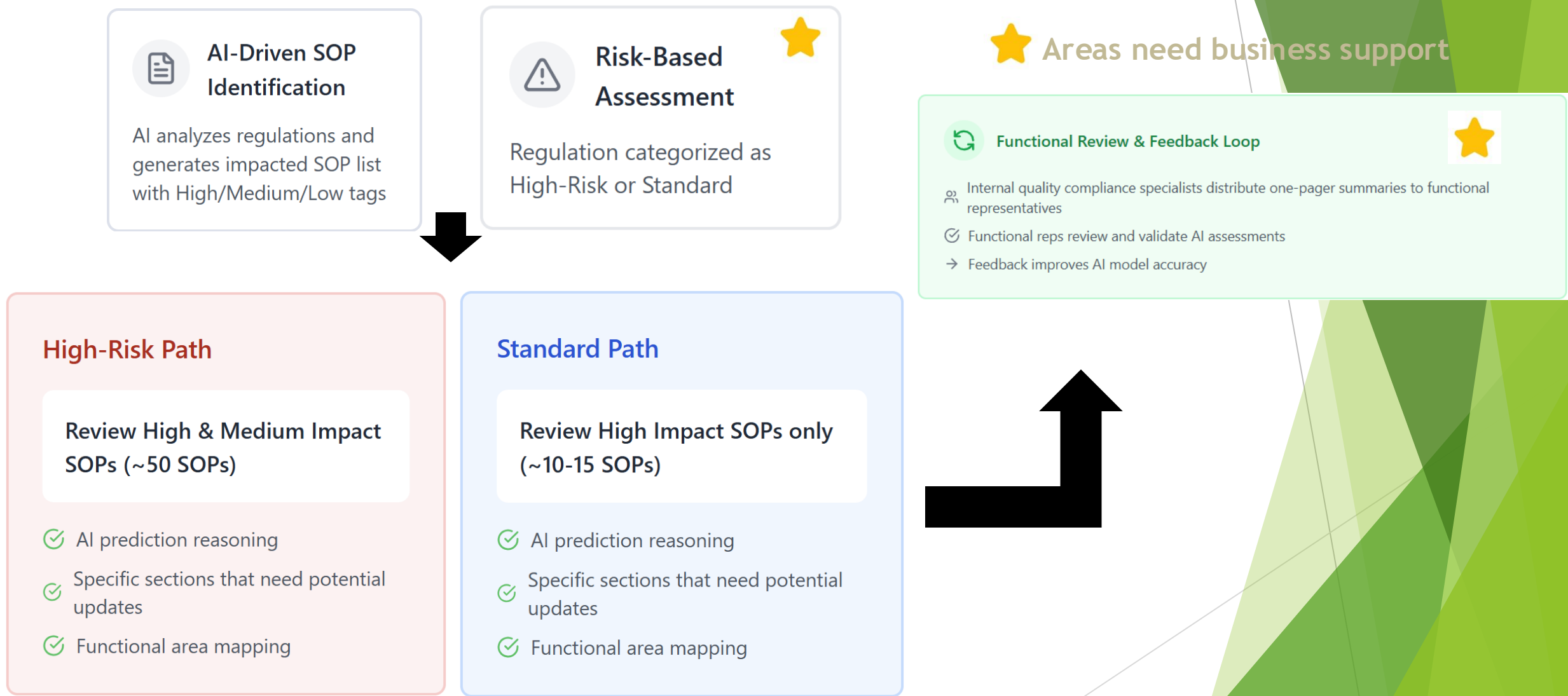
Approach	Accuracy	F1	Precision	Recall	Specificity
Custom Queries	0.6735	0.2000	0.3333	0.1429	0.8857
ReAct Agent	0.4792	0.5283	0.3590	1.000	0.2647
Long Context	0.8163	0.7097	0.6471	0.7857	0.8286

Long Context Approach



- Initial prompt is tuned to align output with SME feedback

Proposed Operational Workflow



Limitations



Small dataset of regulatory changes and impacted SOPs



Some overlap between tuning sets and validation set

Potential bias in our evaluations



Reproducibility assessment of our results

Probabilistic nature of LLMs -> variation in response
92% match when using impact/not impacted determination



Transparency

Black box nature of LLM
Ask model about its reasoning when comparing reg change to our SOPs



AI error rate

Possible that AI could rank SOP as low likelihood of being impacted by reg change, but in reality could be a high likelihood of impact

Summary and Next Steps



The current manual approach is not scalable and may result in impacted SOPs being overlooked



AI-driven approach can compare all 300+ SOPs with a given regulation in a matter of hours for a cost of around \$10



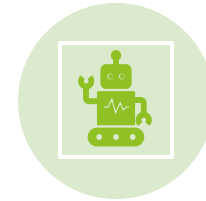
Promising methodology using long context LLM with the highest F1 score



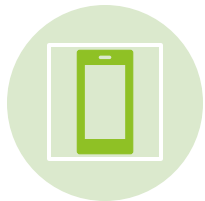
AI's impact determinations align well with SME's, but there is room for improvement with more validation data and additional tuning during a POC



New techniques like prompt caching may provide savings



Operationalize our framework at Biogen with human-in-the-loop to validate AI assessments



Internal streamlit app was developed to gather SME feedback on AI assessments

For more information, please contact:
Jake Gagnon
jake.gagnon@biogen.com