Paper ML05

Enhancing Pharmaceutical Quality Assurance through Large Language Models: A Novel Approach to Intelligent Regulatory Monitoring

Hangyu Liu, Biogen, Cambridge, MA, USA
Yuka Moroishi, Biogen, Cambridge, MA, USA
James Melican, Biogen, Cambridge, MA, USA
Matthew Ryals, Cencora, Conshohocken, PA, USA
Jake Gagnon, Biogen, Cambridge, MA, USA
Haleh Valian, Biogen, Cambridge, MA, USA

ABSTRACT

This study introduces an innovative Generative AI framework to enhance regulatory compliance processes in the pharmaceutical industry. With the continuous evolution of FDA and EMA regulations, pharmaceutical companies often face the challenge of managing manual, error-prone workflows for updating internal documents, such as Standard Operating Procedures (SOPs). We propose a two-step AI-driven solution to streamline regulatory update management. The first step leverages semantic search to identify SOPs potentially impacted by new regulatory updates. The second step utilizes a Large Language Model framework, powered by GPT-4o, to perform detailed comparisons between the new regulations and identified SOPs, confirming their impact and pinpointing specific sections requiring updates. This automated approach aims to reduce manual effort, minimize the risk of oversight, and ensure timely alignment with regulatory changes. Our findings demonstrate the transformative potential of Generative AI in regulatory intelligence, providing a scalable and efficient solution to a critical industry challenge. Future research will focus on refining the model and implementing it in real-world settings to quantify its benefits in terms of time savings and compliance accuracy.

INTRODUCTION

The pharmaceutical industry operates within a complex and ever-shifting regulatory landscape, governed by stringent guidelines from agencies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA). These regulatory frameworks dictate numerous aspects of drug development, manufacturing, distribution, and post-marketing surveillance. They are not static: as scientific knowledge evolves, public health priorities shift, and new technologies emerge, regulatory bodies respond by introducing novel guidelines or updating existing ones. As a result, maintaining organizational compliance is not simply a matter of establishing fixed procedures; it is an ongoing process of monitoring, interpreting, and integrating regulatory changes into internal documents, such as standard operating procedures (SOPs).

Currently, the process of regulatory monitoring is heavily reliant on manual efforts and the expertise of subject matter experts (SMEs). Teams tasked with regulatory monitoring must meticulously review updates from regulatory agencies, interpret their implications, and determine how they apply to internal processes and documentation. This approach often involves combing through lengthy regulatory texts, relying on individual judgment to assess the relevance and impact of changes. While SMEs bring valuable knowledge and experience to this task, the process is inherently subjective and varies based on personal interpretation and expertise. Furthermore, the manual nature of this work makes it time-consuming, prone to human error, and challenging to scale as the volume and complexity of regulatory updates grow. This labor-intensive system places significant pressure on regulatory monitoring teams and creates vulnerabilities in maintaining consistent and timely alignment with regulatory requirements. Figure 1 is an illustration of the current regulatory monitoring process:

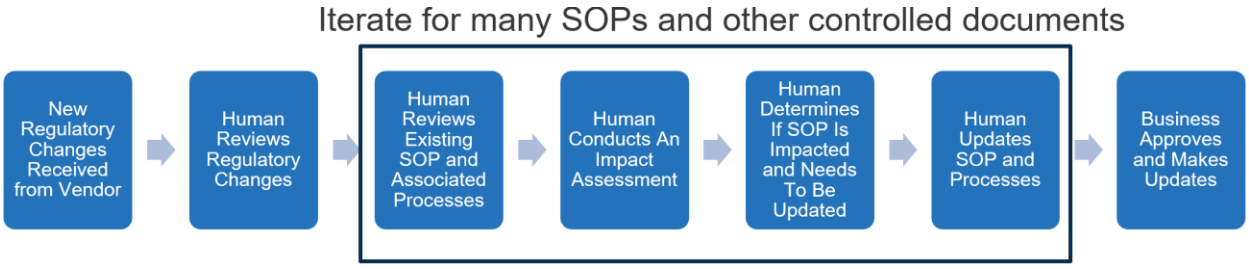

Figure 1: Human-driven Regulatory Monitoring and Impact Assessment

The challenges of the current manual approach are further compounded by the interconnected nature of pharmaceutical processes. A single regulatory update can ripple across multiple documents, workflows, and departments, requiring multiple teams to evaluate dependencies and ensure consistency throughout the organization. In practice, even minor oversights in this process can result in misalignment with regulatory requirements, exposing the organization to risks such as inspection findings, delays in product approvals, or reputational damage. As regulatory frameworks become more nuanced and the pharmaceutical industry increasingly embraces digital transformation, the limitations of manual monitoring and documentation alignment are becoming more apparent.

To address these challenges, we have developed an AI-powered regulatory monitoring system that can automate and streamline the process. By leveraging advancements in natural language processing (NLP) and large language models (LLMs), we built a system that can analyze and comprehend regulatory updates, identify a list of internal documents that may be impacted, and flag discrepancies or omissions for review. This would not only reduce the reliance on manual interpretation and effort but also enhance the accuracy and consistency of regulatory monitoring processes. Our ultimate goal is to empower regulatory monitoring teams to focus on strategic decision-making rather than routine document reviews, enabling faster adaptation to regulatory changes while minimizing the risk of non-compliance. Our AI-driven solution represents a significant step forward in how the pharmaceutical industry approaches regulatory monitoring, offering a scalable, reliable, and efficient alternative to the traditional manual methods.

Related Work in Regulatory Interpretation

Some related works to our regulatory monitoring pipeline include: 1) using LLMs for regulatory intelligence of health guidance documents; 2) financial regulatory interpretation; 3) LLMs applied to compliance checks in the food safety industry; and 4) finding relevant regulatory requirements for given business processes. We briefly summarize each of these related works below.

First, Venkatraman (Venkatraman Balasubramanian, 2024) discusses the application of LLMs to produce regulatory intelligence from health authority regulations in the January 2024 issue of DIA global forum. His goal was to increase productivity in regulatory submissions by making compliance more efficient by utilizing LLMs to summarize new guidelines as well as using LLMs (GPT 3.5 Turbo) to “chat” with the guideline documents. With their benchmark data, 77% of answers were categorized as highly accurate or close enough upon manual review, and the author suggests LLMs be used (to an extent) as a co-pilot not as a replacement for regulatory professionals.

In addition to using LLMs for regulatory intelligence on health guidance documents, Cao et al (Cao & Feinstein, 2024) propose the use of LLMs for financial regulatory documents. Specifically, they performed a case study in using LLMs to apply regulations from the ‘Minimum Capital Requirements for Market Risk’ section of the Basel III document to simulated bank assets holdings. Their framework included key components such as: document loading, prompt engineering, minimum capital requirements calculations, and pipeline assessment. Their assessment compared four LLMs and concluded that GPT-4 has the best overall performance. Additionally, they recommend detailed prompting, converting PDFs to images during document loading, and subdividing their objective into simpler objectives.

Thirdly, Shabnam (Hassani, 2024) discusses the application of LLMs to food safety regulations. Specifically, she was interested in two objectives: classifying each provision in food safety regulations and checking that a particular data processing agreement is compliant with GDPR[1] regulations. She proposes a hybrid approach combining LLM classification and keyword matching to achieve provision classification. As for compliance checking, she recommends content chunking and prompt construction as an input to a zero-shot LLM document comparison. Her experiments illustrated an 87% F-score on classification with BERT and an 81% accuracy for the compliance checking task with GPT-4.

Lastly, Sai et al (Sai, Sadiq, Han, & Rinderle-Ma, 2024) are interested in the automated comparison of business processes with regulatory texts to reduce the burden on domain experts. Their objective was finding the most

[1] The General Data Protection Regulation (GDPR) is a European law that protects the privacy and security of personal data. It applies to how personal data is collected, stored, and used by companies and organizations.

relevant regulatory texts for a given business process and after identifying relevant regulations, they wish to identify which step of the process is impacted. To this aim, the authors compared four approaches (the expert goal standard, embedding based models, GPT-4, and a crowd sourcing option) on two case studies at three levels of process detail. In their benchmark study consisting of travel insurance claim processes and due diligence business processes for new banking customers, their recommendations depended on the anticipated process impact and frequency of regulatory changes. From their benchmark data, they make the following conclusions: embedding methods should be used with high process impact and high regulatory dynamics, GPT-4 should be used for low process impact with high regulatory dynamics, and expert analysis is necessary for high process impact and low regulatory dynamics.

Related Works in Document Comparison

Some previous work in document comparison include: 1) Retrieval-augmented generation (RAG) or LLMs, 2) composable graphs, 3) prompt decomposition, 4) ReACT agents, and 5) semantic search. Below we provide a brief description of these approaches.

One example of using RAG or LLM for document comparison is given by Narendra et al (Narendra, Shetty, & Ratnaparkhi, 2024) who used LLMs for legal document comparison to potentially improve efficiency in contract analysis. They had two objectives: 1) classifying legal hypotheses as entailment, contradiction, or not mentioned in a corpus of NDAs (ContractNLI) and providing the user the related evidence and 2) contract analysis of internal JPMorgan documents. For the 1st task, comparison of multiple LLMs illustrated that GPT-4 had a 70% F1 for contradictions, a 91% F1 for entailment, and a 93% precision for evidence identification. For the 2nd task, a RAG pipeline was applied to extract relevant chunks from a contract related to each template concept. From the extracted context, an LLM with detailed prompting classified each template concept for entailment in the contract document (plus reverse comparison). Overall, GPT-4 had an 96% accuracy on this task.

The Composable Graph in LlamaIndex converts a large document into several subcomponents (LlamaIndex, 2023) (LlamaIndex, 2023). This approach utilizes a tree-based data structure, where nodes represent pieces of information of a document, allowing for efficient querying. Queries recursively traverse the graph, starting at the root index then to sub-indices, to find relevant nodes for retrieval. The graph structure of this approach allows for querying large documents and side-by-side comparisons of related documents.

Prompt decomposition is a technique that breaks down a complex query into simpler components (Zhou, et al., 2022) (Dua, Gupta, Singh, & Gardner, 2022). The use of smaller and more focused queries simplifies the question, reduces ambiguity, and improves accuracy. It also allows the user to identify where the LLM makes an error in order to make adjustments to the prompt where necessary. This approach is especially useful when retrieving information from multiple sources as results from the multiple queries can be synthesized. The sub question query engine in LlamaIndex is one example of prompt decomposition (LlamaIndex, n.d.). It can be paired with a retriever to efficiently retrieve information from a document or multiple documents. Briefly, the question of interest is split into multiple sub-questions that are processed through a query engine. These sub-questions can be answered in parallel or sequentially, if the result of one sub-question depends on the result of another. All answers to the sub-questions are then synthesized to answer the original question.

The idea of a ReAct agent was first proposed in "ReAct: Synergizing Reasoning and Acting in Language Models" (Yao, et al., ReAct: Synergizing Reasoning and Acting in Language Models, 2023) (Yao, et al., n.d.) in a collaboration between Google Research and Princeton. The novel idea of ReAct is for a LLM to combine both reasoning and acting to improve its performance and to generate more interpretable responses compared to a standard LLM. Acting here refers to the LLM using tools such as web search, knowledge base search, or querying a database. The LLM's action plan can be updated dynamically by the model as the model reasons about the user's query. In contrast to standard LLMs, ReAct agents can utilize external information to gather up-to-date information to answer a user's query which reduces the hallucination rate of the LLM. Although the original paper did not apply ReAct agents to the document comparison task, we will describe below how we applied a ReAct agent to the comparison of incoming regulations to SOP documents.

Semantic search, which involves finding documents that are semantically similar to a user's query, has been a popular topic in information retrieval in recent years. Some previous works include embedding methods based on deep learning (refs below), WordNet metrics (O, 1999), Semantic Web relatedness metrics (Gracia & Mena, n.d.), topological similarity metrics (O, 1999) (Pekar & Staab, 2002), latent semantic indexing (LSI) (Deerwester, 1988), point wise mutual information (PMI) (Fano, 1961), and semantic networks (Sowa, 2015). In this manuscript, we focus on deep learning embedding approaches to find semantically similar documents by converting both the user's query and documents into a latent vector representation. Some examples in the literature include: the sentence-t5-xxl model (Ni, et al., 2021), openai's ada002 (openai, 2022), E5 (Wang, et al., 2024), Cohere's embed v3 (Reimers, et al., 2023), word2vec (Mikolov, Chen, Corrado, & Dean, 2013), GloVe (Pennington, Socher, & Manning, n.d.), FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017), ELMo (Peters, Neumann, Iyyer, & Gardner, 2018), BERT (Devlin, Chang, Lee, & Toutanova, 2019), and others. After the embedding step, the semantic "relevance" between the user's query and documents are assessed with metrics such as cosine similarity.

Comparison to related works

Our proposed approach builds upon previous work by incorporating ideas from semantic search as an initial filter and then performing a detailed document comparison. Previous literature methods in document comparison are compared below using methods such as composable graphs, a ReAct agent, a sub-question query engine, custom querying, and long context LLMs. With our benchmark dataset, we illustrate superior performance with long context LLMs.

Compared to past work in regulatory interpretation, our objective differs in a few key aspects. Compared to Venkatraman's work, we wish to identify impacted SOPs based on new regulatory guidance rather than summarizing or "chatting" with regulatory documents. As for Cao et al and Shabnam's work, our domain of applicability is different. We are interested in pharmaceutical regulatory intelligence rather than the financial or food safety domains. Lastly, Sai et al's aim is opposite of ours: they wish to identify relevant regulations for a given business process, whereas we are looking for impacted business processes and SOPs based on a given set of regulations.

METHODOLOGY

Our AI-powered regulatory monitoring and impact assessment is a 2-step approach as shown in Figure 2. The proposed framework comprises an iterative two-stage process for managing regulatory compliance updates across pharmaceutical documentation. When new regulatory changes are received, Step 1 employs a semantic filtering mechanism to identify potentially impacted SOPs and controlled documents from the organization's document repository. Step 2 then conducts a granular document comparison through our LLM framework, which performs three key functions: comprehensive comparison between identified SOPs and new regulations, specification of discrepancies and conflicts, and generation of a confidence score indicating the likelihood of impact. This multi-tiered approach ensures systematic coverage while minimizing false positives and computational cost through semantic pre-filtering. The framework culminates in human expert review, where SMEs validate the AI-generated insights and implement necessary documentation updates. This human-in-the-loop design maintains critical oversight while leveraging AI capabilities to dramatically reduce manual comparison workload. Our later empirical evaluation demonstrates that this approach achieves promising accuracy in identifying truly impacted documents while reducing review time by approximately from weeks to hours compared to traditional manual methods.

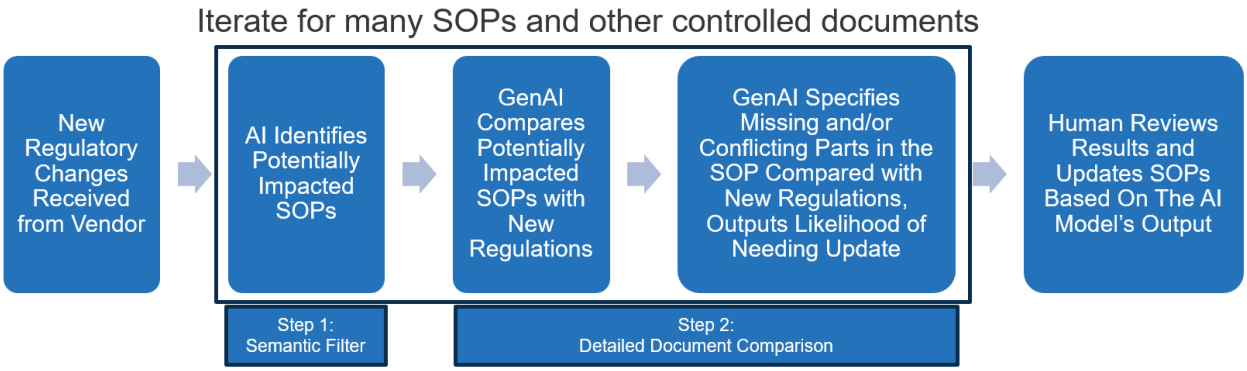

Figure 2: AI-driven Regulatory Monitoring and Impact Assessment

Step 1: Semantic Filtering

A set of regulatory change document summaries and a set of DEV-SOP[2] documents comprise the analyzed dataset. Summaries of each regulatory change are assumed *a priori* to effectively describe the regulatory change they represent. The flowchart in Figure 3 illustrates the workflow for Step 1.

[2] DEV-SOP: Development Standard Operating Procedures (DEV-SOP) are a comprehensive set of documented processes and guidelines established under the Research & Development (R&D) division. These procedures ensure consistency, compliance, and quality across all R&D activities, aligning with regulatory standards and organizational objectives.

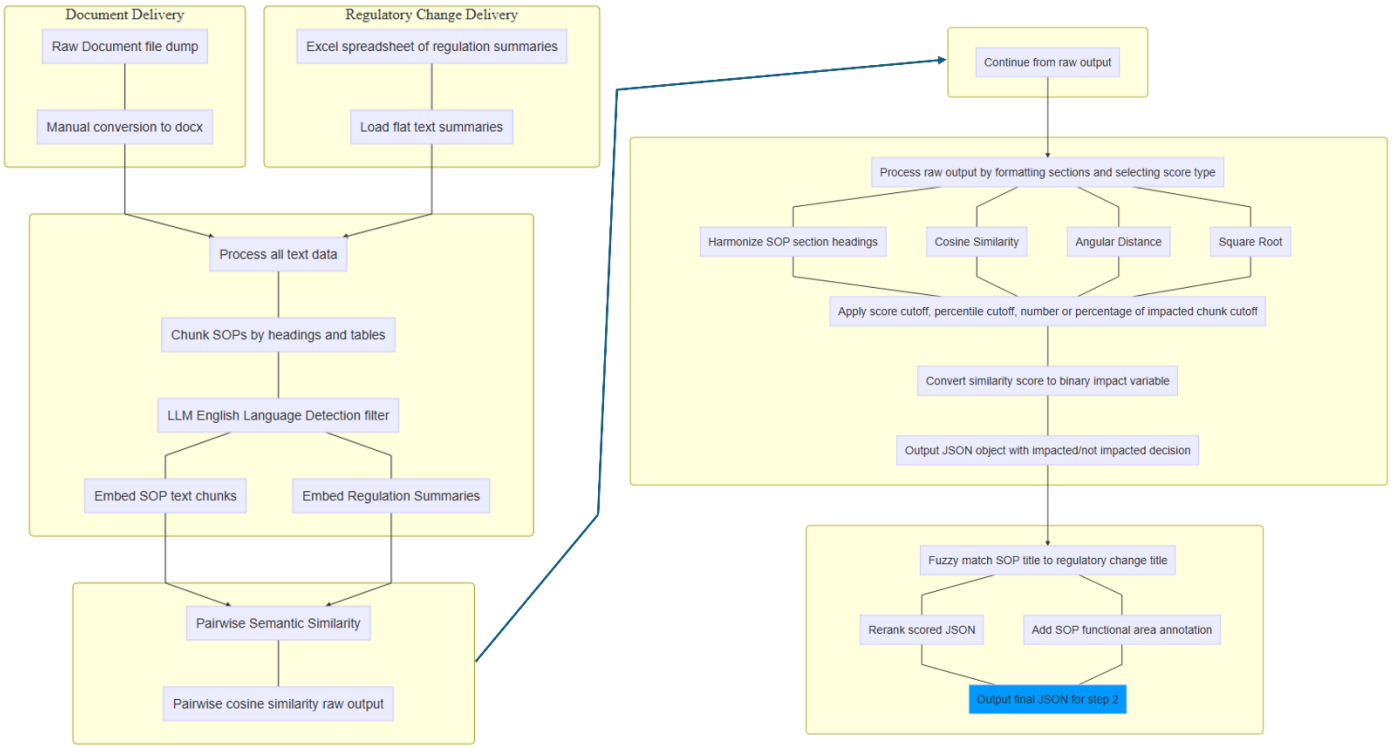

Figure 3: Semantic Filtering (Step 1) flowchart

Each DEV-SOP was either natively in Microsoft Word docx format with formatted headings or coerced to docx using the python *pdf2docx* and *spire.doc* packages. They were parsed using a hierarchical approach with the *docx* python package where each level 1 heading was used to divide text and tables (table contents are parsed row-wise and added to the text) into sections, then each section was sub-chunked to a size of 512 tokens. Non-English chunks were then detected and removed using language detection LLM, juliensimon_xlm-v-base-language-id. Summaries of each regulatory change and all DEV-SOP text chunks were embedded using sentence-t5-xxl and compared to assign pairwise similarity scores. In the following sections, we will use regulation 119 and 146[3] as examples.

Similarity scores were converted to a binary classifier using a rule-based approach where DEV-SOP section headings are harmonized to a set of unified headings, and a DEV-SOP score must be greater than the percentile threshold of a regulatory change's total score distribution in addition to passing a flat similarity score cutoff. A DEV-SOP must have a requisite number of chunks passing those criteria within the harmonized sections in order to be considered impacted. The score used could be either cosine similarity, angular distance, or the square root. Angular distance is given by $((f(x)=1-\arccos(x)/\pi))$ and square root is given as $(f(x)=\sqrt{1-(1-x)/2})$, where x is the cosine similarity score values. Specific parameters, including score type, are optimized with a grid search based on a subset of ground truth for regulation 119 (95^{th} percentile cutoff, 0.77 flat cutoff for angular distance, and 8 chunks required in harmonized headings).

Taxonomy Tree

A taxonomy tree-based approach was explored as an alternative approach to semantic filtering. First, an XML tree of topics and subtopics (2-layer tree) relevant to medical study administration is built using OpenAI model GPT-4o using prompting. Using either GPT-4o or GPT-4o-mini, the XML tree was then passed to a second prompt along with the text of a regulatory change summary to assign a topic or subtopic to that regulatory change. Following that, each DEV-SOP text was concatenated with section headings, and data extracted from tables into a single text chunk and passed to a third prompt along with the XML tree to extract relevant topics and subtopics for that DEV-SOP. Topic extraction for each DEV-SOP was repeated 5 times and a set of topics that appeared at 80% frequency or greater was then used to determine the topic set. Relevancy was determined if any topic of the DEV-SOP topic set matches the regulatory change topic. Topic matching could occur either in a child subtopic or in the parent topic.

[3] Regulation 119 is an EMA guideline on computerized systems and electronic data in clinical trials; Regulation 146 is a FDA guidance document on considerations for the conduct of clinical trials of medical products during major disruptions due to disasters and public health emergencies

Step 2: Document Comparison

We explored three approaches to comparing documents: custom queries, a ReAct agent, and a long context window LLM. Our preliminary assessment of Composable Graphs revealed long runtimes and poor performance. As a result, we did not include this approach in the final experiment. Additionally, our preliminary assessment of the sub question query engine revealed low performance. We instead generated our own set of sub-questions (i.e. custom queries) for the model and synthesized the results. The long context window approach involved combining the full text of both the regulatory document and the SOP into one prompt, and this limited our choice of LLMs to those with a context window of 128k tokens (or larger).

Custom Queries

The sub question query engine produced poor results in our preliminary analysis due to the model's inability to generate appropriate sub-questions from the original query. As a result, we subdivided our main document comparison query to mimic this approach. The same chunking procedure was applied to the DEV-SOP documents as described in Step 1. The pdf file of the regulatory document was parsed using 'SimpleDirectoryReader' from *LlamaIndex*. We deployed GPT-4o as the model for this approach and 'bge-base-en-v1.5' embeddings to index documents. First, we queried the model to list topics and subtopics outlined in the table of contents of the regulatory document. This query was run five times, and topics that were present in all five runs were saved. Then we queried the model to identify parts of the DEV-SOP document that are relevant to the regulatory document topics identified. If at least one of the topics were relevant, the document was labeled as impacted. Otherwise, the document was labeled as not impacted. We prompt engineered this approach on 36 document comparisons, consisting of regulation 119 and 36 DEV-SOP documents.

ReAct Agent

In our study, we implemented a ReAct-based LLM agent using GPT-4o for the agent and OpenAI embeddings within the RAG query tools. We utilized the LlamaIndex implementation to build the agent and its tools, facilitating seamless integration of the reasoning and retrieval processes (LlamaIndex; ReAct Agent, 2024). The agent was tasked with identifying the main topics in both the SOP and the regulatory document, reasoning about potential conflicts, and performing targeted queries to compare specific sections. One hyperparameter in this setup is the maximum number of iterations or tool uses. Initially set to 20, we increased this limit to 30 after observing that the agent occasionally required additional steps to complete its task.

Despite the promise of this approach, we encountered challenges with the RAG retrieval mechanism, which was configured to return the top 5 or 10 document chunks per query. This retrieval often lacked precision, as the most relevant sections were not always included, leading to less accurate comparisons. Additionally, the agent occasionally produced outputs that were inconsistent with its intermediate reasoning steps, indicating limitations in integrating reasoning and action phases.

An additional difficulty we faced was engineering the initial prompt to achieve consistent outputs, given the 20+ intermediate reasoning and querying steps performed by the agent. This highlights the sensitivity of ReAct-style agents to prompt quality when tasked with multi-step, complex reasoning workflows. Our findings align with prior research showing that the effectiveness of ReAct prompting can vary depending on the coherence of reasoning traces and the precision of retrieval mechanisms. (Verma, Bhambri, & Kambhampati, 2024) These challenges underscore the need for further optimization of both retrieval strategies and prompt design to improve consistency and performance of ReAct agents in long-context reasoning tasks.

Long Context Approach

To address the limitations observed with the ReAct and RAG-based methods—such as imprecision in retrieval, inconsistencies between reasoning steps and outputs, and difficulties in prompt engineering—we adopted a long context approach. This method involved combining the full text and tables of both the regulatory document and the SOP into a single input, preceded by an initial prompt (Supplementary Material 1) that explained the task and provided instructions on how to compare the documents. This approach required the use of LLMs with extended context windows, specifically GPT-4o, which supports inputs up to 128k tokens.

The key advantage of the long context approach is its ability to process both documents simultaneously within a unified context window. This avoids the fragmentation inherent in retrieval-based methods, where only subsets of the documents (e.g., top 5 or 10 chunks) are queried at any one time. In this way, the long context approach enables the model to identify relationships and conflicts at both a global and a token-level scale, without risking the omission of critical information.

This improvement can be understood through an analogy to cross-encoders and bi-encoders (ntongana, n.d.) in NLP. While bi-encoders process two inputs independently and compare their embeddings, cross-encoders process both

inputs together, allowing the model to evaluate interactions between tokens directly. Similarly, by placing the full texts of both documents into the same prompt, the long context approach allows the LLM to directly compare and analyze the content across the two texts without losing detail. We found that the total token usage and time required to generate a comparison was similar to the ReAct agent, but the quality of the summaries and classifications generated were improved, in addition to enabling more intuitive prompt engineering.

EXPERIMENTS

Comparison of Semantic Filtering versus Taxonomy Tree approach in Step 1

Using the subset of the 49 DEV-SOPs that were validated as ground truth by SMEs as impacted/not impacted by regulation 146, a comparison was performed to see whether a marked difference was observed in the error matrix performance between the methods.

A)

Metric	GPT-4o	GPT-4o-mini	Semantic Filtering
TP	9	12	12
TN	11	0	4
FP	24	35	31
FN	5	2	2

B)

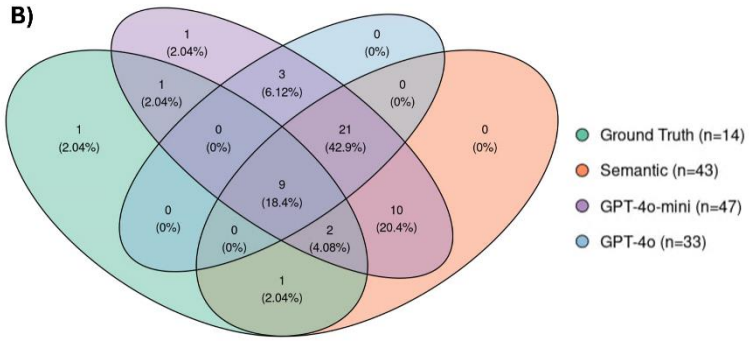

Figure 4: A) Error Matrix result[4] for Step 1 method comparison[5]. B) Venn diagram overlap of DEV-SOPs identified as impacted between ground truth and Step 1 methods.

The error matrix in Figure 4A did not show a large difference in the TP or FN rate between methods. The topic tree approach when topics were identified by GPT-4o had an improved TN and FP rate compared to the tree approach using GPT-4o-mini, and the semantic filtering approach. The result of the TP DEV-SOP overlaps is shown in Figure 4B, demonstrating that only 1 of the ground truth impacted DEV-SOPs remained unidentified across the methods for this set.

Assessment of Document Comparison Approaches in Step 2

We assessed the performance of the three document comparison approaches on regulation 146 and a subset of DEV-SOP documents. The subset of DEV-SOP documents consisted of both documents selected by Step 1 and those that were not selected by Step 1. A SME validated the comparisons with 49 DEV-SOP documents, and the documents were labeled dichotomously as “impacted” or “not impacted". Of the 49 DEV-SOP documents in this testing set, 28.6% (n=14) were labelled as impacted and 71.4% (n=35) were labelled as not impacted. The Custom Queries and ReAct Agent approaches generated dichotomous conclusions of “impacted” and “not likely impacted”. On the other hand, the Long Context approach generated conclusions as High, Medium, or Low impact. We grouped the High and Medium impacts and tagged them as “impacted” and tagged Low impacts as “not impacted”.

Overall, the Long Context approach had the highest performance compared to the Custom Queries and ReAct Agent approaches, with accuracy of 0.8163 and F1 of 0.7097 (Table 1). While the recall and specificity are higher in ReAct Agent and Custom Queries respectively, the Long Context approach demonstrated much higher precision. We note that prompt engineering for the Long Context approach was conducted on a subset of the 49 documents we tested.

[4] **TP (True Positive):** Cases correctly identified as positive by the model. **FP (False Positive):** Cases incorrectly identified as positive by the model. **FN (False Negative):** Cases incorrectly identified as negative by the model. **TN (True Negative):** Cases correctly identified as negative by the model.

[5] The second and third columns in Figure 4 Table (A) are results from the taxonomy tree approach

Table 1: Performance Metrics of Various Document Comparison Approaches on 49 Document Comparisons of Regulation 146

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

CONCLUSION

Our study highlights the significant potential of Generative AI frameworks to transform regulatory compliance processes in the pharmaceutical industry. By introducing a two-step approach that combines semantic filtering with advanced document comparison using LLMs, we have demonstrated how the identification and assessment of regulatory updates on internal documents, such as SOPs, can be automated effectively. Empirical evaluations validate the promise of this approach, with the Long Context method emerging as the most robust among the tested strategies, achieving the highest overall F1 score while maintaining strong recall and specificity. Notably, the study also emphasizes the indispensable role of human oversight in validating AI-driven insights, ensuring that the framework remains both scalable and reliable in the context of complex, high-stakes regulatory workflows. These findings illustrate the practical utility of integrating AI solutions to address long-standing inefficiencies and risks associated with traditional compliance methods. In the next phase, the focus will shift to real-world implementation and refinement, aiming to streamline the regulatory monitoring process further and resolve the critical challenges that manual approaches have historically posed.

To build on the insights from this study, one of our next steps is to operationalize this framework. The proposed operational workflow integrates a sophisticated risk-stratified approach with continuous learning capabilities (Figure 5). The process initiates with AI-driven SOP identification, where our model analyzes regulatory changes and automatically assigns High/Medium/Low likelihood of impact tags to potentially affected SOPs. This classification feeds into a risk-based assessment framework. SMEs categorize regulations as either High-Risk or Standard based on the importance and urgency of regulations, which determines subsequent review pathways. The High-Risk path encompasses review of both high and medium impact SOPs (approximately 50 documents), while the Standard path focuses exclusively on high-impact SOPs (approximately 10-15 documents). Both paths benefit from AI-generated insights, and functional area mapping. A critical step in our workflow is the Functional Review & Feedback Loop, where internal quality compliance specialists distribute concise one-page summaries of AI determination and insights to functional representatives. This human-AI collaborative approach not only validates AI assessments but also generates valuable feedback data that continuously enhances model accuracy, expecting to achieve significant efficiency gains and accuracy lift.

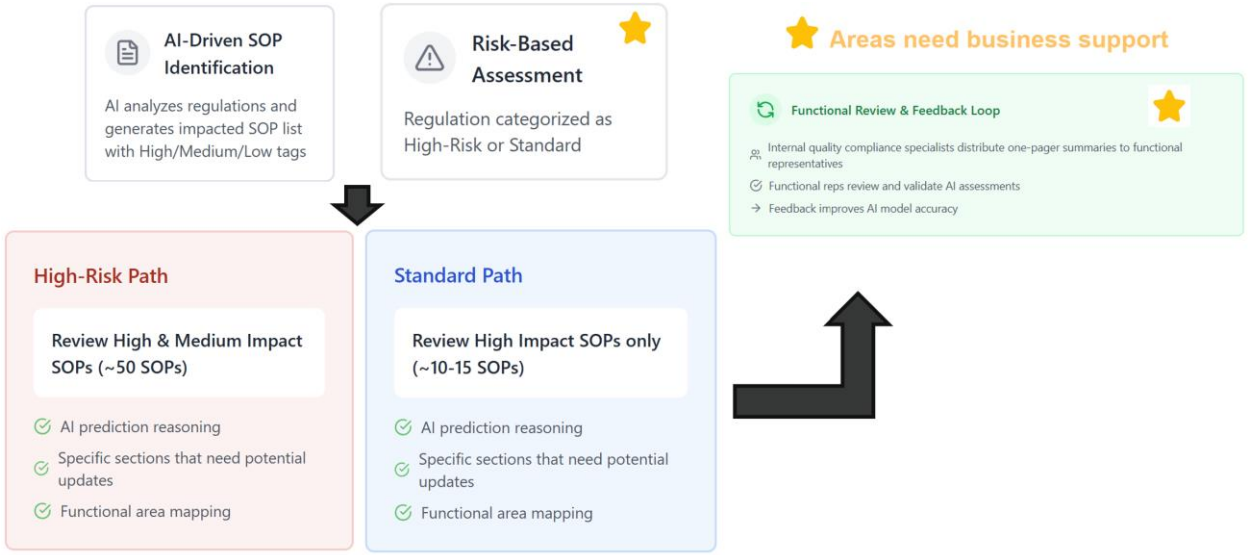

Figure 5: Proposed human-AI collaborative Operational Process

A web application was developed using the Streamlit package in Python to facilitate the collection of feedback from SMEs. The app allows the user to select a regulation, navigate through potentially impacted SOPs, and indicate whether they agreed with the AI-generated impact assessment and its explanation. An example is attached in the supplementary material. We are working with business stakeholders to embed this application into our operational workflow.

Despite promising results, several limitations exist. First, since regulatory monitoring process is relatively new at Biogen, we do not have a large number of historical data points. Our study was constrained by a relatively small dataset of regulatory changes and their corresponding impacted SOPs. The ground truth annotations were accumulated gradually during model development, preventing us from implementing an ideal train/validation/test split typically expected in machine learning research, and these details have been discussed in experiments section. This limitation potentially introduces bias in our performance metrics and may affect the generalizability of our findings to broader regulatory contexts. Second, while our reproducibility assessment demonstrated robust overall performance, we observed inherent variability in the High/Medium/Low likelihood of impacted generated by the LLM component. Running the approach three times on the testing set yielded a 91.8% match in all three runs when assessing the variability of impacted versus not impacted determinations and an 81.6% match when assessing the variability of High/Medium/Low determinations. This variability highlights the probabilistic nature of LLM-based decision-making systems. Additionally, the use of LLMs in Step 2 of our framework presents challenges in terms of complete model transparency. Although we implemented explicit reasoning output to enhance result interpretability, the underlying decision-making process of the LLM remains partially opaque. This "black box" aspect could pose challenges in regulated environments where full algorithmic transparency might be required. Lastly, the AI system is not infallible. There is a risk that the framework could classify a regulatory update as having "Low" impact when, in reality, it is of "High" importance. Such misclassifications could lead to oversights in compliance, posing significant risks in regulated environments. Addressing this requires continuous optimization, iterative testing, and integration of human expertise to mitigate potential consequences of these errors. Future work should focus on increasing the sizes of training and testing datasets, extending application to other document types (e.g. Job Aids, etc.), implementing more rigorous validation protocols, and exploring approaches to enhance the deterministic nature of impact classifications while maintaining the advantages of LLM-based analysis.

REFERENCES

1. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching Word Vectors with Subword Information*. Retrieved from arxiv: https://arxiv.org/abs/1607.04606
2. Cao, Z., & Feinstein, Z. (2024). *Large Language Model in Financial Regulatory*. Retrieved from arxiv: https://arxiv.org/pdf/2405.06808
3. Deerwester, S. e. (1988). Improving Information Retrieval with Latent Semantic Indexing. *Proceedings of the 51st Annual Meeting of the American Society for Information Science 25*, (pp. 36–40).
4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Retrieved from arxiv: https://arxiv.org/abs/1810.04805
5. Dua, D., Gupta, S., Singh, S., & Gardner, M. (2022). Successive Prompting for Decomposing Complex Questions. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1251-1265. doi:10.18653/v1/2022.emnlp-main.81
6. Fano, R. (1961). Transmission of Information: A Statisticial Theory of Communications. MIT Press.
7. Gracia, J., & Mena, E. (n.d.). *Web-Based Measure of Semantic Relatedness*. Retrieved from Universita di trento: https://disi.unitn.it/~p2p/RelatedWork/Matching/Gracia_wise08.pdf
8. Hassani, S. (2024). *Enhancing Legal Compliance and Regulation Analysis with Large Language Models*. Retrieved from arxiv: https://arxiv.org/abs/2404.17522
9. LlamaIndex. (2023). *Composability*. Retrieved from LlamaIndex 0.9.48: https://docs.llamaindex.ai/en/v0.9.48/module_guides/indexing/composability.html
10. LlamaIndex. (2023). *Composable Graph*. Retrieved from LlamaIndex 0.9.48: https://docs.llamaindex.ai/en/v0.9.48/examples/composable_indices/ComposableIndices.html
11. LlamaIndex. (n.d.). *Sub Question Query Engine*. Retrieved from LlamaIndex: https://docs.llamaindex.ai/en/stable/examples/query_engine/sub_question_query_engine/
12. *LlamaIndex; ReAct Agent*. (2024). Retrieved from https://docs.llamaindex.ai/en/stable/examples/agent/react_agent/
13. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in*. Retrieved from arxiv: https://arxiv.org/pdf/1301.3781
14. Narendra, S., Shetty, K., & Ratnaparkhi, A. (2024). Enhancing Contract Negotiations with LLM-Based Legal Document. *Proceedings of the Natural Legal Language Processing Workshop 2024* (pp. 143-153). Association for Computational Linguistics.
15. Ni, J., Ábrego, G., Constant, N., Ma, J., Hall, K., Cer, D., & Yang, Y. (2021). *arxiv*. Retrieved from

Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models: https://arxiv.org/abs/2108.08877
16. ntongana, J. (n.d.). *Decoding Sentence Representations: A Comprehensive Guide to Cross-Encoders and Bi-Encoders*. Retrieved from https://plainenglish.io/community/decoding-sentence-representations-a-comprehensive-guide-to-cross-encoders-and-bi-encoders-3a675a
17. O, R. (1999). Semantic Similarity in a Taxonomy: An information based measure and its application to problems of ambiguity and natural language. *Journal of Artificial Intelligence Research Vol 11*, 95-130.
18. openai. (2022). *New and improved embedding model*. Retrieved from openai: https://openai.com/index/new-and-improved-embedding-model/
19. Pekar, V., & Staab, S. (2002). Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. *Proceedings of the 19th international conference on Computational linguistics*, (pp. 1-7).
20. Pennington, J., Socher, R., & Manning, C. (n.d.). *GloVe: Global Vectors for Word Representation*. Retrieved from nlp.stanford.edu: https://nlp.stanford.edu/projects/glove/
21. Peters, M., Neumann, M., Iyyer, M., & Gardner, M. (2018). *Deep contextualized word representations*. Retrieved from arxiv: https://arxiv.org/pdf/1802.05365
22. Philip, R. (1995). *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*. Retrieved from arxiv: https://arxiv.org/pdf/cmp-lg/9511007
23. Reimers, N., Choi, E., Kayid, A., Nandula, A., Govindassamy, M., & Elkady, A. (2023). *Introducing Embed v3*. Retrieved from cohere.com: https://cohere.com/blog/introducing-embed-v3
24. Sai, C., Sadiq, S., Han, L. D., & Rinderle-Ma, S. (2024). *Identification of Regulatory Requirements Relevant to Business Processes: A Comparative Study on Generative AI, Embedding-based Ranking, Crowd and Expert-driven Methods*. Retrieved from arxiv: https://arxiv.org/abs/2401.02986
25. Sowa, J. (2015). *Semantic Networks*. Retrieved from jfsowa: https://www.jfsowa.com/pubs/semnet.htm
26. Venkatraman Balasubramanian. (2024). *Large Language Models: Extracting and Summarizing Regulatory Intelligence from Health Authority Guidance Documents*. Retrieved from DIA global forum: https://globalforum.diaglobal.org/issue/january-2024/large-language-models-extracting-and-summarizing-regulatory-intelligence-from-health-authority-guidance-documents/?utm_source=chatgpt.com
27. Verma, M., Bhambri, S., & Kambhampati, S. (2024). On the Brittle Foundations of ReAct Prompting for Agentic Large Language Models.
28. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., . . . Wei, R. (2024). *Text Embeddings by Weakly-Supervised*. Retrieved from arxiv: https://arxiv.org/pdf/2212.03533
29. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. *International Conference on Learning Representations (ICLR)*. Retrieved from https://arxiv.org/abs/2210.03629
30. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (n.d.). *ReAct: Synergizing Reasoning and Acting in Language Models*. Retrieved from react-lm.github.io: https://react-lm.github.io/
31. Zhou, D., Scharli, N., Hou, L., Wei, J., Scales, N., Wang, X., . . . Chi, E. (2022, May). Least-to-most prompting enables complex reasoning in large language models. *ICLR*. doi:10.48550/arXiv.2205.10625

ACKNOWLEDGMENTS

We would like to thank our Biogen colleagues Kerensa Ransom, Diane Scaffidi, Jennifer Giannetti, Abner Vazquez, Jay Timmerman, Sindhu Gurajala, Suhas Mahadev Pawar, MD Mainul Islam, Archana Gundamraju, and our Pharmalex colleague, Jialie Luo, for their collaborative work to help validate the results and provide insight on the application development.

CONTACT INFORMATION

Contact the corresponding author at:
Author Name: Hangyu (Cedric) Liu
Company: Biogen, Inc.
Address: 701 Pennsylvania Avenue, NW Suite 715. Washington, DC 20004
Email: Hangyu.liu@biogen.com

SUPPLEMENTARY MATERIAL

Supplementary Material 1: Prompt for the LLM in the long context approach of Step 2

"""**INSTRUCTIONS:** The text of two documents follows; first is a newly published regulation, and after that is an internal SOP from Biogen. Your job is to make a determination of HIGH, MEDIUM, or LOW for the likelihood that the SOP needs to be updated in order to be in compliance with the regulation. In addition to providing your determination, identify the specific sections in the regulation and the SOP that appear to be in conflict with each other. The rubric for making your determination is as follows: If the topics covered in the SOP directly overlap with those in the regulation and there are sections that are potentially in conflict, the likelihood of needing an update should be HIGH. If the SOP and regulation topics do not directly overlap but are conceptually related and potentially in conflict, it should be MEDIUM. If the topics covered by the SOP are not related to the regulation, the likelihood of needing an update should be LOW. The title of the SOP should be an important factor in determining if it is likely to be relevant to the regulation. """

Supplementary Material 2: Example of Streamlit Application