Benchmarking Large Language Models for GxP Healthcare Compliance

Data Science Institute
COLUMBIA UNIVERSITY

Vishweshwar Tyagi, Daoxing Zhang, Siqi He, Siwen Xie, Yihao Gao

Industry Mentors: Frank Janssens & Majd Mustapha
Faculty Mentors: Adam S. Kelleher & Cathy Li

Data Science Capstone Project
with Johnson & Johnson

Background & Business Need

GxP regulations are thousands of pages of text files (pdf or HTML) posted in several internet locations. These regulatory requirements must be manually parsed, analyzed, and classified to develop the J&J quality requirements. This is a time-consuming process where some automation is desired.

In this project, we team up with J&J's Quality Assurance wing to benchmark large language models and assess their performance in classifying regulatory requirements from health authorities into J&J quality policy.

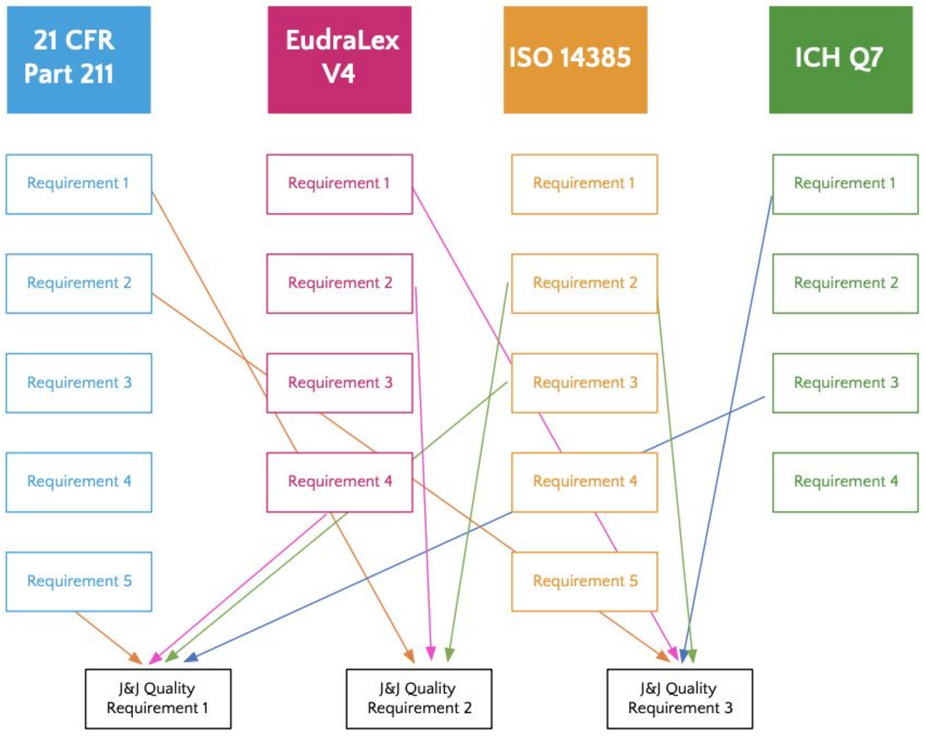

Figure 1. Policy Document Creation Workflow

Figure 2. Common Quality Topics

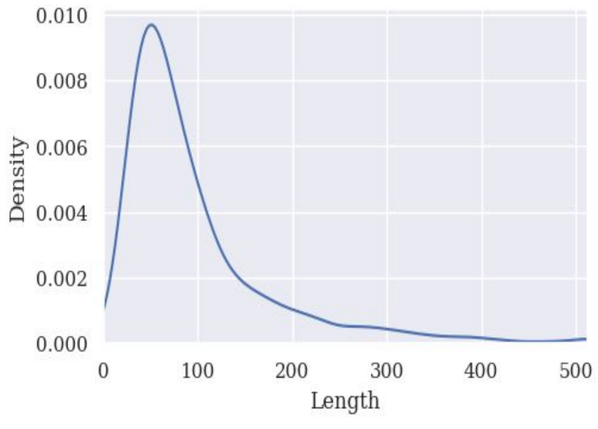

Figure 3. Regulations Length

Methods

Figure 1. gives an overview of the *multi-label topic classification* problem. Our goal is to predict, for a regulation, quality topics which are not mutually exclusive.

We use Random Forest Classifier trained on TF-IDF features as our baseline. We also fine-tune and benchmark pre-trained large language models. These include BERT from Hugging Face and Ada, Curie and Davinci from OpenAI. We then take an ensemble of BERT, Ada and Curie using a majority vote.

Figure 2. shows out of the total 40 J&J quality topics, the most common ones, along with their proportion in the test set. Figure 3. shows the length distribution of regulations.

Benchmark Results

We summarize the overall sample-wise results in Table 1. Table 2 lists the label-wise benchmark results of the best-performing ensemble method for quality topics from Figure 2.

	Random Forest	BERT	Ada	Curie	Davinci	Ensemble
Hamming Loss	0.032	0.013	0.014	0.013	0.014	**0.012**
Exact Matching Ratio	0.359	0.724	0.730	0.740	0.727	**0.753**
Precision	0.511	0.868	0.863	0.875	0.855	**0.880**
Recall	0.551	0.836	0.828	0.839	0.837	**0.839**
F1 Score	0.506	0.834	0.830	0.842	0.832	**0.845**
F2 Score	0.525	0.831	0.825	0.837	0.831	**0.839**

Table 1. Benchmark Results - Overall

	Precision	Recall	F1 Score	F2 Score
Production Process Controls	0.977	0.689	0.808	0.732
Clinical Research	0.966	0.966	0.966	0.966
Facilities	0.885	0.885	0.885	0.885
Utilities and Equipment	0.885	0.885	0.885	0.885
Labeling & Packaging	0.893	0.962	0.926	0.947
CAPA	0.842	0.801	0.821	0.808
NC	0.842	0.801	0.821	0.808
RCA	0.834	0.750	0.789	0.765
Material & Product Controls	0.760	0.826	0.792	0.812
Risk Management	0.778	0.609	0.683	0.636

Table 2. Ensemble - Label-wise evaluation

Conclusion

Large Language Models achieved significant improvement above the baseline. Of all the pre-trained models, *Curie* achieves the best performance across all metrics, although we expected Davinci to outperform Curie.

Finally, an ensemble of BERT, Ada and Curie attained the best scores.

Acknowledgments

We thank our sponsors from Johnson & Johnson, Frank Janssens & Majd Mustapha, and faculty mentors Adam Kelleher & Cathy Li for their guidance and support.