Guiding Principles for Building Scalable, Safe, Secure and Compliant Document Intelligence Systems in GenAI Era

Abstract

Documents have been a key medium through which important information is shared by users of a wide variety of popular applications across industries around the world. Document intelligence systems aim to automate the identification, extraction and processing of the key information embedded in documents. Document intelligence using AI (also known as document AI or DocAI) has become one of the most highly sought after technologies in many sectors, such as financial services, real estate, insurance, government, legal, and healthcare.

A central goal of document intelligence services is to categorize/classify, extract, and organize key information embedded in documents and securely deliver that information instantly for downstream tasks. Documents for which this goal is relevant include financial documents (receipts, invoices, contracts, mortgage documents, loan applications, purchase orders, etc.), government documents (tax forms, licenses, certificates, etc.), and more. Without automation, the amount of manual effort for document-intensive tasks – such as medical case review, evidence processing, tax preparation, small business account – can take hours.

The primary challenge with document extraction is that embedded data can be present in a combination of unstructured text; semi-structured content, such as multi-column formats, tables, and key-value pairs; and graphical content, such as figures and vector graphics. The ability to understand and interpret documents' various formats is a critical and challenging application for AI.

Potential Discussion Points

In this talk, we'll share guiding principles and key insights for building high performance, scalable document AI systems for real-world applications in highly regulated domains (with security, safety and compliant requirements). Attendees will come away with an understanding of critical success factors, including:

1. Building pipelines: how to build a system to serve high stake document intelligent use cases (with high accuracy and coverage requirements) versus long tail/extreme categories of documents/tasks.
2. Curating evaluation datasets: sampling methods for broad coverage and synthetic data generation methods to build evaluation datasets while keeping highly sensitive information secure.

3. Curating training datasets: techniques to and best practices to scale up document intelligent tasks labels for model training when human annotated fully labeled data are labor intensive and expensive and error prones. Generation of synthetic datasets that are business relevant while avoiding AI safety and security issues such as data leakages and adversarial attacks.
4. Integrating AI/Human Feedback: how to approach UI design, evaluate/take action from confidence scores, and drive continuous model improvement for specific tasks based on human and AI feedback to
5. Approaching zero-shot Inference: how to address evolving business needs over time with zero-shot label generation and inference.
6. Incorporating Large Language Models (LLMs): how to apply unimodal and multimodal LLMs in document intelligence in various components of document AI pipeline (dataset generation, weak labeling (AI judge), zero-shot inference, etc.) and mitigate challenges with LLM-based document AI in practice.

Importantly, we will describe foundational AI components in the document intelligent systems – including state-of-the-art techniques vs. the use of general purpose LLMs – and the principles on the path to a step change in the rapidly evolving landscape for document AI, including unimodal and multimodal LLMs.

Relevance to the Workshop

High performance ML systems and pipelines such as document AI require an investment in continual deep research with rapid/iterative benchmarking of new technologies, while simultaneously delivering impactful business outcomes with safety and security of highly sensitive data. Using Intuit's evolution of document AI as a case study, we'll share guiding principles and lessons learned from building and evolving document intelligence services – with iteratively improvement via rapid benchmarking of ML systems using business relevant datasets and the adoption of unimodal and multimodal LLMs. These lessons can be applied to ML systems other than document AI.

Presenter's Bio

Joy Rimchala is a Principal Data Scientist leading AI research and innovation in domains including document services and AI safety. Joy's current focus is to scale and enhance AI native capabilities – enabling scalable, secure, and compliant document intelligent service. In addition to document understanding, Joy has served as one of the lead architects in AI safety. Joy has presented her works at WiDS, NVIDIA GTC, and O'reilly AI Conference and NeurIPS workshops. Joy has served as a program committee and a reviewer at leading conferences including ACL, EACL, EMNLP, and AACL-IJCNLP. Joy holds a PhD in Biological Engineering from MIT with a focus on parameter estimation methods in cell decision processes.

Key Words

Best practice in ML workflow
ML projects life cycle management
Document Understanding