hanzo

AI-enhanced eDiscovery: **balancing security, costs, and transparency for legal**

Guide

Introduction

Welcome to our comprehensive guide on optimizing eDiscovery with AI, tailored for the legal industry. This guide is divided into three parts, each addressing crucial aspects of AI integration in legal technology. We will explore the challenges and solutions related to data security, cost efficiency, and transparency and conclude with strategies for scaling AI tools for large datasets.

Key Learning Points:

Part I:
Data security and cost control

Understand how to maintain high data security while controlling costs in eDiscovery processes.

Part II:
Enhancing transparency

Learn about the importance of transparency in AI-driven eDiscovery and how it can minimize errors and enhance the accuracy of legal document processing.

Part III:
Overcoming scaling challenges

Discover strategies to efficiently scale AI tools in eDiscovery without compromising security, cost-efficiency, or transparency.

Problems Addressed:

- The increasing volume and complexity of legal data.
- The high cost and security concerns of using AI in legal processes.
- The risk of AI hallucinations leading to incorrect or misleading information.
- The challenges of scaling AI to handle large datasets efficiently.

Part I:
Ensuring data security while controlling costs

Introduction to AI in eDiscovery

In the opening segment of our three-part look at optimizing eDiscovery with generative AI, we explore the pressing need for data security and cost-effective strategies in legal tech. The shift towards larger datasets has amplified the demand for tools that not only expand capabilities but also safeguard sensitive information and manage expenses wisely. We will delve into how Hanzo's unique approach to using Large Language Models (LLMs) in legal contexts ensures high data security and cost efficiency.

The evolution of ediscovery and the role of generative AI

For decades, the world of eDiscovery has faced multiple challenges related to processing large amounts of varied data. The volume of digital data continues to grow exponentially, with estimates from IDC suggesting that the global datasphere will reach 175 zettabytes by 2025[1].

When dealing with increasingly larger datasets, we need tools that scale efficiently and safely, making the best use of the available resources. In recent years generative AI has opened up new possibilities to decrease the workload on overworked legal professionals dealing with huge datasets.

Large Language Models (LLMs), which are language models that utilize deep learning techniques on extensive datasets as a foundation for predicting and generating natural-sounding text, are excellent at summarizing large documents and generating answers to simple questions. At the same time, generative AI has introduced new potential risks, such as hallucinations or the ability to generate text that contains an incorrect answer and deliver this text with apparent confidence. (We will discuss hallucinations later.) The use of LLMs can also be expensive, meaning that every failure costs time and money.

What is Generative AI? "Artificial intelligence that is capable of generating new content (such as images or text) in response to a submitted prompt (such as a query) by learning from a large reference database of examples."[2]

Security concerns with generative AI

It is possible to use LLMs to process large datasets in a manner that is safe, fast, cheap, and accurate, but this requires care and attention to detail. Perhaps the most important concern when dealing with large datasets for eDiscovery is data security. Today's datasets likely contain confidential information that must be secured with minimized opportunities for data leakage. Data protection is one of the hallmarks of Hanzo's platform and why Hanzo has adopted a single tenant policy whereby each customer gets its own environment, and while their data is with Hanzo, it does not leave that environment. This "walled garden" approach, a closed ecosystem in technology and computing, where the provider controls all operations, is also applied to our use of LLMs. When LLMs are needed, they can be spun up within a customer's environment, ensuring that datasets are never mixed or use shared resources.

"Only 24% of generative AI projects are being secured." [3]

Cost implications of using LLMs

On the topic of cost, we know that our customers do not have infinitely deep pockets, and any solution must come at a reasonable price. Keeping costs low is largely a matter of choosing the right tools for the right problems and using those tools as effectively as possible. Keeping LLMs "hot" and running on machines is expensive, and the GPUs required are a limited, and precious resource, which is why Hanzo only spins up what it needs when it is needed.

Hanzo's approach to data security and cost efficiency

Hanzo understands this problem and created an infrastructure for LLMs when, and only when, called upon for use. This means we don't need to count API requests or tokens, we can count GPU hours instead. By using this approach, we can focus on setting up the workload to make the best use of those GPU hours, addressing the root source of costs, and incentivizing real savings as directly as possible.

[1] Source. Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, November 2018.

[2] Merriam-Webster. (n.d.). Generative AI. In Merriam-Webster.com. Retrieved June 25, 2024, from https://www.merriam-webster.com/dictionary/generative%20AI

[3] Rodgers, Clarke, Moumita Saha, Dimple Ahluwalia, Kevin Skapinetz, and Gerald Parham. Securing generative AI, What matters now. IBM. May 2024

Part II:
Enhancing transparency

Recalling our discussion on data security and cost management, this second part focuses on the critical element of transparency in the use of Large Language Models (LLMs). Understanding how AI tools derive their outputs is essential for legal professionals to trust and effectively use the technology.

In eDiscovery, transparency in AI applications is crucial because it ensures the reliability and fairness of AI-generated evidence. Without transparency into how AI algorithms categorize and prioritize documents, courts may question the accuracy and potential biases of AI-generated findings. This transparency is essential for upholding due process, allowing parties to effectively challenge and validate evidence, and maintaining the integrity of legal proceedings.

We will discuss how Hanzo ensures that users are fully informed about the workings and outputs of LLMs, enhancing reliability and reducing the risk of errors.

Understanding LLMs and their training data

LLMs are trained on huge "foundational datasets" of public data. These foundational datasets include (but are not limited to) all of Wikipedia, the works of Shakespeare, billions of public web pages, legal and financial documents, and other content. When LLMs generate content, they can draw on patterns from these foundational datasets to decide what text to generate. This means that when using common LLM tools in today's marketplace and asking the LLM to generate content, it can generate text based on what we ask it, as well as the content of the foundational dataset.

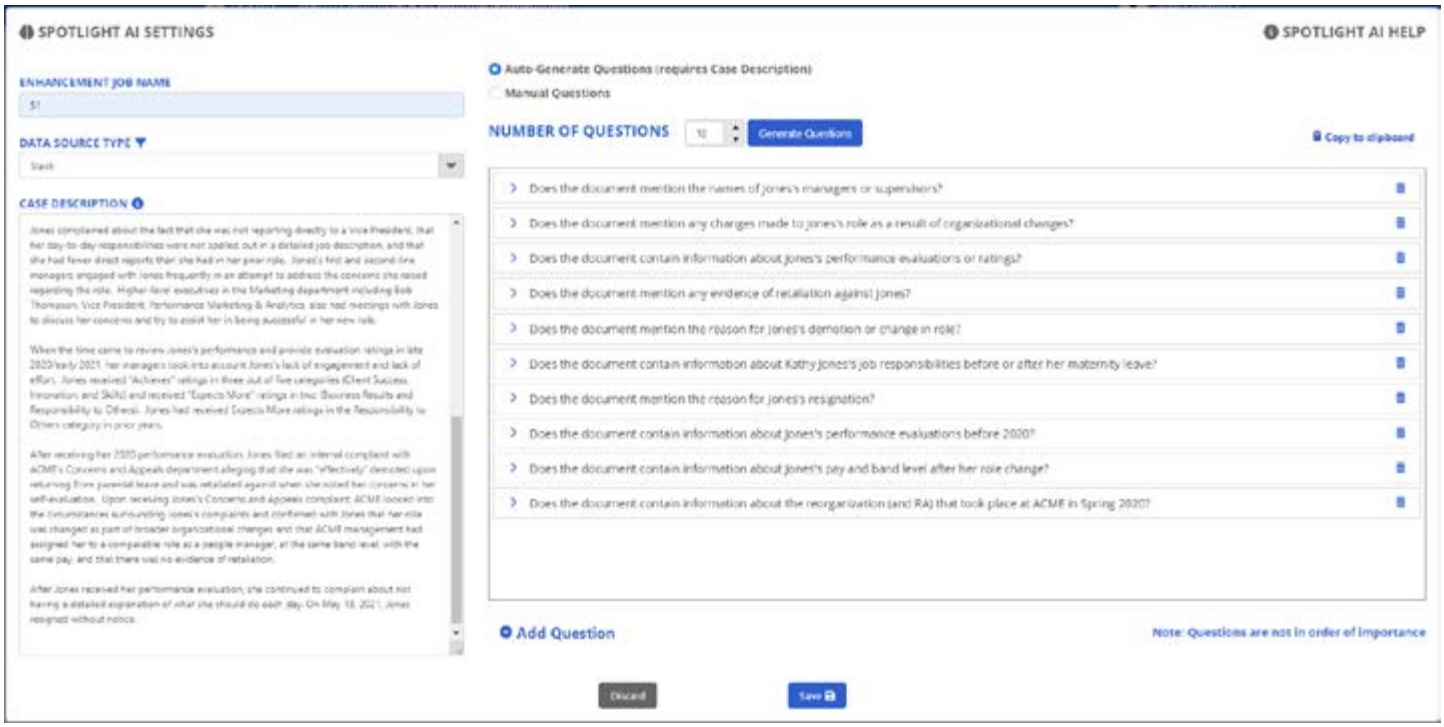
The problem of hallucinations in LLMs

When an LLM generates text based on the user's input and the foundational dataset, this process can lead to "hallucinations," where the LLM produces text that is either factually incorrect or not relevant to the query. Such hallucinations can present significant risks in the legal industry, as they can result in misleading or inaccurate information being considered credible, often without adequate verification.

Transparency measures to avoid hallucinations

When using the common LLMs, there is the possibility of hallucinations, and we need to be very careful to avoid them. Hanzo's strategy is to transparently expose responses with the largest scope for hallucination and bring them to the user's attention before the dataset is analyzed. In this context, "scope" means the extent to which the LLM can generate false, misleading, or irrelevant content. Hanzo's Spotlight AI feature generates questions to help users understand how the AI is going to determine if parts of the dataset are relevant to a given case.

Transparently showing this information to the user provides understanding and oversight of the data analysis and guidance to the LLMs for identifying the most relevant content. These questions are easy to understand and can be adjusted before any user data is run through the process, thus saving valuable time and cost by providing full transparency upfront before any work is performed, resulting in better results.



Spotlight AI question panel within Hanzo Illuminate allows users to easily review and modify before running them across the dataset.

Spotlight AI: enhancing decision-making

Spotlight AI's workflow uses methods to reduce the scope of hallucinations to a problem of traditional Type I (False positive) and Type II (False negative) errors. Rather than having responses that are a synthetic mix of fact and fantasy, each answer is tuned to be either correct or incorrect. This process makes it easier for users to understand and easier to measure success. When the user evaluates Spotlight AI's relevant content, they see the rationale for the decision, including the questions that led to it.

Importance of transparency in complex cases

Dealing with legal cases and datasets is rarely straightforward, and answers need to be complex. This is why transparency is so important: Given the complexities of a given case, at least the decisions about each message or document can be black and white and traced back to a clear rationale. In the case of Spotlight AI, once the Spotlight AI enhancement process is run, Hanzo returns unique, item-level details as to why the Spotlight relevancy engine deemed a message, email, or document as relevant. This leaves the user in control of the discovery process and with a clear understanding of the output. Any steps that make the discovery process more opaque or put unnecessary distance between the user and the dataset will introduce a risk of content being missed. Hanzo provides exported data intelligence along with content that can be viewed outside of the Hanzo platform to make things easier. This shared intelligence provides enrichment details that can be very helpful to outside reviewers.

Part III:
Overcoming Scaling Challenges

Introduction to scaling challenges

In earlier sections, we explored the application of large language models (LLMs) in eDiscovery, emphasizing the significance of data security, cost efficiency, and transparent data analysis. We also examined a unique risk associated with LLMs known as "hallucinations," where the model generates inaccurate or irrelevant text. Now, we will integrate these insights and address the scalability challenge.

Meeting multiple requirements with LLMs

As we have seen, Hanzo effectively meets several critical requirements in the deployment of Large Language Models (LLMs) for legal eDiscovery:

- **Data security:** Datasets are segregated safely in customer environments to maintain a single-tenant policy. This method minimizes the risk of data breaches and unauthorized access, providing a robust security framework that is essential in handling sensitive legal information.
- **Cost:** To manage costs effectively, Hanzo utilizes LLMs in a targeted manner. Hanzo ensures that computational resources are used efficiently by deploying the smallest appropriate model for each specific task. Furthermore, the operational model is designed so that machines remain active only for the time needed to complete the task, significantly reducing unnecessary expenditures on processing power and energy.
- **Transparency:** To address the issue of hallucinations—where LLMs might generate inaccurate or irrelevant information—Hanzo has established stringent controls. The system is designed to either return the AI-generated content to the user for a thorough review or limit responses to simple yes/no answers. This method not only mitigates the risk of misinformation but also enhances the transparency of the AI processes, enabling users to understand and trust the results provided by t he LLMs.

Challenges of scaling LLMs in eDiscovery

The main challenge we face with LLMs is scale. LLMs are expensive to run due to the high-capacity computing resources like GPUs, which are required for their efficient operation. Sometimes, the necessary hardware isn't available. This mismatch of supply and demand leads to delays in dataset analysis and raises concerns about cost. Solving this scale issue is critical because a secure, cost-effective, and transparent solution is pointless without the essential hardware to analyze your datasets. However, even though LLMs require more expensive hardware compared to traditional machine learning models used in eDiscovery, the overall cost remains lower than CAL/TAR due to the elimination of human costs.

Strategic scaling solutions with the right LLM

By choosing the best LLM for the task and tuning the deployments, Hanzo is able to engineer a solution that requires affordable and abundant hardware. When we need extra capacity, we can horizontally scale up more machines for as long as necessary. Understanding how datasets and tasks scale is crucial to understanding how the workload and hardware demands scale. Doubling the size of the dataset should double the workload, and so should doubling the number of questions posed.

Impact of scaling on data processing

Being able to scale up data processing can be crucial when time is limited, datasets grow in size, or additional processing is needed at a later time. By keeping the LLM-based data processing within customer environments, we ensure that customers do not compete for shared resources and that processing can be scaled up as long as resources are available. This also means that there are no quotas, rate limits, or API tokens to worry about and that costs scale linearly with processing time.

Challenges of scaling LLMs in eDiscovery

The main challenge we face with LLMs is scale. LLMs are expensive to run due to the high-capacity computing resources like GPUs, which are required for their efficient operation. Sometimes, the necessary hardware isn't available. This mismatch of supply and demand leads to delays in dataset analysis and raises concerns about cost. Solving this scale issue is critical because a secure, cost-effective, and transparent solution is pointless without the essential hardware to analyze your datasets. However, even though LLMs require more expensive hardware compared to traditional machine learning models used in eDiscovery, the overall cost remains lower than CAL/TAR due to the elimination of human costs.

Wrapping Up:
key takeaways from our journey through AI-enhanced eDiscovery

Throughout this guide, we've examined the intricate balance of cost, scale, and transparency in deploying AI in legal eDiscovery. Hanzo's strategic integration of LLMs puts data security at the heart of data processing with LLMs, makes the solution cost-effective and scalable, and keeps the process as transparent as possible. This approach underscores our commitment to providing secure, cost-efficient, and transparent AI solutions essential for navigating the complexities of modern legal challenges. By aligning closely with the needs of corporate legal teams and legal service providers and by prioritizing these principles, Hanzo not only enhances the efficiency of eDiscovery processes but also improves relevancy assessments.

hanzo

Contact us
today

Sales
contact@hanzo.co

Customer Support
support@hanzo.co