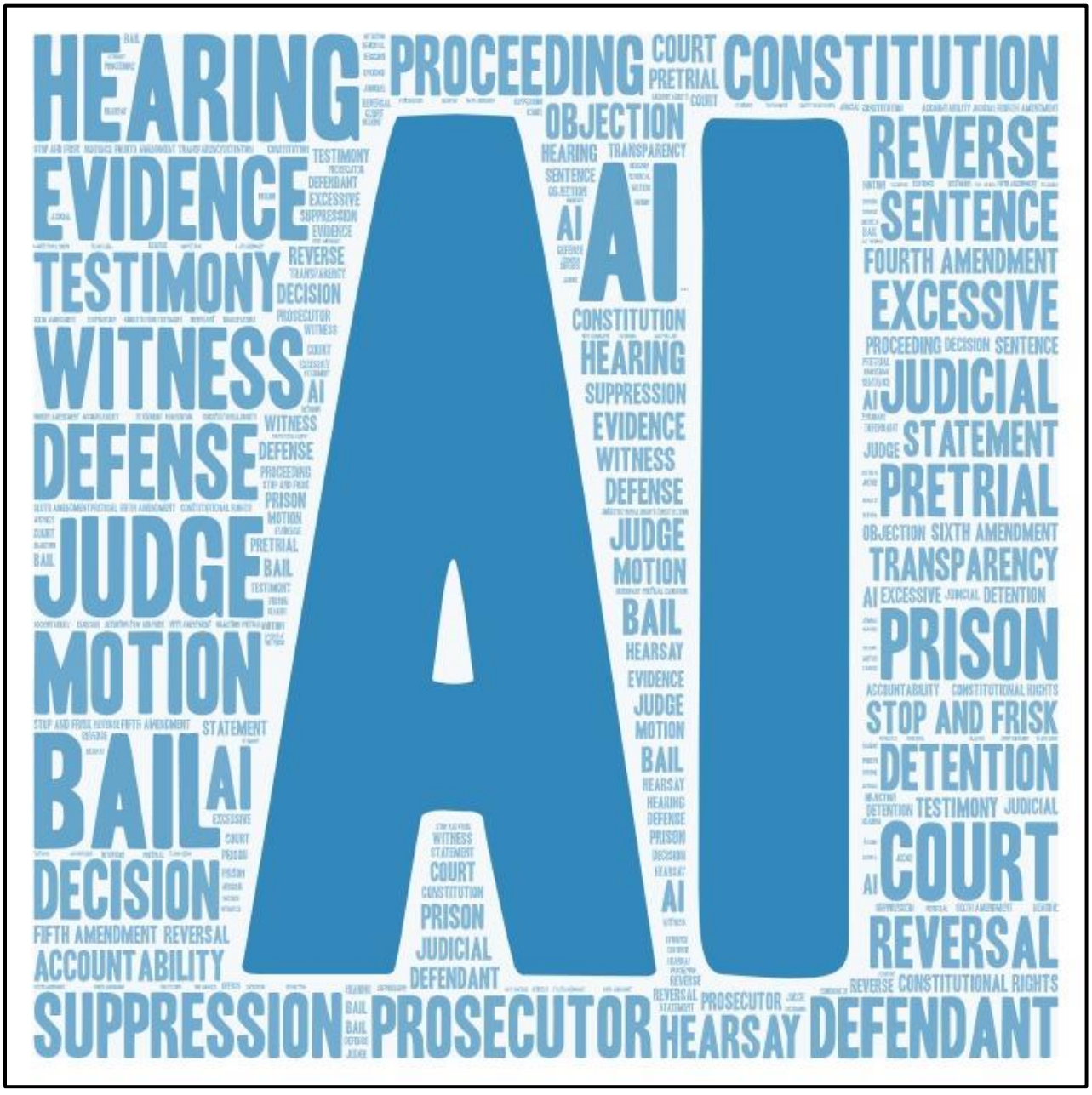
Harnessing AI To Enhance Judicial Transparency



Scrutinize | September 2024

Harnessing AI To Enhance Judicial Transparency

This factsheet introduces a new method for enhancing judicial transparency using large language models (LLMs) to analyze judicial texts. By converting appellate decisions into data, this method identifies complex legal issues and their outcomes, offering valuable insights for policymakers and the public.

Our previous work demonstrated the significance of such judicial data when reports we produced with algorithmic methods fueled several campaigns calling for judicial accountability in New York.

This factsheet explains how LLM-powered analysis can further accelerate data development to enhance judicial transparency. We have documented our development process and methodology in greater detail internally and would be happy to collaborate with researchers or organizations interested in conducting their own LLM-based analysis to enhance judicial transparency.

Appellate Decisions Can Be Used as Data to Evaluate the Judiciary

Over the past year, Scrutinize has developed a novel approach to judicial transparency by using algorithms to extract quantitative metrics from appellate court decisions. This method transforms legal texts into structured, analyzable datasets, enabling researchers to quantify judicial practices across multiple levels—from individual judges to entire court systems.

This hybrid approach has successfully classified key outcomes in appellate decisions, as shown in our previous reports, *Excessive Sentencers* and *Reverse and Reassign*.

Algorithmic Analysis Is Constrained for Complex, Context-Specific Classification

Our recent report, *Unprotected*, on judicial errors in protecting constitutional rights revealed a constraint in our analysis method. The report required identifying appellate decisions that overturned lower court rulings specifically due to suppression errors. Our algorithm could identify appellate decisions that reversed lower courts and discussed suppression, but it could not reliably determine if suppression was the cause of the reversal due to the

lack of uniform signaling language.

As a result, we manually classified over 1,000 decisions, a labor-intensive process that required carefully reading lengthy and complex legal texts.

Enhancing Appellate Decision Classification Using LLMs

To address the linguistic challenges encountered in *Unprotected*, we developed a proof of concept for an LLM-powered methodology to replace manual classification of suppression decisions. This methodology involves a two-step process:

1. Summarization: GPT-3.5-turbo-1106 generates a summary of each appellate decision.[1]
2. Classification: GPT-4-1106-preview uses summaries to determine if the appellate decision overturned lower court rulings specifically on suppression issues.[2]
3. Repetition: Steps 1-2 are applied to each decision five times, producing five different classification (this leverages non-deterministic variation, as discussed below).
4. Review of Non-Consensus Outcomes: For decisions where Step 3 does not result in a consensus—meaning that all five runs do *not* yield the same classification—we manually classify the case.

Steps 3 and 4 of our methodology improve accuracy by leveraging the inherent randomness in GPT-3.5 model outputs. Non-deterministic systems like GPT models are probabilistic, producing varying outputs for the same input. Our tests showed that output variation increases with the complexity of the decision text. While most correctly classified decisions consistently yield the same result across runs, more complex decisions tend to have lower average classification scores. To capitalize on this variation, we classify each decision five times (Step 3). We accept decisions that reach consensus but manually check those with divided results (Step 4). This approach allows us to harness the GPT models' non-deterministic nature, enhancing overall accuracy, particularly for more challenging cases.

---

[1] At the time of testing, GPT-4-1106-preview was the most advanced model available. To reduce costs, we used GPT-3.5 for initial summaries and then fed those into GPT-4-1106-preview instead of processing full decisions. However, with falling API costs and improved models now on the market, this approach is likely no longer needed.

[2] Our prompts to the LLMs include specific instructions for summarization and classification, including guidelines for classifying edge cases.

In Test, The LLM Methodology Achieves 99% Accuracy Rate

Our LLM methodology achieved a 99% accuracy rate when tested on 1,035 appellate decisions from *Unprotected*. 89% of cases (921) were correctly, unanimously classified across all five runs, 10% of cases (104) had divided classifications and required manual review, and 1% of cases (10) were incorrectly classified by all five runs.

Thus, the LLM-powered methodology we developed demonstrates significant potential for time savings while maintaining a high accuracy rate.

Expanding Judicial Transparency with Flexible LLM Analysis

LLM-powered analysis can significantly enhance judicial transparency. It does not rely on uniform legal phrasing, allowing for context-based textual analysis. LLMs' inherent adaptability and scalability streamline labor-intensive data development, reducing both time and cost.

These capabilities open new possibilities for quantitative analysis in judicial transparency. LLMs enable large-scale examination of judicial texts and make it easier for organizations to create their own metrics with less programming expertise than algorithmic methods.