# LLM data extraction: Unstructured records of police violence & police misconduct

Tristan Chambers
*Epic Retreat 2024*

Police violence and misconduct remain a stubborn issue in the United States. Landmark California legislation SB 1421 and SB 16 promise to make big changes in policing. These records may shed light on the issue, promoting more accountability. However, the records released by agencies are **long** and **disorganized** and **laborious** to interpret, even for experienced subject matter experts. Moreover, they contain highly graphic depictions that can be **traumatizing** for data entry staff. These challenges threaten to keep data about policing in the dark. How do we go from these opportunities and challenges to actionable data?

## Large Language Models?

Large Language Models (LLMs) have an impressive ability to interpret natural language. Can LLMs be used to automate the extraction of data?

### LLM Challenges

- Limited context window
- Limited reasoning
- Maintaining state between chunks
- Off-the-shelf models lack domain knowledge & concepts
- Concerns about LLM “hallucinations”

### LLM Tactics

- Extract high level summaries from large blocks of text that focus only on very basic facts.
- Extract detailed summaries only from smaller chunks.
- Focus on natural language summaries not structured data when parsing the messy source text.
- Don’t prompt the model to provide exact sources at the same time as summarization or extraction tasks. Make this a separate post-facto step.

## Autofolio

Autofolio takes a batch of disorganized files and extracts basic data points and clusters on these points to associate files together into cases.

### Approach

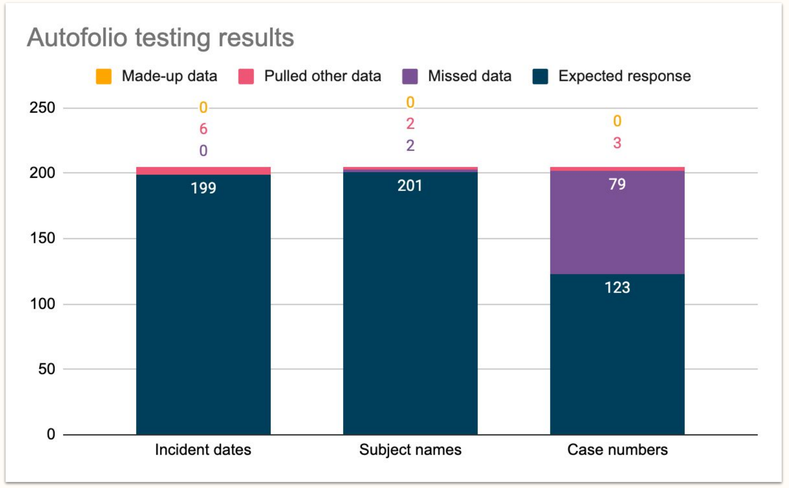
The high level attributes used for clustering like incident date, subject name, and case numbers appear frequently and are often made clear throughout the text. This makes it possible to use large context windows to scan over entire documents, or large portions of them.

### Results

**Incident dates**:
- 97% accurate
- 0% false negative rate
- 0% hallucination rate

**Case numbers**:
- 99% accurate
- 39% false negative rate
- 0% hallucination rate

**Subject names**:
- 99% accurate
- 1% false negative rate
- 0% hallucination rate




## Extractor

The Extractor combines all of the files into a concise summary, extracts structured data from the summary, and validates assertions by citing the original source documents.

### Approach

#### 1. Summarization

High Level Summary

First extract “high level” summary. This summary is produced by doing a summary over large chunks of text, up to 100 pages, but only seeking the most obvious details.

Detailed Running Summary

Fifteen page chunks are fed into the model’s context window, prepended with the “high level” summary, as well as a “running” summary. The “running” summary accumulates more information after each chunk is processed. The result is a 1:100 “compression” of source text to final summary.

#### 2. Structured data extraction

Extract structured data from the summarized case files. Use few-shot learning examples to illustrate domain concepts to the model.

#### 3. Citation & Validation

Prompt model for quotes from original source text. Search for cited text accounting for typographic variations.

### Results

Proof of concept code ran on a limited data set of three cases successfully identified: Incident date, incident location, names of officers directly involved, officer badge numbers and ranks, actions taken by the involved officers, subjects involved, subject injuries, and produced zero hallucinations on these fields. Larger sample coming soon!