Can LLMs Extract Useful Data from News Articles for Police Accountability?

Jesse Loi, MS Data Science

Faculty Advisor Ariana Mendible, PhD

SEATTLE UNIVERSITY

Community Need

Police Accountability Data Is Not Commonly In Structured Form

Many community organizations struggle to find structured data on police misconduct. We turn to large language models (LLMs) for assistance.

Key Questions

- Can an LLM successfully label a location from news reports of instances of police misconduct?
- If so, how well do the LLM's location labels match a human's?
- How does the human-LLM match compare to human-human match?

Hypothesis

A large language model will match human responses when asked to identify the location of an incident.

Data Collection

Data Source: **CUAPB** Database: Contains ~1300 articles on police misconduct in Minneapolis.

COMMUNITIES UNITED AGAINST POLICE BRUTALITY

Data Scraping: Used **PyMuPDF** to scrape the data from a majority of them.

Data Labeling: Set up tools to facilitate human labeling, **Baserow** and **Zooniverse.**

Methodology

Human Data Labeling

We crowdsourced article labels in two **Datathon events** across Seattle U, Carleton College, and Hamline University. Across both, over 40 volunteers labeled **860 articles** and revalidating 126 of those for human-human comparison.

LLM Data Labeling

We used Ollama to implement Meta's **Llama 3.1B model** to label locations from all articles. This model is small and lightweight, increasing its accessibility.

Comparison Metric

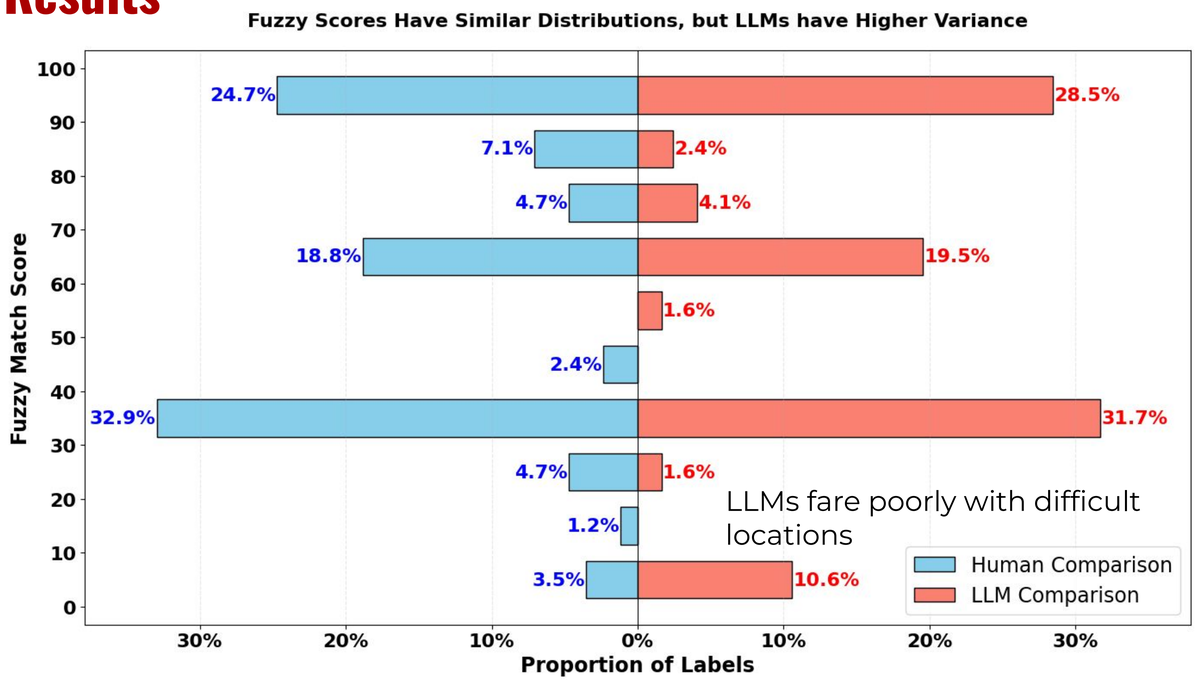
Fuzzy matching (0-100) compares subset matches between strings and is more generous than exact string matches.

53rd and Emerson Avenues N.	**93%**	Intersection of 53rd and Emerson Avenues N
1235 Reform St	**63%**	520 Reform St., Norwood Young America MN
N/A	**0%**	Minnesota City Court

Analysis

We compare distributions of fuzzy matching scores between **human-LLM** data and the **human-human** data.

Results



This figure shows how humans compare to each other in terms of response similarity with how the LLM compares to a human response with respect to similarity.

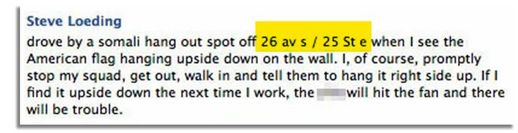
Discussion

Limitations

- **Encoding Errors:** Private use area (PUA) unicode text was improperly formatted, leading to errors.

  ``` ``` Reform St. → 530 Reform St.

  Encoding errors were a main source of model hallucination.
- **Text Embedded Images**: Screenshots and images with words could not be read by the scraper, leading to worse off responses, e.g.:

  
  Steve Loeding
  drove by a somali hang out spot off 26 av s / 25 St e when I see the American flag hanging upside down on the wall. I, of course, promptly stop my squad, get out, walk in and tell them to hang it right side up. If I find it upside down the next time I work, the will hit the fan and there will be trouble.
- **Overconfidence**: Increased difficulty for LLMs to give a human-like response when an answer is not readily available, instead giving a vague answer.

Strengths

- Fuzzy score distribution between human-human scores and LLM-human scores is similar, explaining some noise.
- Lack of hallucination outside of encoding errors, for which solutions are available.
- Human-LLM score is, on average, higher than human-human scores.

Future Work

- Community mapping: Use the data to map instances of police misconduct to identify trends.
- OCR implementation: Implement optical character recognition for improved scraping accuracy.
- LLM Guardrails: Utilize agentic AI to first detect if there is a location, then next deploy the model to attempt to locate.
  ```