Improper Conduct: A Human-AI Collaboration Framework for Identifying Prosecutorial Misconduct

Students: *Begum Gokmen, Dina Blachman*

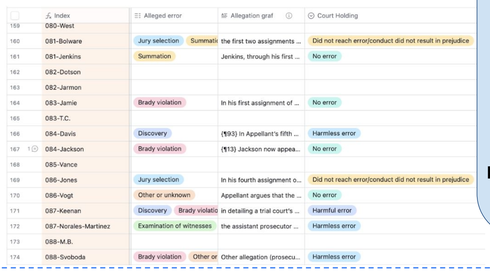
Supervisor: *Prof. Smaranda Muresan*

Computer Science Department, Barnard College

Introduction & Background

Collaborators: Columbia Journalism Investigations, Columbia Law School

Motivation: Prosecutorial misconduct is difficult to identify systematically. Columbia Journalism Investigations' fellows spent two years to manually compile a database of appellate Ohio court rulings from 2018 to 2021 (989 cases) and to label cases of prosecutorial misconduct.



	Index	Alleged error	Allegation graf	Court Holding
159	080-West			
160	081-Bolware	Jury selection Summatio	the first two assignments ...	Did not reach error/conduct did not result in prejudice
161	081-Jenkins	Summation	Jenkins, through his first ...	No error
162	082-Dotson			
163	082-Jarmon			
164	083-Jamie	Brady violation	In his first assignment of ...	No error
165	083-T.C.			
166	084-Davis	Discovery	(¶93) In Appellant's fifth ...	Harmless error
167	084-Jackson	Brady violation	(¶13) Jackson now appea...	No error
168	085-Vance			
169	086-Jones	Jury selection	In his fourth assignment o...	Did not reach error/conduct did not result in prejudice
170	086-Vogt	Other or unknown	Appellant argues that the ...	No error
171	087-Keenan	Discovery Brady violatio	in detailing a trial court's ...	Harmful error
172	087-Norales-Martinez	Examination of witnesses	the assistant prosecutor ...	Harmless error
173	088-M.B.			
174	088-Svoboda	Brady violation Other or	Other allegation (prosecu...	Harmless error

What is Prosecutorial Misconduct?

Unethical or illegal tactics by prosecutors in a criminal case. This **alleged misconduct** can satisfy **any** of these **7** broad categories: Brady violation, Discovery, Jury selection, Opening statements, Examination Of Witnesses, Summation, Plea Deal, or other/unknown types of error.

Why do we track it?

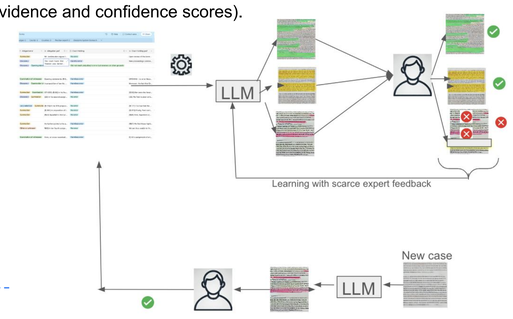
To identify **misconduct hotspots** and **repeat offender prosecutors**.

Goal

Develop a human-AI collaboration framework by training an open-source Large Language Model to assist journalists in identifying prosecutorial misconduct from appellate court rulings. Two main tasks:

(1) Identify candidate cases;

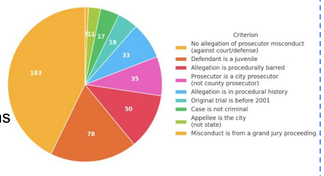
(2) Classify candidate cases by type of misconduct ruling (providing evidence and confidence scores).

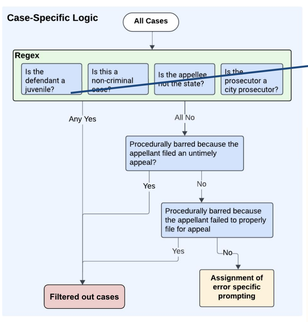


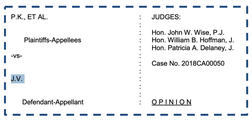
Methods

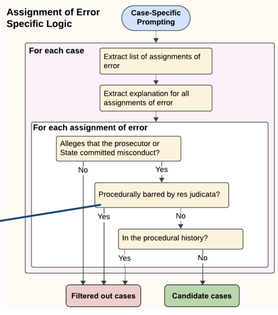
Task 1: Identifying Candidate Cases

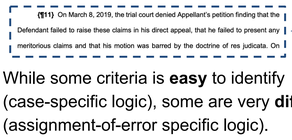
- Expert-guided pipeline to identify candidate appellate court cases that might contain prosecutorial misconduct
- Combination of regular expressions (regex) and LLM inference (**Ministral-8B-Instruct-2410**) depending on complexity of criteria.











While some criteria is **easy** to identify (case-specific logic), some are very **difficult** (assignment-of-error specific logic).

Subtask: Evidence extraction

For each allegation, in each case:

Allegation: The trial court abused… + **Case text** → **Allegation graf:** Appellant next argues… / **Holding graf:** …error is overruled.

To evaluate performance, we used **AlignScore** to compare the human-extracted gold label against the LLM extracted graf. We tested both **Ministral** and **GPT-4.1**.

Task 2: Identifying Error Types

Court Holding
Harmful error ← Reversal
Harmless error
No error
"Did not reach"

Alleged error: Jury selection, Brady violation, Other

Alleged error column lists 1+ labels corresponding to types of errors mentioned.

Allegation graf: "In his first AOE, Appellant argues an error in jury selection…in his second AOE…in his third AOE… in his fourth AOE…"

Allegation graf does not have a consistent structure between cases.

Court Holding graf: "... We find that [Brady violation] constitutes a harmless error…"

Court holding will often leave non-harmful errors unaddressed, or address errors out of order → **difficult for LLMs to understand chain of thought**

Subtask: Extract individual error grafs for training

The new database maps each alleged error in a case to excerpts of corresponding grafs. Only the most egregious error includes a court holding, enabling clearer model training on error types and outcomes.

Alleged error: Brady violation → **Allegation graf:** "In his third AOE, Appellant argues…" → **Court Holding graf:** "... We find that [Brady violation] constitutes a harmless error…"

…

Alleged error: Jury selection → **Allegation graf:** "In his first AOE, Appellant argues…" → **Court Holding graf:** [BLANK]

Results

Task 1: Identifying Candidate Cases

Accuracy	0.71520
Precision	0.68279
Recall	0.89647
F1	0.77517

We prioritize minimizing false negatives, which can be measured by the **recall** score.

"Extract the text that further discusses and analyzes the allegation of error." → "You are given the full text of a court opinion with one or more allegations of prosecutorial misconduct. Your responses must be direct quotes from the case text. For each allegation identified in the case, locate two distinct, contiguous blocks of paragraph(s): The ALLEGATION DISCUSSION: all paragraph(s) that further discuss or analyze the alleged error. The COURT HOLDING: all paragraph(s) in which the court explicitly rules on that error."

Subtask: Evidence extraction

Index	Allegation_N	LLM_Extraction	Human_Extraction	AlignScore
008-Groce	1	In his first assignment of error, Groce argues his conv	(¬ 55) In his fourth assignment of error,	0.012059428
008-Groce	2	In his second assignment of error, Groce argues the t	(¬ 55) In his fourth assignment of error,	0.038004696
008-Groce	3	In his third assignment of error, Groce argues the tria	(¬ 55) In his fourth assignment of error,	0.040942501
008-Groce	4	In his fourth assignment of error, Groce argues the pr	(¬ 55) In his fourth assignment of error,	0.980377257
008-Groce	5	In his fifth assignment of error, Groce argues he recei	(¬ 55) In his fourth assignment of error,	0.01283892
008-Groce	6	In his sixth assignment of error, Groce argues the tria	(¬ 55) In his fourth assignment of error,	0.063163079

Total Errors	Cases Sampled	Unique Matches	Correct Matches	Retrieval Accuracy
46	10	11	8	72 %

Average AlignScore for correct LLM-Human matches was **0.74**.

Ministral vs. GPT evidence extraction

Pipeline results for 20 sample cases, evenly split between classes eveal using evidence extracted by Ministral yield better results than GPT. Additionally, **Macro-average recall** for Ministral is **61**% while GPT is **50**%.

Task 2: Identifying Error Types

Identify court holding label

Classification Report				
	Precision	**Recall**	**F1**	**Support**
No Error	0.92	0.87	**0.89**	52
Harmful	0.62	0.81	**0.70**	16
Harmless	1.00	0.50	**0.67**	4

Ministral-8B-Instruct-2410 is fine-tuned given **most egregious** allegation graf and court holding. The **model struggles with nuance of harmless errors**, in which an error has occurred but it is not egregious enough to overturn the court ruling.

Ex. 241-Rouzier

A prosecutor's use of "I believe" and "I think" to preface credibility arguments **should be avoided**…

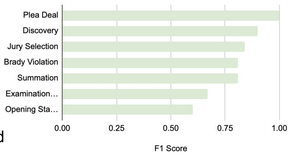
Whether improper prosecutorial commentary during closing argument rises to reversible error involves a separate analysis. **….those comments did not affect the trial's result and constitute plain error**.

→ Harmless error ✓ / No error X

Identify type of error given allegation graf

"[INST] Allegation: <text>. Respond with one label: Brady violation, Discovery, …, Other. [/INST]"

Model performs with **~79% accuracy** on the test dataset. The model performs stronger on error types which are typically procedurally narrow and weaker on those with semantic overlap.



Conclusions

The primary challenge was handling the unstructured and inconsistent nature of text across court opinions and manual allegation extractions. Future work will involve ongoing prompt refinement and interdisciplinary collaboration with journalism students to address remaining data gaps.

Next Steps

(1) Extract evidence with Ministral for all cases with optimized prompt and re-run models.

(2) Fine-tune on more difficult tasks with longer context lengths:

Input: All allegation grafs (+ Court holding grafs) → **Output**: Court holding

Input: Extracted text from Task 1 → **Output**: Court holding

References

"Mistralai/Ministral-8B-Instruct-2410." Hugging Face, huggingface.co/mistralai/Ministral-8B-Instruct-2410. Accessed 25 July 2025.

Yuh-Zha. "ACL2023 - AlignScore, a Metric for Factual Consistency Evaluation." GitHub, github.com/yuh-zha/AlignScore. Accessed 25 July 2025.

GPT-4.1 Model, platform.openai.com/docs/models/gpt-4.1. Accessed 25 July 2025.

Acknowledgements

Barnard College and the Office of the Provost, "Magic Grant" from the David and Helen Gurley Brown Institute for Media Innovation, faculty and student collaborators from Columbia Journalism Investigations and Columbia Law School