From Manual Coding to AI Assistance

Using LLMs to Classify and Summarize Judicial Evaluation Comments

John M Charles, MS – Brian Robertson, PhD – Cooper Kelley – Xiaolei Pan · Market Decisions Research

Background & Problem Statement

State judicial performance evaluation (JPE) programs collect thousands of open-ended survey responses from attorneys, jurors, and court users each evaluation cycle. These comments inform both **judicial professional development** and **voter decision-making** prior to retention elections.

Processing these comments manually is resource-intensive. Staff must determine: (1) which comments contain meaningful performance feedback, (2) whether comments are correctly attributed to the intended judge, and (3) what themes emerge across performance domains — all before reports can be written or reviewed.

The Colorado Office of Judicial Performance Evaluation (COJPE) provided the operational context for this proof-of-concept study.

Research Objectives

1 Attribution Validation

Detect comments likely misattributed to the wrong judge by comparing comment text against judge-level metadata: name, gender, pronouns, and judicial title.

2 Constructiveness Classification

Apply a structured rubric to separate behavior-based, actionable feedback from vague, irrelevant, placeholder, or discriminatory content.

3 Thematic Summarization

Generate judge-level narrative summaries across performance domains: communication, fairness, legal reasoning, demeanor, and case management.

4 Accuracy Evaluation

Benchmark automated model outputs against manual coding to measure accuracy for each classification task.

Data Source

Dataset: Colorado JPE open-ended survey responses (2025-2026 Survey Administration)

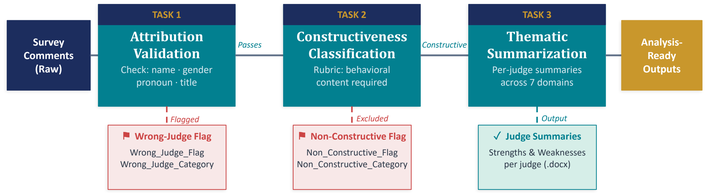
Unit of analysis: Individual comment rows, each linked to a named judge with associated metadata

Metadata fields: CommentID, judge name (full and last), gender, title, court, and performance category

Volume: **9,397 comments** across **387 judges**

Methodology

A custom LLM-based agent was developed within **Microsoft Copilot** and deployed across three sequential analytic components:



1 Attribution Validation (Task 1)

Each comment is evaluated against judge-level metadata to detect likely misattribution. The agent checks for spelling errors in the judge's name, name mismatches, gender pronoun mismatches, and title mismatches. Flagged cases are marked with a structured category label.

2 Constructiveness Classification (Task 2)

Unflagged comments are evaluated using a structured rubric. A comment is constructive only if it contains behavioral or performance-related information — positive or negative. Non-constructive comments (placeholders, vague praise/criticism, irrelevant content, discriminatory language) are labeled and excluded from summarization.

3 Thematic Judge Summaries (Task 3)

Constructive comments are grouped by judge and performance category. The agent produces two-paragraph summaries per judge (Strengths and Weaknesses) across domains including: communication clarity, fairness, legal reasoning, demeanor, case management, preparation, and timeliness.

Prompt Engineering & Validation Logic

Agent Instruction Design

Prompts were engineered to minimize false positives in name and pronoun detection. Common dictionary words (e.g., "long," "king") were explicitly excluded from name-match logic. Pronoun validation was restricted to pronouns clearly referring to the judge, not third parties. Title mismatches were flagged only when strongly indicating a different person.

Output Structure

Tasks 1 & 2 produce structured CSV outputs with four new binary/categorical variables: `Wrong_Judge_Flag`, `Wrong_Judge_Category`, `Non_Constructive_Flag`, and `Non_Constructive_Category`.

Human Review

Automated decisions are subject to human review. The workflow is designed so that reviewers can efficiently inspect flagged cases without re-processing the full dataset.

Results

328 Comments Flagged as Potential Wrong-Judge	**4%** Wrong-Judge Flag Rate	**13%** Human Agreement
278 Comments Flagged as Non-Constructive	**3%** Non-Constructive Flag Rate	**31%** Human Agreement

Metric	Value	Interpretation
Accuracy	49%	Correct classifications in reviewed subset
Precision	21%	Of flagged comments, 21% confirmed by human
Recall	55%	Of human-identified flags, AI caught 55%
F1 Score	31%	Harmonic mean of precision and recall
Specificity	49%	Correct identification of true non-flags (AI=No, Human=No)

Notes on Accuracy: These results reflect **ONLY ONE processing run against one dataset**. Prompt **adjustments and re-testing are recommended before treating the precision/recall figures as stable benchmarks**. Prompt engineering constraints, particularly conservative name/pronoun detection, can substantially reduce false positives without sacrificing recall.

Classification Findings

Wrong-Judge Detection: The AI was over-sensitive to minor textual proximity between judge names and ordinary words or abbreviations in the comment text. It also confused possessives and informal name references with genuine misattribution.

Constructiveness Classification: The AI appears to over-trigger on negative sentiment language without adequately distinguishing relevant criticism of judicial conduct from personal attacks.

Summary Quality: Qualitative review of generated judge summaries revealed a tendency toward generic, meta-descriptive language in place of the behavior-specific synthesis required by the prompt, indicating that summary quality represents an area for further prompt refinement.

> *"Many respondents described the judge as respectful, and patient in this domain, demonstrating a clear commitment to excellence and skill. Few remarks suggested negative impressions within this area of performance, allowing focus to remain upon the positives instead."*

Conclusions & Implications

This proof-of-concept demonstrates that structured LLM-based agents can process high-volume qualitative survey data with **measurable accuracy *(~50% in one processing)* and methodological transparency**. Key takeaways:

Scalability: The pipeline processes thousands of comments in a fraction of the time required by manual coding, enabling faster reporting cycles.

Human oversight: Structured flag outputs allow efficient human review of automated decisions, maintaining accountability without re-doing full processing.

The authors thank the Colorado Office of Judicial Performance Evaluation Committee for access to the operational context for this work.
AAPOR 81st Annual Conference · Market Decisions Research