

ANALYTICAL FINDINGS REPORT

From Manual Coding to AI Assistance:

Using Large Language Models to Classify and Summarize Judicial Performance Evaluation Comments

Colorado Judicial Performance Evaluation Program

Market Decisions Research | May 2026

MARKET DECISIONS RESEARCH

John M. Charles, MS
Research Director
jcharles@marketdecisions.com

Market Decisions Research
511 Congress Street, Suite 801
Portland, Maine 04101
207-767-6440

Executive Summary

This report presents the analytical findings from a proof-of-concept study evaluating the use of large language models (LLMs) to automate the processing of open-ended survey comments collected through the Colorado Judicial Performance Evaluation (JPE) program. The study applied a three-task AI pipeline to a dataset of **9,397 comments spanning 387 judges** across four court types. Tasks 1 and 2 automated attribution validation and constructiveness classification; Task 3 generated narrative summaries for judges with sufficient constructive feedback.

The pipeline successfully processed the full dataset. The large majority of comments, **93.7%**, passed both screening tasks and were eligible for summarization. Flags were generated for 592 comments (6.3%), all of which were routed to a human reviewer for adjudication. A stratified sample of 512 unflagged records was also reviewed to assess false negative rates.

Among the reviewed records, the AI demonstrated stronger performance on Task 2 (constructiveness classification) than on Task 1 (attribution validation). Precision was notably low across both tasks, driven primarily by over-sensitive spelling error detection in Task 1 and a tendency to conflate strong critical language with personal attacks in Task 2. These findings point to specific, addressable prompt refinements rather than fundamental limitations of the approach. Human reviewer thresholds may have also been more conservative than the AI's, which could mean the observed false positive rates are somewhat overstated.

The results support the conclusion that LLM-assisted processing is a viable and scalable approach for JPE comment workflows, provided that human oversight remains an integral component and that prompt logic is iteratively refined based on reviewer feedback.

1. Dataset Overview

The dataset contains **9,397 survey comments** submitted as part of the Colorado Judicial Performance Evaluation program. Comments were linked to **387 judges** representing District, County, Court of Appeals, and Supreme Court levels. Each record included judge-level metadata (full name, last name, gender, title, court type, and JBar identifier) alongside the comment text and a performance category label. The dataset was approximately gender-balanced, with 51.8% of comments associated with male judges and 48.2% with female judges.

Dimension	Count	Share of Total
Total comments	9,397	100.0%
Total unique judges	387	
AI-flagged (either task)	592	6.3%
Human-reviewed	1,104	11.7%
Constructive and eligible for summarization	8,805	93.7%

1.1 Comment Categories

Comments were organized into 11 performance categories at the time of data collection. Strength comments were the most common, representing 29.5% of the total, followed by Weakness comments (20.2%) and Final comments (15.3%). Domain-specific categories including Case Management, Application of Knowledge of Law, Demeanor, Communication, Fairness, and Diligence together accounted for 35.0% of the dataset. The non-constructive flag rate varied considerably across categories, with Final comments (7.0%) and Appearance categories (7.1% to 9.5%) showing the highest rates of flagged content, while Strength comments had the lowest at 1.0%.

Category	N	% of Total	Non-Constructive Flagged	NC Rate
Strength	2,771	29.5%	27	1.0%
Weakness	1,899	20.2%	62	3.3%
Final	1,434	15.3%	101	7.0%
Case Management	742	7.9%	15	2.0%
Application of Knowledge of Law	653	6.9%	12	1.8%
Demeanor	622	6.6%	26	4.2%
Communication	537	5.7%	10	1.9%
Fairness	414	4.4%	18	4.3%
Diligence	290	3.1%	4	1.4%
Appearance (General)	21	0.2%	2	9.5%
Appearance (Written)	14	0.1%	1	7.1%

1.2 Dataset Composition by Court Type

District Court comments represented the vast majority of the dataset at 73.4%, followed by County Court (24.8%). Court of Appeals and Supreme Court comments were comparatively small in volume. Flag rates were broadly consistent across court types, with slight elevation at the Court of Appeals and Supreme Court levels, likely reflecting the smaller sample sizes in those groups.

Court Type	N	% of Total	Wrong-Judge Flagged	WJ Rate	Non-Constructive Flagged	NC Rate
District	6,895	73.4%	261	3.8%	196	2.8%
County	2,327	24.8%	59	2.5%	77	3.3%
COA	142	1.5%	6	4.2%	3	2.1%
Justice	33	0.4%	2	6.1%	2	6.1%

2. Task 1: Attribution Validation

Task 1 evaluated each comment against judge-level metadata to detect potential misattribution. The agent checked for four types of attribution errors: spelling errors in the judge's name, name mismatches (where a different individual appeared to be referenced), gender pronoun mismatches, and title mismatches. Of 9,397 comments evaluated, **328 (3.5%)** were flagged as potential wrong-judge attributions and routed for human review.

2.1 Flag Distribution by Category

Spelling errors were the most frequently flagged type, representing 42.4% of wrong-judge flags, followed by name mismatches at 38.1% and gender mismatches at 19.5%. All 328 flagged comments were included in the human review process.

Wrong-Judge Category	N	% of Flagged	Human-Reviewed	Agreement Rate
Spelling Error	139	42.4%	139	5.0%
Name Mismatch	125	38.1%	125	16.0%
Gender Mismatch	64	19.5%	64	25.0%
Total Flagged	328	100.0%	328	13.1%

2.2 Human Review Results

The human reviewer confirmed 43 of 328 wrong-judge flags, yielding an overall agreement rate of **13.1%**. The remaining 86.9% of flags were overturned. Agreement rates differed substantially by subcategory, reflecting distinct failure patterns in each.

Spelling errors produced the lowest agreement rate at 5.0%. The dominant reviewer note, appearing in 121 cases, was simply that no spelling error existed. An additional 49 cases were described as misspellings that did not constitute true attribution errors. The AI appears to have flagged cases where a judge's name had textual similarity to common words or abbreviations in the comment, rather than detecting genuine name errors. This category requires the most substantial prompt revision.

Name mismatches yielded a 16.0% agreement rate. A recurring pattern involved comments that mentioned a second judge in passing, typically for comparison or context. Reviewers noted in 18 cases that the AI treated these references as attribution errors when they were not. The agent also had difficulty distinguishing possessive constructions and informal name usage from factual misidentification.

Gender mismatches had the highest agreement rate at 25.0%, but the majority of flags in this category were still overturned. Notably, a substantial share of the false positives traced back to errors in the judge roster metadata rather than in the comments themselves. Reviewers documented 34 cases where the source file recorded an incorrect gender for the judge, meaning the flag reflected a data quality issue rather than an AI error. Correcting the roster file before any future processing run would materially improve precision on this subcategory.

Prompt refinement note: *Spelling error detection requires the most targeted adjustment. Name proximity matching should require the matched string to function as a proper name in context, not simply resemble one. The gender mismatch logic would benefit from a pre-processing step that validates roster metadata before evaluation begins.*

3. Task 2: Constructiveness Classification

Task 2 applied a structured constructiveness rubric to each comment. A comment was classified as constructive only if it contained behavioral or performance-related content, whether positive or negative. Comments lacking such content were flagged across four categories: placeholder responses, overly general statements, irrelevant topics, and discriminatory or personal attacks. Of 9,397 comments evaluated, **278 (3.0%)** were flagged as non-constructive.

3.1 Flag Distribution by Category

Discriminatory or personal attack flags represented 63.3% of all non-constructive flags, making this the largest and most consequential subcategory. Overly general comments accounted for 28.1%, while placeholder responses (4.0%) and irrelevant topics (4.7%) together represented a small share of the total.

Non-Constructive Category	N	% of NC Flags	Human-Reviewed	Agreement Rate
Discriminatory or Personal Attacks	176	63.3%	176	15.3%
Overly General Positive or Negative	78	28.1%	78	57.7%
Placeholder	11	4.0%	11	100.0%
Irrelevant Topics	13	4.7%	13	23.1%
Total Flagged	278	100.0%	278	30.9%

3.2 Human Review Results

The human reviewer confirmed 86 of 278 non-constructive flags, for an overall agreement rate of **30.9%**. Performance varied widely by subcategory, ranging from perfect agreement on placeholder responses to 15.3% agreement on discriminatory or personal attack flags.

Placeholder responses achieved 100% agreement across all 11 cases. This is the most rule-governed category, and the AI's performance was commensurate with that clarity. Simple non-response text is reliably identified without ambiguity.

Overly general comments yielded 57.7% agreement, the second-strongest result. The majority of flags in this category were confirmed, though some edge cases were overturned where the reviewer judged a brief comment to contain sufficient implicit performance context. This category is functioning reasonably well and requires only modest prompt adjustment.

Irrelevant topics produced a 23.1% agreement rate. Several flagged comments were found to be irrelevant to the judge's performance category when read in full context, suggesting the AI was applying an overly narrow interpretation of what constitutes relevant feedback.

Discriminatory or personal attacks was the primary source of non-constructive false positives, with only 15.3% of flags confirmed. The most common reviewer note, appearing in 118 cases, described comments as containing negative language that did not cross the threshold into a personal attack because the criticism remained connected to observable judicial conduct. An additional 23 cases were flagged erroneously for comments that were, in fact, positive in tone. The AI is conflating strong critical language about judicial behavior with attacks on personal characteristics, which the prompt defines as the operative distinction.

Prompt refinement note: *The discriminatory/personal attack classification requires a more explicit threshold. The agent should flag only content where offensive language is directed at personal characteristics unrelated to judicial conduct, or where the comment is entirely disconnected from observable behavior. Strong negative criticism of a judge's courtroom conduct should not meet that threshold.*

4. Classification Accuracy

A total of **1,104 comments** were reviewed by a human analyst to assess AI accuracy. The review pool included all 592 AI-flagged records plus a stratified sample of 512 AI-unflagged records, the latter included as a false negative check. This mixed-design approach allows calculation of a full confusion matrix and supports estimation of false negative rates across the broader dataset.

4.1 Confusion Matrix

	Human: Flag (Yes)	Human: No Flag (No)	Row Total
AI: Flag (Yes)	125 (True Positive)	467 (False Positive)	592
AI: No Flag (No)	102 (False Negative)	410 (True Negative)	512
Column Total	227	877	1,104

Of the 512 AI-unflagged records reviewed, 102 (19.9%) were flagged by the human reviewer, indicating that approximately one in five comments the AI passed without a flag may warrant attention. Extrapolating this false negative rate across the full pool of 8,805 unflagged records suggests that roughly 1,760 additional comments could contain content a human reviewer would flag. This finding does not diminish the value of the AI screening layer, but it does reinforce the importance of maintaining human oversight beyond review of flagged records alone.

4.2 Performance Metrics

The metrics below are computed from the 1,104 reviewed records. Because the review pool oversampled AI-flagged cases, these figures best characterize performance on the records most likely to contain errors and should not be interpreted as representative of the full dataset.

Metric	Value	Interpretation
Precision	21.1%	Of all comments the AI flagged, 21% were confirmed as flags by the human reviewer
Recall	55.1%	Of all comments the human reviewer would flag, the AI identified 55%
F1 Score	30.5%	Combined measure of precision and recall; reflects the imbalance between the two
Specificity	46.8%	Of comments the human reviewer passed, the AI also passed 47%

Precision and recall are the most informative metrics here and they tell a consistent story. The AI's precision of 21.1% means that most flags it generates are overturned by a human reviewer. Its recall of 55.1% means it captures just over half of the comments a human reviewer would identify. Both figures point to a classification threshold that is too permissive, particularly in the wrong-judge detection task. The F1 score of 30.5%, which is the harmonic mean of precision and recall, reflects this imbalance. A score in this range is typical for a first-generation prototype applied to nuanced qualitative content, and it establishes a useful baseline against which future prompt iterations can be measured.

It is also worth noting that the overall false-positive burden is operationally manageable. The 592 flagged records represent only 6.3% of the full dataset, a volume that can be reviewed by a human analyst without exceptional time investment. The pipeline therefore reduces the review workload substantially even at its current accuracy level.

4.3 Agreement Rates by Flag Type

Flag Type	AI Flags (N)	Human Confirmed	Agreement Rate	Primary FP Driver
Wrong-Judge (all)	328	43	13.1%	Spelling error over-sensitivity
<i>Spelling Error</i>	139	7	5.0%	Name proximity false matches
<i>Name Mismatch</i>	125	20	16.0%	Multi-judge references in comments
<i>Gender Mismatch</i>	64	16	25.0%	Metadata errors in judge roster
Non-Constructive (all)	278	86	30.9%	Negative language over-flagging
<i>Placeholder</i>	11	11	100.0%	N/A
<i>Overly General</i>	78	45	57.7%	Edge cases with implicit context
<i>Irrelevant Topics</i>	13	3	23.1%	Context misread
<i>Discriminatory / Attack</i>	176	27	15.3%	Conduct criticism vs. personal attack

5. Summarization Eligibility and Yield

Comments that were not flagged by either Task 1 or Task 2 were classified as constructive and eligible for judge-level summarization in Task 3. A total of **8,805 comments (93.7%)** met this threshold, spanning **386 of the 387 judges** in the dataset. All but one judge had at least one constructive comment available for summary generation.

Funnel Stage	N	% of Total
All survey comments	9,397	100.0%
Pass attribution check (Task 1)	9,069	96.5%
Pass constructiveness check (Task 2)	8,805	93.7%
Eligible for summarization (Task 3 input)	8,805	93.7%
Judges with constructive comments	386	99.7% of 387 judges

5.1 Comment Volume per Judge

The median number of constructive comments per judge was **20**, with a mean of 22.8. The range was wide, from a minimum of 1 to a maximum of 115. Sixty-four judges (16.6%) had five or fewer constructive comments, a threshold below which summary depth may be limited. The Task 3 prompt instructs the agent to note when comment volume is insufficient to support an informed summary, and this provision applies to the lower end of the distribution.

Comments per Judge	Number of Judges	Cumulative % of Judges
1 to 5	64	16.6%
6 to 10	48	29.0%
11 to 25	132	63.1%
26 to 50	114	92.7%
51 to 100	26	99.5%
101 and above	2	100.0%

6. 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. The structural output was consistent across judges, with the two-paragraph format for strengths and weaknesses generally followed. However, some summaries described what respondents said rather than synthesizing the substantive content of what they reported, producing language that could apply to any judge rather than the individual being evaluated. This pattern is most common in categories with lower comment volumes, where the agent had less material to work with. Targeted prompt adjustments and a post-generation review step are recommended before summaries are shared with judges or made public.

7. Identifying Information

A separate review identified **34 comments (0.4% of the total)** containing identifying information. Of these, 28 were confirmed and 6 were marked as possibly identifying. The flagged content included case-specific details such as hearing dates, party names, and references to specific proceedings that could identify the respondent or the individuals involved in a case. All 34 comments require manual redaction before judge-level summaries are finalized or disseminated.

Given the false negative rate observed in Tasks 1 and 2, it is reasonable to expect that a small number of identifying comments were not caught by the automated process. A targeted manual review of a random sample from the constructive comment pool is recommended as a quality assurance step prior to publication or distribution of any summaries.

8. Key Findings

The following findings summarize the most important results from this evaluation cycle.

Scale and Coverage

The pipeline processed all 9,397 comments without error and produced structured, analysis-ready output for all 387 judges. The constructive yield of 93.7% confirms that the Colorado JPE program generates predominantly useful feedback, with a relatively small proportion of comments requiring removal before summarization.

Attribution Validation (Task 1)

Task 1 is the weaker of the two classification tasks. The 13.1% agreement rate on wrong-judge flags reflects systematic over-sensitivity, particularly in spelling error detection. The most actionable finding is that a significant share of gender mismatch false positives were caused by errors in the judge roster metadata rather than in the comment text itself. Correcting the source data and tightening name proximity thresholds in the prompt are the two highest-priority interventions for this task.

Constructiveness Classification (Task 2)

Task 2 performed considerably better overall, with a 30.9% agreement rate. Placeholder detection achieved perfect accuracy, and overly general comment classification was reliable at 57.7%. The primary weakness is the discriminatory or personal attack subcategory, where the AI consistently over-flagged comments that contained critical but conduct-related language. Introducing a clearer behavioral standard in the prompt for this subcategory would likely produce substantial improvement.

False Negatives

Among the 512 AI-unflagged records sampled for review, 102 (19.9%) were flagged by the human reviewer. This rate suggests that a meaningful volume of comments warranting attention are passing through the pipeline undetected. The AI catches a majority of true flags, but the false negative rate is high enough that a purely automated workflow would be insufficient for a program with the accountability requirements of judicial performance evaluation.

Human Oversight

The workflow's design, routing all AI-flagged records to a human reviewer, is well-calibrated for the current accuracy level. The 592-record flagged pool represents a manageable review workload and allows reviewers to focus their attention on the cases most likely to require judgment. As prompt precision improves in future iterations, this workload will decrease further.

9. Limitations

Several methodological considerations should be kept in mind when interpreting the findings in this report.

Review Sample Design

The 1,104-record reviewed pool was not a simple random sample of all 9,397 comments. It oversampled AI-flagged records by design, which means the precision and recall metrics reported here apply to the reviewed subset and cannot be directly generalized to the full dataset. The figures are best understood as characterizing performance on the cases most likely to contain errors.

Reviewer Threshold Variability

Human reviewer judgments may reflect a more conservative threshold for what constitutes a flag than the AI's classification logic, particularly in the discriminatory or personal attack category. If the reviewer and the agent were applying different implicit standards for the constructiveness rubric, the AI's apparent false positive rate may be somewhat overstated. Establishing explicit reviewer calibration criteria would reduce this source of variability in future evaluations.

Single-Cycle Evaluation

The results presented here reflect one processing run against one dataset. Precision and recall figures should be treated as a performance baseline for this iteration of the prompt rather than as stable benchmarks for the approach as a whole. Iterative prompt refinement followed by re-evaluation against the same or a comparable dataset is the appropriate next step.

Roster Metadata Quality

At least 37 false positive gender mismatch flags were traceable to incorrect gender entries in the judge roster file. Downstream classification accuracy is directly dependent on the accuracy of the metadata provided to the agent. Verification of roster data prior to each processing run should be treated as a required preparatory step.

Category Label Inconsistency

The Category field in the source data contained two variant spellings of the Weakness comment label. These were consolidated for purposes of this analysis but should be standardized in future data exports to ensure consistent downstream processing.

Identifying Information Coverage

The identifying information flag captured 34 comments at a rate of 0.4%, which is likely an undercount given the false negative rates observed in the primary classification tasks. Manual spot-checking of a random sample from the eligible comment pool is recommended before any summaries are distributed externally.

10. Recommendations

Based on the findings presented in this report, the following steps are recommended for the next iteration of this work.

Revise the spelling error detection logic in Task 1. Name proximity matching should require the candidate string to function as a proper name within the sentence, not simply share characters with the judge's name. This single adjustment is likely to have the largest impact on overall precision.

Validate the judge roster metadata. Gender entries should be verified against an authoritative source before the next processing run. This will eliminate a known and correctable source of false positives in gender mismatch detection.

Refine the discriminatory or personal attack threshold in Task 2. The prompt should specify that flagging requires offensive language directed at personal characteristics unrelated to judicial conduct. Strong negative assessments of a judge's courtroom behavior should be treated as constructive regardless of tone.

Establish reviewer calibration criteria. A shared decision guide for human reviewers, aligned with the AI prompt rubric, would reduce variability in adjudication and produce more comparable agreement metrics across review cycles.

Conduct a post-generation review of Task 3 summaries. A structured quality check against the prompt's specificity and behavioral-content requirements should be applied to a sample of generated summaries before they are shared with judges or published. Generic and meta-descriptive language should be flagged for regeneration.

Standardize category labels in source data exports. The Weakness comment label variant should be corrected at the source to ensure consistent processing in future cycles.