

A Dataset of Motion-Response Text Pairs from US District Court Cases for Legal Judgment Prediction

Damian Curran¹, Lea Frermann², Eduard Hovy¹

¹The University of Melbourne, Australia ²Universität Tübingen, Germany
dcurran@student.unimelb.edu.au

Abstract

Many existing Legal Judgment Prediction (LJP) datasets are based on test conditions which do not resemble real-world contexts, in that both the input to the task and the predicted case outcome labels are drawn from the same document. In other words, the outcome was known at the time the input was written. By utilizing the chronological event logs of court cases in US district courts, called dockets, we create a 1.6k instance outcome prediction dataset called MOP-2026 where the text inputs predate the outcome. We describe the collection, filtering and labelling of the pairs of opposing argument text and define a new subtask of LJP called motion-outcome prediction with a ternary label set. We provide baseline results and measure the performance of a variety of LLMs on the task. While exceeding naive baselines, even the latest LLMs perform below most reported metrics in LJP, supporting recent commentary that valid test conditions greatly impact predictive performance, and that many earlier LJP systems appear to overstate predictive capability in real world settings. We close with a discussion of the inherent challenges of LJP.

1 Introduction

The tasks in a number of commonly used Legal Judgment Prediction (**LJP**) benchmarks do not utilise inputs that would be available to a real user at the time of need. Rather than using input text that predates the judgment, these benchmarks are often constructed by taking both the “facts” of a case, and the predicted label, from the same judgment text. However, no real-world predictive system will ever have those “facts” inputs available because once the judgment text is published there is no longer any need for a prediction.

The construction of LJP benchmarks in this way is understandable when in many jurisdictions, judgments are the only substantive case texts available publicly and in bulk. Nonetheless, this construction

method raises questions about whether predictive performance on such benchmarks generalises to real-world settings — sometimes referred to as questions of ‘external’ (Salaudeen et al., 2025) or ‘ecological’ (Bronfenbrenner, 1977) validity.

A survey by Medvedeva and McBride (2023) found that only 7% of the 150 LJP papers surveyed use data as input to the prediction tasks which predate the judgments from which the labels were extracted, suggesting that there is a strong need for LJP datasets which more closely resemble the real-world data constraints of legal practitioners.

Our main contribution is the release of a new dataset for LJP which addresses these validity concerns. We introduce **MOP-2026**, the Motion-Outcome-Prediction dataset.¹ MOP-2026 is comprised of motion and response pairs in contract and property rights law cases in US district courts, labelled with the court’s ruling on the motion. Both the motion and response texts used as input to the predictive task predate the ruling from which the label is extracted. An example is shown in Figure 1. We describe the collection, filtering and labelling of this data. We introduce a new prediction task, provide some baseline results and measure performance of various LLMs on the task.

MOP-2026 is an important addition to the limited pool of LJP datasets in which the inputs predate the outcome label. It also provides, for the first time, full text pairs of opposing legal argumentation, permitting examination of a predictive system’s ability to reach reasoned predictions in the face of competing interpretations of the same underlying facts. Beyond the dataset and quantitative analysis of model performance, we provide a qualitative assessment of the LLM chain of thought reasoning. We use that analysis to advance discussion on the general utility of LJP and in particular

¹A script to reconstruct the dataset can be accessed here: <https://github.com/dzcurran/MOP-2026>

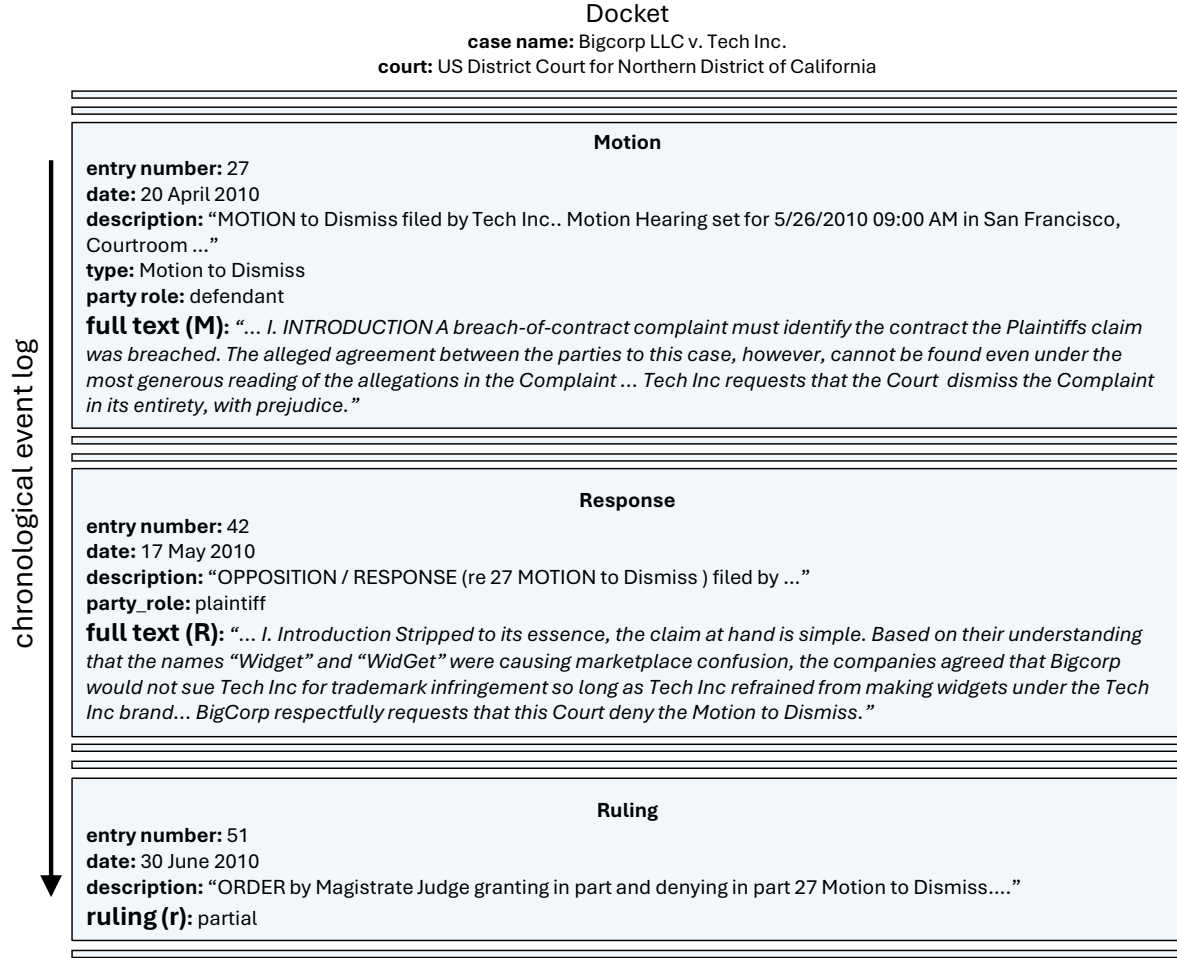


Figure 1: A fictionalized docket example, illustrating the docket structure and the variables in the motion-outcome prediction task M , R and r .

whether there are certain ‘hard’ cases in which the LJP paradigm of predicting a single outcome label is an appropriate goal for legal NLP at all.

2 Related Work

2.1 LJP and its critics

Legal Judgment Prediction, a subset of legal NLP, is a range of subtasks involving the prediction of a case outcome label based on part of the case input (Feng et al., 2022). Benchmark datasets include the widely used corpora from the European Court of Human Rights (ECHR) (Chalkidis et al., 2019, 2021; Kaur and Bozic, 2019; Xu et al., 2023), and other case data drawn from courts in China (Luo et al., 2017; Wang et al., 2019; Xiao et al., 2018), India (Malik et al., 2021; Nigam et al., 2025; Vats et al., 2023), Switzerland (Niklaus et al., 2021) and the United States (Katz et al., 2017). LJP subtasks include binary or multi-label classification of an input text into a range of label sets, including violated

law article, criminal charge or case outcome. Early work used feature-based machine learning models (such as Katz et al., 2017). More recent work has explored using supervised fine-tuned neural models and LLMs with retrieval and a range of prompting techniques (including Nigam et al., 2025; Niklaus et al., 2021; Shui et al., 2023). Performance is typically reported by accuracy and macro F1 score. Results vary across corpora and tasks. Some reported scores have suggested that the systems are highly capable and that the task is near-solved, including 96% reported accuracy on outcome prediction task on ECHR data (Quemy and Wrembel, 2022), 94.9% accuracy on charge prediction task on the CAIL2018 data (Wu et al., 2023) and 98% macro f1 score on ruling prediction in French Supreme Court cases by (Sulea et al., 2017). Other results are more sanguine yet nonetheless suggest strong predictive capability of the systems, such as 78% accuracy in outcome prediction of Indian Supreme

Court cases (Malik et al., 2021) and 70% macro-f1 on Swiss judgment prediction (Niklaus et al., 2021).

Recent commentaries have critiqued some LJP dataset construction methodologies. They observe that much of the early work in legal judgment prediction relied on judgments, issued at the end of a legal case, to build tools designed to provide utility at an earlier stage in the case. Many tasks extract, for instance, the ‘facts’ section of a judgment and use that as input to a system which predicts the outcome later in that same document. However, those ‘facts’ sections may be “highly selective summaries of the original case record, written by the decision-makers themselves and tailored for consistency with the decision” (Tippett et al., 2021). Medvedeva et al. (2021) showed that LJP models that use case summaries that predate the judgment from the ECHR performed much worse than when modelled on the inputs from the judgment text itself (with F1 score reduction from 0.92 to 0.64). In other words, the predictive task became more difficult when using input data that predates the judgment.

In a survey of the field, Medvedeva and McBride (2023) found that only 7% of the 150 LJP papers surveyed use data sourced from a document produced earlier than the judgment it seeks to predict. These works explore a range of creative data sources. For instance, Semo et al. (2022) uses the text of complaints to predict the outcome of class action cases in the US. However, the label set combined several case outcomes which may be unrelated to the substantive merits of the claim, such as settlements and dismissals, into a binary label scheme. McConnell et al. (2021) also used US court complaint texts to predict subsequent case outcomes, namely the outcome of subsequent motions to strike. It also used a binary label. Both works involved the use of case materials filed by one party only (the plaintiff) and so were unable to explore a system’s predictive or reasoning capability when presented with competing interpretations of the same underlying facts. The works also suggest that when using input text that predates the judgment, LJP remains a challenging, open question, with their best performing approaches achieving accuracies of 63% and 64% respectively on the binary label sets. Other LJP works identified in the survey used oral arguments (Kaufman et al., 2017) and the full text of lower court decisions on appeal (Jacob De Menezes-Neto and Clementino, 2022)

as input. However, oral argument transcripts are not widely available and lower court decisions are only useful inputs to predicting appeals and hence of little interest to the vast majority of litigants at first instance.

These validity challenges are not unique to legal NLP. Ramadan et al. (2025), for example, identified structural flaws in machine learning experiments which sought to predict admission outcomes from medical diagnostic codes, where both the codes and the patient outcomes were extracted from patient discharge summaries produced only at the end of the patient’s hospital visit. Their critique highlighted inflated reported performance of the prediction models. They framed the problem as one of “data leakage, in which information unavailable in deployed clinical settings is improperly used during model training and evaluation.”

2.2 LLMs and prompting strategies in LJP

Several works have explored prompting strategies for LLMs on LJP tasks. Deng et al. (2024) used a guided reasoning framework in prompts on the CAIL LJP dataset task, finding superior performance. Zhang et al. (2025) explored trichotomous reasoning in LLM prompts on the CAIL task, notably enhancing accuracy. Xie et al. (2024) explored the use of GPT-4 on UK employment law LJP task, finding that fine-tuned models outperformed LLMs but that performance could be improved by integrating task-related information into the prompt. Prompting the model to solve a problem using the IRAC method of legal reasoning (which involves reasoning over a problem by identifying the applicable Issue, Rule, Application and Conclusion) has also been used in various legal NLP prompting task to improve performance (Kang et al., 2023; Yu et al., 2022).

2.3 Chain of thought

An LLM’s ‘Chain of Thought’ (CoT) are the intermediate tokens, or reasoning steps, generated by the model prior to reaching a final answer (Wei et al., 2022). Many modern LLMs default to CoT prior to providing an answer, including the gemini family of models used in this study. This is commonly referred to as the ‘thinking’ mode on the commercial LLM products. While the CoT is not always a faithful record a model’s underlying reasoning (Turpin et al., 2023), it can nonetheless be an informative signal for other ends, such as agent monitoring (Korbak et al., 2025).

2.4 NLP analysis of US federal court system texts

Several large datasets of bulk US case law material have been made available to the NLP community. The Pile of Law dataset includes judicial opinions as well as briefs, exhibits and other case records (Henderson et al., 2022). A smaller number have been structured into NLP tasksets. Most prominent are the various datasets compiled from US Supreme Court proceedings (Bergam et al., 2022; Fang et al., 2023; Katz et al., 2017). As noted above, Semo et al. (2022) and McConnell et al. (2021) use US court complaint texts to predict case outcomes. However, given the vast size of the US court systems and the relative availability of case texts, there appears to be significant opportunities for further exploration of these texts by the legal NLP community.

3 The US federal court system, dockets and motions

3.1 The docketing system

The documents in our dataset are from the US federal court system. The system includes 94 first instance district courts, 13 circuit courts of appeal and the US Supreme Court. It has a broad subject matter jurisdiction from bankruptcy and intellectual property to constitutional law and federal crimes. Close to 1 million new cases are initiated across district, bankruptcy and appeal courts each year.

Cases in the system are managed on a record called a **docket**. A docket is a chronological event log. Each case has its own docket. The docket itself contains metadata relating to the case as a whole, including the party names and court details. The docket’s event log is comprised of a list of **docket entry** items. There is one docket entry per legal event, such as the filing of a document by a party, or the making of an order by the court. A docket entry includes metadata such as a docket entry number (a unique id), date of the docket entry, and a description of the legal event. This description must “briefly show the nature of the paper filed” or the “substance” of each order and judgment (per Fed. R. Civ. P. 79(a)(3)). Each docket entry may also link to one or more full text documents, such as a party’s motion or response text.

Relevantly to the question of benchmark validity, docket entries and documents are typically not retrospectively edited. We are therefore able to utilise data from docket entries before a given date

to safely predict an event on a future docket entry.

3.2 Motions, Responses and Rulings

Various types of legal events are recorded on docket entries. These include the filing of a **motion** by a party. Motions are applications made by a party during a court proceeding. The party bringing the motion (*the moving party*) is asking the court to make a specific order (*the relief sought*). The relief sought by the moving party ranges from procedural requests (such as a change to the administrative timetable) through to substantive relief (such as judgment on the entire claim). A motion document filed by the moving party and annexed to the relevant docket entry typically makes a complete, atomic argument in favour of the granting of the motion, including a recitation of the relevant background facts, the relevant law and a plea for the specific relief sought.

Motions may be opposed by another party in the case. If so, the other party will file a **response** to the motion. The response will appear as a separate docket entry and typically be supported by another full text document. However, unlike the motion document, the response-in-opposition document sets out an argument *against* the granting of the motion.

The court will then, subsequently, make a **ruling** on the motion, typically either granting, denying, or granting-in-part the motion. The ruling may also be accompanied by a full text document setting out the court’s reasons for the ruling. There may be multiple motions filed on the one docket at different times by different parties. In a small number of instances there are different responses filed in respect to the same motion, such as where two different defendants separately oppose a plaintiff’s motion for their own reasons.

4 Dataset construction

MOP-2026 comprises motion and response-in-opposition pairs drawn from contract law and property rights cases in US district courts. The texts are sourced from the RECAP archive by the nonprofit Free Law Project,² a mirror database of the US Public Access to Court Electronic Records (PACER) system, a searchable archive providing public access to US federal court records.³

We built a multi-step pipeline to locate the texts.

²<https://www.courtlistener.com/recap/>

³<https://pacer.uscourts.gov/>

We began by filtering dockets by court and dispute type. We selected contract law and property rights disputes in the district courts. These are first instance civil disputes covering a diverse range of scenarios and parties including private and public corporations, individuals and pro se litigants. We link motion, response and ruling docket entries on those dockets using regular expressions. The example in Figure 1 illustrates how the docket entry descriptions contain numbers which correspond to earlier docket entries, permitting this linking. (This linking was validated with an LLM filter, below.) We then obtain the full text documents for those docket entries, where available, from the RECAP archive. These full text files are OCR versions and may include artefacts such as headers and footers, cover pages and noise. OCR quality varies and is not post-processed. We removed all instances where either a motion or response exceeded 20,000 words.

While that process reduced the universe of candidate instances to many thousands of pairs, a majority of the remaining texts were unsuitable. Given the scale, we developed a series of LLM-based classifiers for further fine-grained filtering. These classifiers identified any instances where the entry was not a motion or a response (i.e. the docket entry type had been misidentified above), where the full text motion or response did not contain the legal submissions of the filing party (such as where the full text merely refers to an ‘attached’ brief which was not itself included in the text), where the OCR scan was illegible or incomplete, and where the response was not in opposition to the motion (as some responses were in support of the motion). This filtering step also filtered out instances where the regex linking of docket entries, above, had misattributed a link across motion, response or ruling. To ensure the quality of this step, we manually iterated the LLM classifier prompts on small samples before scaling to all candidates. We manually reviewed every remaining full text pair to ensure quality of the final set.

We manually annotated the remaining motion and response pairs based on the text of the docket entry description for the associated ruling (as in the ‘Ruling’ docket entry in Figure 1). These docket entry descriptions are typically written by court staff. They are not necessarily the verbatim text from the judge’s order on the motion. They are summaries. High confidence in the mapping of the docket descriptions to the actual court order is warranted

given the Federal Rules of Civil Procedure (the applicable law) governing the maintenance of the civil docket states that the court clerk must show the substance of each order and judgment in the docket entry.⁴ Our checks of the full text orders, where available, suggest the docket entry description text closely maps to either the introductory or concluding section of a full text order.

We used the following ternary label set to annotate the ruling docket entry descriptions: **granted** (the motion was granted), **denied** (the motion was denied) and **partial** (being any combination of granted in part and denied in part). These ternary labels were consistent with other LJP tasks (such as Nigam et al. (2025)). They mapped cleanly to the typical docket entry descriptions for the motion rulings. An example annotation for each label is shown in Table 1. The manual annotations were cross-checked against the outputs of an LLM annotation and agreement exceeded 99%.

The ‘partial’ label would typically apply if the motion sought multiple items of relief, but not all were granted. For example, a motion may seek to have two claims dismissed, but the court ruling only dismissed one. This label inevitably collapses the complexity of such ruling and hides meaningful information about *which* relief sought in the motion, among several, was granted. A more complex annotation scheme would have instead individually articulated the relief sought in the motion and annotate the ruling on that relief individually. Unfortunately, establishing such a complex annotation scheme was beyond the scope of this work.

We had also annotated a *not-decided* label covering instances, for example, where a motion was withdrawn or became unnecessary to decide on its merits (often described on the docket entry as ‘denied as moot’). We removed those instances from the final dataset because that outcome did not involve an assessment by the court of the substantive legal merits of the motion, but rather extraneous factors such as the parties settling, or the outcome of a parallel motion by another party. Prediction of that label would therefore not be a reflection of the legal reasoning capability of the system.

Finally, we add some further annotations to each instance, such as motion type and the role of the authoring party of each document (either ‘plaintiff’, ‘defendant’ or ‘other’), to allow us to explore some feature-based classifiers.

⁴Fed. R. Civ. P. 79(a)

Docket entry description	ruling
"ORDER by Judge granting 10 Motion to Dismiss for Lack of Jurisdiction."	granted
"ORDER GRANTING IN PART AND DENYING IN PART 13 Defendants' Motion to Dismiss Counts II, IV, and V of Plaintiff's Complaint."	partial
"OPINION and ORDER denying 4 Motion to Dismiss for Failure to State a Claim."	denied

Table 1: Sample docket entry descriptions from the Ruling docket entry, and the associated annotation label.

Label	Count	%
denied	777	47.7
granted	524	32.1
partial	330	20.2
Total	1631	100.0

Table 2: Distribution of ruling labels (r).

5 Dataset statistics

The final dataset contains 1631 pairs of full text motions and responses in opposition, each labelled with the ruling outcome of granted, denied or partial. There are 1597 unique motions drawn from 1007 separate dockets. There are slightly fewer motions than pairs because in some multi-party cases, parties will prepare their responses independently, each making separate arguments in opposition to the motion. We retain those motion-response pairs as separate instances. The pairs are roughly equally split between contract law and property rights cases (at 803 and 828 pairs, respectively). The motions date between 2000 and 2020.

The motions are typically described on the docket entry as "Motion for [type]", which permitted us to classify them into types. The largest subgroup was Motion to Dismiss, which are requests typically brought by a defendant, asking the court

	Min	Max	Mean	Median
M-words	160	19713	3934	2811
M-pages	1	92	13	10
R-words	158	19963	4349	3351
R-pages	2	87	14	12

Table 3: Word and page count statistics for Motion (M) and Response (R).

Motion type	Count
Motion to Dismiss	323
Motion for Summary Judgment	185
Motion to Compel	136
Motion to Strike	80
Motion in Limine	71
Motion for Extension of Time	68
Motion for Leave to File	64
Motion for Protective Order	60
Motion for Reconsideration	55
Other Motion Types (Various)	589

Table 4: Distribution of motion types.

to throw out (or 'dismiss') the plaintiff's claims because they fail to state a valid legal basis. The judge in those motions will need to evaluate whether the plaintiff makes a claim with a valid legal basis. The second largest subgroup was Motion for Summary Judgment, in which a party claims there is no genuine dispute on the facts of the case, and the matter can be resolved without a trial. The judge in those motions will need to evaluate whether the evidence is one-sided enough not to warrant a trial. There is a long tail of other motion types, some of which are shown in Table 4.

6 Task

We illustrate a use of the data with a task called motion-outcome prediction. Formally, given a pair of the full text of a motion (M) and the full text of a response in opposition to that motion (R) for a US district court case, predict the ruling of that motion as indicated in the subsequent docket entry, r (where $r \in [\text{'granted'}, \text{'denied'}, \text{'partial'}]$):

$$r = \text{CLS}(M, R)$$

The task maps directly to the real-world requirement for attorneys to advise clients on the prospects of success of a particular motion. More abstractly, the task is a proxy for the key legal skill of reaching reasoned legal conclusions in the face of competing arguments and diverging interpretations of the same underlying facts.

7 Experiments

We establish the following naïve baselines: random choice, majority-class and majority-class-by-motion-type. We also explore whether simple document features are predictive. It could plausibly be that a court grants motions of a particular type

when filed by a particular party (for example, a Motion for Extension of Time filed by a plaintiff), merely as a matter of course and irrespective of the merits of the response in opposition. Or, a court could plausibly be swayed for or against a motion merely by the length of the submissions. To test the impact of such features, we encode the vector using one-hot encoding with 3 dimensions per text for page count, 3 dimensions per text for the party roles and 10 dimensions for the motion type, fit a support vector machine classifier and evaluate it using 4-fold cross validation.

We were most interested in the ability of leading-edge LLMs to perform the task. We wrote two zero-shot prompts guided by the phrasing in related LJP papers. Each prompt provided the LLM with the full text of both the motion and the response and asked it to make a prediction on the outcome of the motion. The first prompt (*‘simple’*) was a simple instruction asking the model to make the prediction and provided some basic context for the texts. It was similar in style to the zero-shot prompt used by Xie et al. (2024) in the UK CLC outcome prediction task. The second prompt (*‘IRAC’*) built on the first prompt but further instructed the model to reason using the IRAC method (Kang et al., 2023; Guha et al., 2024). We were curious whether explicitly asking the model to follow those reasoning steps would improve performance. See Appendix A for the full text of both prompts. We opted not to experiment with few-shot prompting because some of the input texts had tens of thousands of tokens and, given the variety of motion types, it was not clear which exemplar instances would have been appropriate for inclusion in a few-shot prompt.

We run the prompts on a range of open- and closed- weight LLMs of different sizes, all with reasoning capability, namely: qwen3.6-35b-a3b, deepseek-v4, claude-sonnet-5 and gemini-3.1-pro-preview, all in ‘high’ thinking mode. We record the predicted label and the model’s CoT, where available, across both prompts. We report on accuracy and macro-F1.

We performed a preliminary qualitative examination of the LLMs CoT. We selected a small sample of CoTs from instances where the gemini model misclassified on the simple prompt, across a mixture of motion types. We manually compared the CoT to the court’s reasoning. We reflected on its hallucinations, confidence and its mapping to the court’s reasoning.

Method	Acc.	M-F1
maj-overall	0.48	0.22
random	0.34	0.33
maj-by-type	0.50	0.34
SVM	0.41	0.41
qwen-3.6-35b (irac)	0.55	0.47
qwen-3.6-35b (simple)	0.55	0.48
deepseek-v4 (simple)	0.55	0.50
claude-sonnet-5 (simple)	0.56	0.51
deepseek-v4 (irac)	0.56	0.51
claude-sonnet-5 (irac)	0.57	0.52
gemini-3.1-pro (irac)	0.61	0.56
gemini-3.1-pro (simple)	0.62	0.58

Table 5: Model performance ranked by macro-F1 score.

8 Empirical results

The accuracy and macro-F1 score of our baseline and LLM approaches are shown in Table 5. Models ranked as expected, with naïve baselines with lowest performance, then the SVM classifier, the open-weight LLMs, the mid-size closed LLMs and topped by the latest, largest closed-weight LLM in the group. There was no notable pattern of the preferable prompt across the models. The simple prompt on gemini performed best, with a macro-F1 score of 0.58, marginally out-performing the IRAC reasoning prompt on the same model.

The only motion type with notably higher F1-score with the gemini model (simple prompt) was the *Motion for Leave to File* (with an F1 of 0.74 in that category), perhaps reflecting the model’s understanding that such requests are relatively minor procedural steps and are often granted. The models performed slight better in the contract law domain than the property rights domain, with average macro-F1 scores across all models of 0.46 and 0.45 respectively. The confusion matrix of the simple prompt method on the gemini model is shown in Figure 2.

The per-class evaluation metrics are shown in Table 6 for the highest performing prompts on gemini-3.1-pro-preview, claude-sonnet-5, and deepseek-v4. They each perform relatively well for the ‘denied’ label with F1 scores of 0.70. This suggests that motions which are going to be denied are easiest to identify. The other models degrade, however, by hedging away from ‘granted’ towards the ‘partial’ label more often than gemini. This tendency might plausibly be addressed with prompt engineering, such as including examples

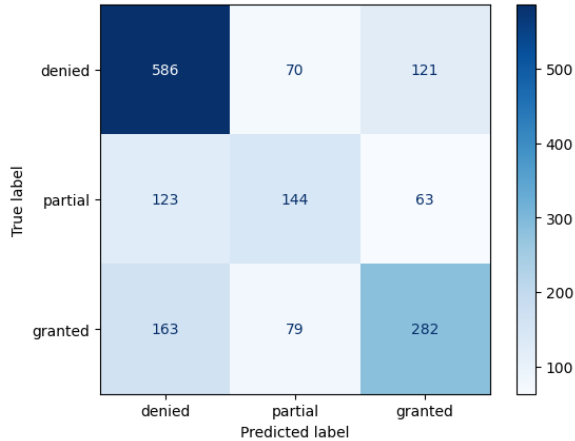


Figure 2: Confusion matrix from the gemini model using the simple prompt.

Label	P	R	F1	prd.	true
<i>gemini-3.1-pro (simple)</i>					
denied	0.67	0.75	0.71	872	777
partial	0.49	0.44	0.46	293	330
granted	0.61	0.54	0.57	466	524
<i>claude-sonnet-5 (irac)</i>					
denied	0.66	0.73	0.69	851	777
partial	0.38	0.65	0.48	573	330
granted	0.70	0.28	0.40	207	524
<i>deepseek-v4 (irac)</i>					
denied	0.63	0.77	0.69	942	777
partial	0.38	0.54	0.44	470	330
granted	0.67	0.28	0.40	219	524

Table 6: Per-class precision (P), recall (R), F1, number of predicted instances (prd.) and true instances (true) for highest performing prompts across three model families.

of the types of decisions that should be given a ‘partial’ label or an explicit instruction to defer to a ‘granted’ label if unsure of the outcome. However, extensive prompt engineering of this nature was beyond the scope of this work.

9 Qualitative observations

The model CoT referred to conducting searches of its knowledge base when such functionality was not available to it. On several occasions it was confidently incorrect about the outcome of the motion. However, in other instances its reasoning mapped onto the court’s own opinion in that it identified the same relevant issues, rules and key legal test as the court. It failed merely because the motion outcome turned on a discretionary judgment made by the court and it happened to fall on the wrong

side of that discretionary judgment. For example, one instance involved a plaintiff’s motion to file an amended complaint. Part of the relevant legal test for such a motion was whether there had been ‘undue’ delay by the plaintiff. The notion of whether delay is ‘undue’ is discretionary (or ‘open-textured’), distinct from a rule which might, for instance, specify a hard deadline after an exact number of days have elapsed following a particular event. Given the LLM was not given a fourth label it could apply to such indeterminate or discretionary cases, it perhaps should not be faulted for incorrect predictions in those instances.

10 Discussion

Our analysis reinforces prior work indicating that LJP using text that pre-dates the judgment is challenging. Leading edge LLMs across a range of prompting approaches exceeded naïve baselines but fell short of reported F1 scores of other LJP systems and far short of what would be demanded in a real-world legal deployment. This is so despite, on one view, this task being easier than other LJP tasks, because it merely requires the selection of the *preferable* argument from the motion-response pair rather than reasoning, from scratch, over the bare ‘facts’ of a case through to a legal outcome. And yet the task remains challenging and performance metrics are lower than on many of those other benchmarks.

Nonetheless, direct comparison across datasets from different jurisdictions and dataset construction, such as between the ECHR and US District Courts, is not straightforward. The underlying tasks involve different legal knowledge bases. The ternary label set in this work is also inherently more challenging than a binary label set (as was used in some of the ECHR dataset tasks) and will itself explain some of the performance gap. In order to fully explore the challenges of LJP using texts which predate the judgment, we would need pre-judgment and judgment input texts from the same cases to be observable simultaneously. That may be possible with this US court data. We hope to explore this in future work.

Our analysis also suggests that simple label prediction may not be an appropriate optimisation target for some instances *at all*. Our examination of the CoTs identified instances where the decision turned on a discretionary choice of the judge. While we are not in a position to conclusively say

whether a more advanced reasoning engine or human specialist may have correctly made that judgment in the particular examples we reviewed, legal theorists suggest that there may always be a subset of cases for which a correct prediction can never be made in advance. These works include those characterising legal reasoning as defeasible rather than monotonic (Carlos, 2001), the exploration of the open-texture of legal concepts such as what is ‘reasonable’, ‘fair’ or ‘sufficient’, and the existence of ‘hard cases’ over whose resolution even experts disagree (Rissland and Skalak, 1991). Such cases suggest that perhaps the entire framework of LJP — namely the classification of legal outcomes into a small set of class labels — is not an appropriate or helpful exercise. In that vein, Santosh et al. (2024) recommended moving away from statistically classifying legal texts into ‘impoverished’ case outcome labels and towards the more traditional symbolic and argumentative analysis form of legal NLP. They suggest that greater effort should be made by the legal NLP community to provide useable justifications for predictions, connected to the appropriate legal concepts.

Nevertheless, the MOP-2026 dataset remains a valuable contribution to the field. It adds much needed volume to a limited pool of LJP datasets in which the inputs predate the outcome label. It provides, for the first time, full text pairs of opposing legal argumentation with an authoritative, adjudicated label, permitting examination of a predictive systems’ ability to reach reasoned predictions in the face of competing interpretations of the same underlying scenario. The wide performance gaps across a range of tested models suggests that the task is well gradated and improvements in performance on the task should reflect a genuine improvement in a system’s abilities. As seen above, it is also a useful vehicle to facilitate discussion on the utility and challenges of LJP.

11 Conclusion

We contribute MOP-2026, a dataset of over 1.6k motion and response pairs in contract law and property rights cases in US district courts, labelled with the court’s ruling on the motion. Both the motion and response texts used as input to the predictive task predate the ruling from which the label is extracted, marking an important contribution to the limited pool of LJP datasets with this construction. It also provides, for the first time, full text pairs of

opposing legal argumentation, permitting examination of a predictive system’s ability to reach reasoned predictions in the face of competing interpretations of the same underlying facts. We describe the collection, filtering and labelling of text and define a new subtask of LJP called motion-outcome prediction with a ternary label set. We provide baseline results and measure the performance of a variety of LLMs on the task. We illustrate that while exceeding naive baselines, even the latest LLMs perform below most reported metrics in LJP, confirming recent commentary that test conditions greatly impact predictive performance, and that many earlier LJP systems appear to overstate predictive capability in real world settings. We also provide a qualitative assessment of the LLM failure modes and use that analysis to advance discussion on the general utility of LJP, in particular whether there are certain ‘hard’ cases in which the LJP paradigm of predicting a single outcome label is an appropriate goal for legal NLP at all.

12 Limitations

Our preliminary analysis did not include baseline comparison to human experts in the motion outcome prediction task. This limits the extent to which we can make meaningful comments on whether the task is achievable at the limit.

The prediction task is limited to two document inputs, the motion and response. In reality, the human decision maker has access to significantly greater context, including evidence, exhibits and other documents on the docket. Some of the motion types, such as motions for summary judgment, turn on an examination of documentary evidence, which in this dataset was not provided to the classifier system (outside of text excerpts that may be cited in the motion or response) but may otherwise be available on the docket. Future work may explore whether the classification performance of those systems improves when given that additional context as input.

Not all motions that are filed in the court will end up having a clear ‘granted’, ‘denied’ or ‘partial’ outcome. Sometimes a moving party will voluntarily withdraw a motion. Sometimes a motion is unopposed. Other times the entire case is settled out of court before a motion is heard by the judge and ruling on the motion becomes unnecessary. By excluding motion and response pairs from the dataset that do not have a clear ruling label, we

introduce a selection bias into the model inputs. In effect this shifts the question from "how will this arbitrary motion be ruled upon?", to "*given that this motion was ruled upon*, what was the outcome?". This challenge of selection bias, of course, afflicts existing LJP case outcome prediction tasks in which the datasets consist entirely of cases which have been ultimately determined. But it is not unique to LJP. It has long been recognised by legal academic scholars seeking to draw generalisable lessons from reported judgments that cases which are litigated are neither a random selection nor a representative sample of all disputes (Priest and Klein, 1984). However, what the docket structure of the US federal court system may uniquely allow us to do in future works is to identify motions which were *not* ruled upon and to extend the label set and data accordingly. That option is not open in jurisdictions where the only case text available is a final judgment.

13 Ethics

The public policy interest of making the law broadly available may occasionally conflict with personal privacy interests. Accordingly, at this time we do not make the full text documents in the dataset directly available for download. Full text documents referenced in our dataset are public documents freely available on the website of the Free Law Project, a US not-for-profit organization. We provide a list of document ids and a script to access the documents using their free API service. All documents in the dataset are available at those end points at the time of writing.

References

- Noah Bergam, Emily Allaway, and Kathleen Mckeown. 2022. [Legal and Political Stance Detection of SCOTUS Language](#). In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 265–275, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Urie Bronfenbrenner. 1977. Toward an experimental ecology of human development. *American psychologist*, 32(7):513. Publisher: American Psychological Association.
- Juan Bayón Carlos. 2001. Why is legal reasoning defensible? In *Pluralism and law*, pages 327–346. Springer.
- Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. [Neural Legal Judgment Prediction in English](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.
- Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. [Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.
- Chenlong Deng, Kelong Mao, Yuyao Zhang, and Zhicheng Dou. 2024. [Enabling Discriminative Reasoning in LLMs for Legal Judgment Prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 784–796, Miami, Florida, USA. Association for Computational Linguistics.
- Biaoyan Fang, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2023. [Super-SCOTUS: A multi-sourced dataset for the Supreme Court of the US](#). In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 202–214, Singapore. Association for Computational Linguistics.
- Yi Feng, Chuanyi Li, and Vincent Ng. 2022. [Legal Judgment Prediction: A Survey of the State of the Art](#). In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 5461–5469, Vienna, Austria. International Joint Conferences on Artificial Intelligence Organization.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and others. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.
- Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234.
- Elias Jacob De Menezes-Neto and Marco Bruno Miranda Clementino. 2022. [Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from Brazilian federal courts](#). *PLOS ONE*, 17(7):e0272287.
- Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Zhuo, Patrick Emerton, and Genevieve Grant. 2023. [Can ChatGPT Perform Reasoning Using the IRAC Method in Analyzing Legal Scenarios Like a](#)

- [Lawyer?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13900–13923, Singapore. Association for Computational Linguistics.
- Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the Supreme Court of the United States. *PloS one*, 12(4):e0174698. Publisher: Public Library of Science San Francisco, CA USA.
- Aaron Kaufman, Peter Kraft, and Maya Sen. 2017. Machine learning, text data, and supreme court forecasting. *Project Report, Harvard University*.
- Arshdeep Kaur and Bojan Bozic. 2019. Convolutional Neural Network-based Automatic Prediction of Judgments of the European Court of Human Rights. In *AICS*, pages 458–469.
- Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, and others. 2025. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*.
- Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. [Learning to Predict Charges for Criminal Cases with Legal Basis](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736, Copenhagen, Denmark. Association for Computational Linguistics.
- Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. [ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.
- Devin J. McConnell, James Zhu, Sachin Pandya, and Derek Aguiar. 2021. [Case-level prediction of motion outcomes in civil litigation](#). In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 99–108, São Paulo Brazil. ACM.
- Masha Medvedeva and Pauline McBride. 2023. [Legal Judgment Prediction: If You Are Going to Do It, Do It Right](#). In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 73–84, Singapore. Association for Computational Linguistics.
- Masha Medvedeva, Ahmet Üstün, Xiao Xu, Michel Vols, and Martijn Wieling. 2021. Automatic Judgement Forecasting for Pending Applications of the European Court of Human Rights. In *ASAIL/LegalAIIA@ ICAIL*, pages 12–23.
- Shubham Kumar Nigam, Deepak Patnaik Balaramamanthi, Shivam Mishra, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2025. NYAYAANUMANA and INLEGALLAMA: The largest Indian legal judgment prediction dataset and specialized language model for enhanced decision analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11135–11160.
- Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the natural legal language processing workshop 2021*, pages 19–35.
- George L Priest and Benjamin Klein. 1984. The selection of disputes for litigation. *The journal of legal studies*, 13(1):1–55. Publisher: The University of Chicago Law School.
- Alexandre Quemy and Robert Wrembel. 2022. [ECHR-OD: On building an integrated open repository of legal documents for machine learning applications](#). *Information Systems*, 106:101822.
- Bashar Ramadan, Ming-Chieh Liu, Michael C. Burkhart, William F. Parker, and Brett K. Beaulieu-Jones. 2025. [Diagnostic Codes in AI Prediction Models and Label Leakage of Same-Admission Clinical Outcomes](#). *JAMA Network Open*, 8(12):e2550454.
- Edwina L Rissland and David B Skalak. 1991. CABARET: rule interpretation in a hybrid architecture. *International journal of man-machine studies*, 34(6):839–887. Publisher: Elsevier.
- Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. *arXiv preprint arXiv:2505.10573*.
- T.y.s.s Santosh, Kevin Ashley, Katie Atkinson, and Matthias Grabmair. 2024. [Towards Supporting Legal Argumentation with NLP: Is More Data Really All You Need?](#) In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 404–421, Miami, FL, USA. Association for Computational Linguistics.
- Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. 2022. ClassActionPrediction: A challenging benchmark for legal judgment prediction of class action cases in the US. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 31–46.
- Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. [A Comprehensive Evaluation of Large Language Models on Legal Judgment Prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7337–7348, Singapore. Association for Computational Linguistics.

- Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef Van Genabith. 2017. Exploring the Use of Text Classification in the Legal Domain. *Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2017)*.
- Elizabeth C Tippet, Charlotte S Alexander, Karl Branting, Paul Morawski, Carlos Balhana, Craig Pfeifer, and Sam Bayer. 2021. Does lawyering matter? Predicting judicial decisions from legal briefs, and what that means for access to justice. *Tex. L. Rev.*, 100:1157. Publisher: HeinOnline.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.
- Shaurya Vats, Atharva Zope, Somsubhra De, Anurag Sharma, Upal Bhattacharya, Shubham Kumar Nigam, Shouvik Guha, Koustav Rudra, and Kripabandhu Ghosh. 2023. LLMs—the Good, the Bad or the Indispensable?: a use case on legal statute prediction and legal judgment prediction on Indian court cases. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12451–12474.
- Pengfei Wang, Yu Fan, Shuzi Niu, Ze Yang, Yongfeng Zhang, and Jiafeng Guo. 2019. [Hierarchical Matching Network for Crime Classification](#). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 325–334, Paris France. ACM.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. [Precedent-Enhanced Legal Judgment Prediction with LLM and Domain-Model Collaboration](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12060–12075, Singapore. Association for Computational Linguistics.
- Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xi-anpei Han, Zhen Hu, Heng Wang, and others. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.
- Huiyuan Xie, Felix Steffek, Joana De Faria, Christine Carter, and Jonathan Rutherford. 2024. [The CLC-UKET Dataset: Benchmarking Case Outcome Prediction for the UK Employment Tribunal](#). In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 81–96, Miami, FL, USA. Association for Computational Linguistics.
- Shanshan Xu, T.y.s.s Santosh, Oana Ichim, Isabella Risini, Barbara Plank, and Matthias Grabmair. 2023. [From Dissonance to Insights: Dissecting Disagreements in Rationale Construction for Case Outcome Classification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9576, Singapore. Association for Computational Linguistics.
- Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*.
- Kepu Zhang, Haoyue Yang, Xu Tang, Weijie Yu, and Jun Xu. 2025. [Beyond Guilt: Legal Judgment Prediction with Trichotomous Reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1815–1826, Suzhou, China. Association for Computational Linguistics.

A Prompt Templates

A.1 Simple Prompt

You are a legal assistant. Your task is to predict the most likely outcome of the motion below, based on the motion and the response in opposition. Please return the most likely outcome using one of these three labels:

- *granted: the entire motion is granted;*
- *denied: the entire motion is denied;*
- *partial: any combination of parts of the motion are granted and parts are denied.*

By way of background, these documents are from a contract disputes case in the US District Court. The outcome label (granted, denied or partial) should correspond to the label that the court administrative staff will enter into the docket entry when this motion is decided by the judge.

=====
HERE IS THE FULL TEXT OF THE MOTION:
 =====
 {M}

=====
HERE IS THE FULL TEXT OF THE RESPONSE IN OPPOSITION:
 =====
 {R}

A.2 IRAC Prompt

You are a legal assistant. Your task is to predict the most likely outcome of the motion below, based on the motion and the response in opposition. Please return the most likely outcome using one of these three labels:

- *granted: the entire motion is granted;*
- *denied: the entire motion is denied;*

- *partial*: any combination of parts of the motion are granted and parts are denied.

By way of background, these documents are from a contract disputes case in the US District Court.

The outcome label (granted, denied or partial) should correspond to the label that the court administrative staff will enter into the docket entry when this motion is decided by the judge.

You should reason through this task using the IRAC legal reasoning framework. That framework involves these four steps:

1. *Issue*: Identify the legal issues;
2. *Rule*: Identify the relevant legal rules for the issues;
3. *Application*: Apply the rules to the facts at hand;
4. *Conclusion*: Reach a conclusion with respect to the application of law to facts and determine what the legal outcome of the facts are.

=====
HERE IS THE FULL TEXT OF THE MOTION:

=====
{M}

=====
HERE IS THE FULL TEXT OF THE RESPONSE
IN OPPOSITION:

=====
{R}