

# CLAIM: An Intent-Driven Multi-Agent Framework for Analyzing Manipulation in Courtroom Dialogues

Disha Sheshanarayana\* Tanishka Magar\* Ayushi Mittal Neelam Chaplot

Manipal University Jaipur, India

disha.229301161@mun.manipal.edu, tanishka.229301736@mun.manipal.edu,  
ayushi.229209033@mun.manipal.edu, neelam.chaplot@jaipur.manipal.edu

## Abstract

Courtrooms are places where lives are determined and fates are sealed, yet they are not impervious to manipulation. Strategic use of manipulation in legal jargon can sway the opinions of judges and affect the decisions. Despite the growing advancements in NLP, its application in detecting and analyzing manipulation within the legal domain remains largely unexplored. Our work addresses this gap by introducing LegalCon, a dataset of 1,063 annotated courtroom conversations labeled for manipulation detection, identification of primary manipulators, and classification of manipulative techniques, with a focus on long conversations. Furthermore, we propose CLAIM, a two-stage, Intent-driven Multi-agent framework designed to enhance manipulation analysis by enabling context-aware and informed decision-making. Our results highlight the potential of incorporating agentic frameworks to improve fairness and transparency in judicial processes. We hope that this contributes to the broader application of NLP in legal discourse analysis and the development of robust tools to support fairness in legal decision-making. Our code and data are available at [CLAIM](#).

## 1 Introduction

Courtroom decisions have significant legal and societal implications, shaping legal precedents and affecting lives. However, the inherently adversarial and strategic nature of legal discourse fosters an environment where linguistic manipulation is prevalent. Tactical orchestration of manipulation can shape perceptions, steer arguments, and ultimately influence judicial outcomes. Over the years, studies like (Gold, 1987), (Lively et al., 2020) and (Wood, 2012) have explored the various techniques, covert and overt, employed to manipulate courtroom dynamics, which can manifest through crafted narratives and psychological attacks. (Vinson, 1982)

\*These authors contributed equally to this work



Figure 1: An example of a courtroom conversation that contains manipulation, but ChatGPT-4o fails to identify the primary manipulator and technique accurately.

presented that defense tactics, such as contextual stimuli, can be used by lawyers to psychologically influence jurors, making it difficult for them to be unbiased or nonaligned.

(Gold, 1987) emphasized that while measures against these tactics such as judicial training in psychology, court-appointed experts, increased jury compensation, and expanded jury panels may be costly, the consequences of flawed jury decision-making can be just as significant. Despite its serious implications for justice, computational approaches for detecting and analyzing manipulation tactics in courtrooms remain significantly underdeveloped.

Paper	Dataset	Detection	Manipulator	Technique
MentalManip (Wang et al., 2024)	MentalManip	Yes	No	Yes
Intent-Aware Prompting (Ma et al., 2024)	MentalManip	Yes	No	No
Advanced Prompting (Yang et al., 2024)	MentalManip	Yes	No	No
Communication is All You Need (Ma et al., 2025)	Multi LLM	Yes	No	Yes
Human Decision-Making and AI (Sabour et al., 2025)	Custom	Yes	No	No
MANITWEET (Huang et al., 2023)	MANITWEET	Yes	No	Yes
CLAIM (Our Work)	LegalCon	Yes	Yes	Yes

Table 1: Comparison of related work on manipulation analysis.

While multiple studies have explored social manipulation—including fake news detection (Zhang et al., 2024), toxic language identification (Li et al., 2024) as well as the detection and categorization of mental manipulation techniques (Wang et al., 2024)—these efforts rarely focus on the legal domain. The complexity of legal language means that manipulation can be concealed behind legal jargon and thus is even more challenging to detect. Current SoTA models struggle to detect manipulation in courtroom debates, particularly in longer conversations, and often fail to capture the nuanced, context-dependent nature of courtroom discourse, as demonstrated in Figure 1.

Our study aims to contribute to this research gap by analyzing manipulation in courtroom conversations, with a focus on long and comprehensive exchanges. We introduce **LegalConflict**, a dataset consisting of conversations and debates sourced from transcripts across various judicial settings. It comprises 1,063 conversations annotated for manipulation detection, identification of the primary manipulator, and classification of manipulation techniques. To evaluate this dataset, we conducted extensive experiments using SoTA models. However, these models struggled to accurately identify manipulation, particularly in complex and context-dependent cases, highlighting the need for a more specialized approach. To address this challenge, we propose **CLAIM** (Courtroom Language Analysis with Intent-driven Multi-agent Framework), a novel two-stage framework that combines an Intent-Driven Chain-of-thought prompting (Wei et al., 2022) with a Multi-Agent framework to provide a comprehensive analysis of manipulation

in courtroom dialogues. Our methodology first processes courtroom transcripts through an Intent-driven and CoT prompting technique, generating preliminary manipulation assessments. These are then passed to the Multi-Agent framework for refinement and evidence gathering. This sequential approach allows for increasingly sophisticated analysis by combining the strengths of intent-specific prompting with the collaborative reasoning capabilities of multiple specialized agents. Experimental results across diverse legal contexts show that our approach achieves significant improvements over baseline methods, particularly in detecting the primary manipulator.

## 2 Related Works

Detecting and analyzing manipulation in conversations has been an emerging research focus, especially with the rise of large language models (LLMs) and their role in social, legal, and media contexts. Several datasets and frameworks have emerged to explore different kinds of manipulation like mental manipulation and, persuasion, misinformation, and toxicity. Table 1 summarizes some existing work on manipulation. (Wang et al., 2024) introduced MentalManip, a benchmark dataset that enables fine-grained classification of manipulation in conversation. This study focuses on identifying various manipulation techniques and vulnerabilities used in conversation, providing a solid foundation for manipulation detection in a conversation. Building on this, (Ma et al., 2024) proposed an Intent-Aware Prompting approach that leverages speaker intent for improved detection, while (Yang et al., 2024) demonstrated the effec-

tiveness of Chain-of-Thought (CoT) prompting for nuanced understanding of manipulation in conversations. However, these approaches do not attempt to identify the primary manipulator or extract manipulative techniques. The study (Ma et al., 2025) presents a multi-LLM framework for generating persuasive dialogues. It includes both persuasion detection and technique identification, demonstrating the utility of collaborative LLM setups.

The increasing prevalence of misinformation has also led to the development of specialized models designed for social manipulation and fake news detection, (Zhang et al., 2024) outlines strategies for mitigating manipulation in the LLM era, emphasizing the need for explainability and interpretability. (Huang et al., 2023) introduces a benchmark dataset MANITWEET for detecting manipulative tweets based on their distortion of news articles, highlighting the limitations of fact-checking systems and the need for better manipulation detection in social contexts.

Agent-based approaches have also shown promise. (Li et al., 2024) leverages LLM agents for fake news detection, while (Jeptoo and Sun, 2024) proposes a multi-agent debate framework, where agents critique each other’s outputs to improve factual accuracy. While these studies provide important insights into manipulation across domains, none of them focus on courtroom dialogues.

### 3 Constructing LegalCon

#### 3.1 Data Sourcing and Pre-processing

The dataset was curated from multiple public sources and includes transcripts from various courts across the United States, such as Supreme Courts, Family Support Courts, Trial Courts, and Small Claims Courts. This selection was made to include multiple judicial contexts and case types, ensuring a comprehensive view of courtroom discourse. Long-form courtroom conversations were prioritized to capture in-depth arguments and interactions. All the transcripts collected are in English language.

A significant portion of the transcripts was sourced from Oyez (Oyez, 2020), a multimedia judicial archive that provides publicly accessible Supreme Court transcripts. Additionally, courtroom interactions were extracted from legal television shows such as Paternity Court (Court, 2013), Support Court with Judge Vonda B. (with Judge

Vonda B., 2018), and The People’s Court (Court, 2014). While these shows are staged, they feature judges and legal professionals, and the dialogue mirrors courtroom conversations. To preserve the integrity and legal accuracy of the dataset, careful verification was conducted to ensure that the transcripts adhered to standard legal frameworks and courtroom protocols.

In total, the dataset comprises 1063 conversations featuring interactions between plaintiffs, defendants, lawyers, and judges. We also placed a special emphasis on collecting long conversations and the majority of the dialogues in the dataset average approximately 1000 words as shown in Figure 3. The distribution of manipulative and non-manipulative dialogues is given in Table 2. Refer Figure 2 and Figure 3 for detailed visualization of the dataset. To eliminate potential biases and standardize the dialogues, the original speakers’ names were replaced with their generic roles: “Plaintiff”, “Defendant”, “Plaintiff’s Lawyer”, “Defendant’s Lawyer”, and “Judge”.

#### 3.2 Labeling Schema and Annotation

Building on insights from (Aldridge and Luchjenbroers, 2007) and (Kadoch, 2000), we developed a multi-level labeling schema constituting three key components: (1) detecting manipulation, (2) identifying the primary manipulator, and (3) categorizing the specific manipulative techniques employed. The labeling schema and definitions are mentioned in the Appendix A.1.

The three questions used for labeling and evaluation have been kept consistent throughout our research:

- Q1: Is the given dialogue manipulative?
- Q2: If yes, identify the primary manipulator.
- Q3: If the dialogue is manipulative, identify the manipulative techniques employed by the primary manipulator.

Dataset	Manipulative	Non-Manipulative
LegalCon	663	400

Table 2: LegalCon Dataset Distribution

Through an extensive review of courtroom dialogues and existing studies on psychological manipulation in legal as well as other domains, we identified 11 frequently used manipulative techniques (Fischer, 2022) (McDowell, 1991) (Aguado,

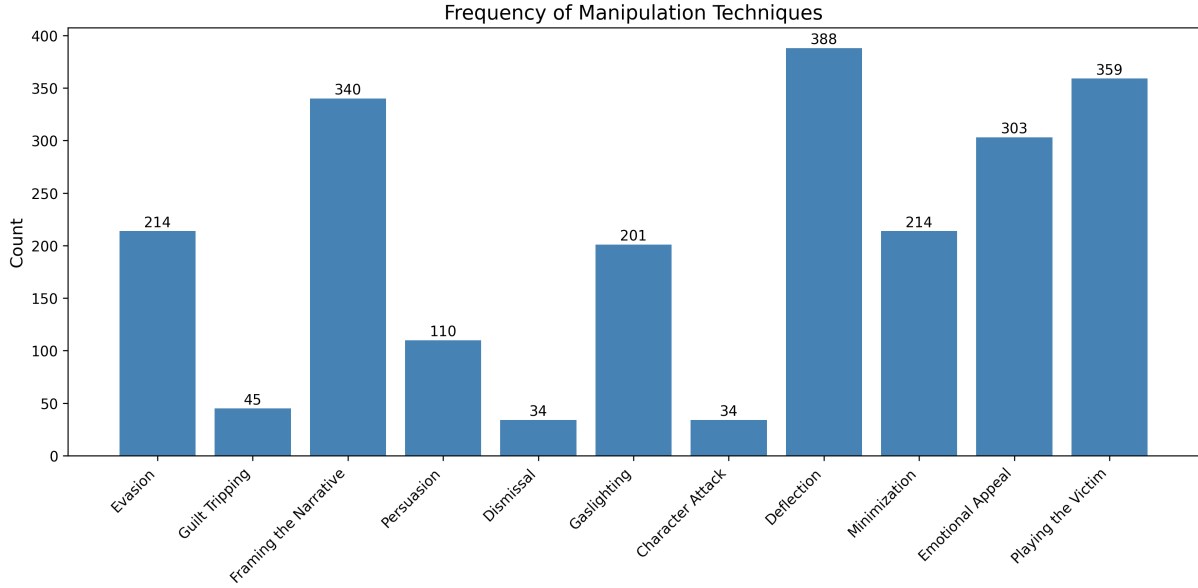


Figure 2: Bar graph showing frequency of different Manipulative Techniques in LegalCon dataset.

Distribution of Primary Manipulator

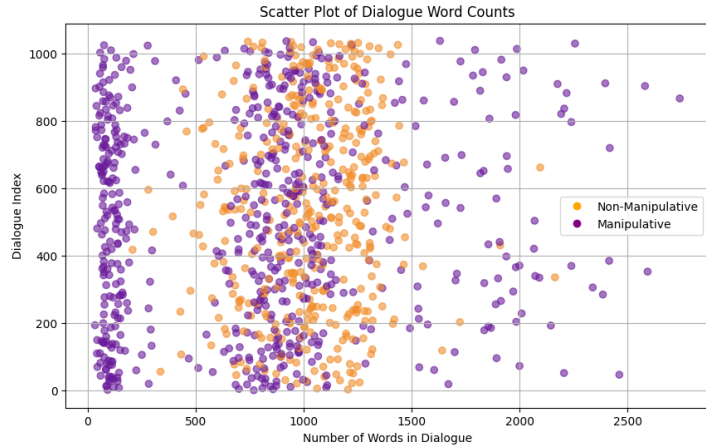
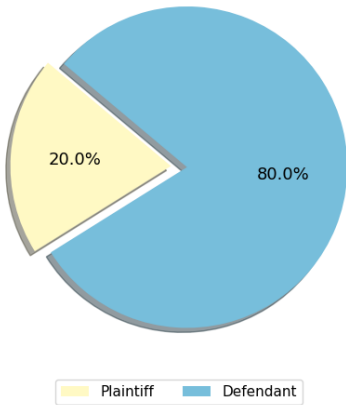


Figure 3: Pie Chart of Primary Manipulator distribution and Scatter Plot of Words Counts in Dialogues in LegalCon.

2015). We then consulted with psychology professors to verify and refine these categorizations. This schema is illustrated in appendix.

While we initially explored using NLP techniques and LLMs for annotation, their performance proved inadequate. Hence, the four of us manually annotated the dataset, leveraging evidences and inferences from LLMs and prior research. To evaluate the reliability of the annotation process, we conducted a post-hoc inter-annotator agreement study on a subset of 100 dialogues. The annotators independently labeled this subset, and agreement was measured across the three tasks. Cohen’s Kappa for Q1 and Q2 was 0.68 and 0.59 respectively. For Q3, which is a multi-label classification task, we used Krippendorff’s Alpha, which yielded a score of 0.41, also indicating moderate agreement. As Q2

and Q3 were only applied when Q1 was marked as manipulative, the number of annotated items was filtered accordingly. These scores proved to be consistent with expectations for subjective annotation tasks in legal and psychological domains.

Since by its inherent nature, manipulation is subjective and manipulation in a legal context especially so, we made an effort to only include the data points that had majority consensus in LegalCon.

## 4 Methodology

In courtroom conversations, manipulation is often quiet and deeply rooted in the speaker’s intent, rhetorical strategy, and power dynamics. Psychological studies show that individuals with a more substantial Theory of Mind (ToM) are better at

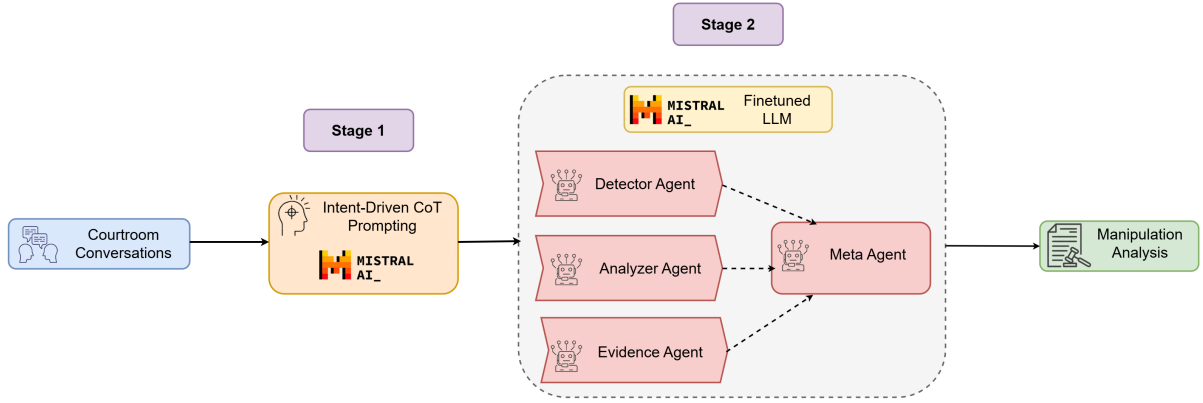


Figure 4: Overview of CLAIM: A two-stage framework for manipulation analysis

interpreting others’ intentions and withstanding manipulation (Chen et al., 2024). A recent study also suggests that LLMs can improve their ToM performance when guided by structured reasoning techniques like Chain-of-Thought (CoT) prompting (Wei et al., 2022).

Building on these studies, we propose **CLAIM**, a two-stage framework to improve manipulation analysis in courtroom dialogues: (1) Intent-Driven Chain-of-Thought Prompting to extract speaker intent and (2) a multi-agent decision framework to analyze manipulation using these inferred intents with the help of agents. This framework allows our method to reason with the speaker’s intent based on contextual evidence. Figure 4 provides an overview of our framework, illustrating how each stage contributes to the final result.

#### 4.1 Intent-Driven CoT Prompting

In the first stage of our framework, we aim to uncover the underlying intent of each speaker. To achieve this, we implement Intent-Driven Chain-of-Thought (CoT) Prompting, which infers each speaker’s intents throughout the courtroom dialogue. This stage is inspired by the approach presented in the study (Ma et al., 2024). The resulting intent summaries provide a structured representation of speaker intent and act as intermediate reasoning scaffolds for subsequent analysis.

#### 4.2 Multi-Agent Framework

Manipulation analysis in long-form courtroom conversations is a complex task that requires context-aware decision-making. Relying on a single model to manage the full complexity of such discourse often leads to brittle outputs and poor interpretability. To address this, the second stage of our framework adopts a multi-agent architecture, where each

agent is responsible for a specific subtask within the manipulation analysis pipeline. LLM agents are particularly effective for such multi-step decision-making processes, as they support decomposition of reasoning, evidence aggregation, and inter-agent communication. In our framework, each agent operates independently but shares intermediate outputs with other agents to collaboratively arrive at a final judgment.

To optimize the performance of our agents for legal-domain reasoning, we fine-tuned the Mistral-7B (Jiang et al., 2023) language model on LegalCon, a curated dataset of courtroom and legal exchanges explained in Section 3. We use QLoRA (Dettmers et al., 2023), a memory-efficient parameter-efficient fine-tuning (Xu et al., 2023) method that enables low-resource adaptation of large models. This framework results in lightweight, high-performance agents optimized for courtroom manipulation analysis, organized into four specialized components, each responsible for a specific subtask:

- **Detector Agent:** This agent is designed to determine whether the courtroom dialogues contain manipulation.
- **Analyzer Agent:** This agent is designed to identify the primary manipulator and classify the manipulation techniques used by the primary manipulator in the dialogue.
- **Evidence Agent:** This agent is designed to extract evidence from the dialogue that substantiates the manipulator and techniques used by them.
- **Meta Agent:** This agent is designed to aggregate the outputs from all agents, generating a final set of labels.

Agents receive the courtroom dialogue along



Experiment Setting	Llama-3.1 8B				Mistral 7B			
	$P$	$R$	ACC	$F_1$	$P$	$R$	ACC	$F_1$
Zero-shot prompting	.713	.600	.609	.614	.618	<b>.937</b>	.609	.518
Few-shot prompting	.653	.811	.622	.598	.717	.747	.667	.664
CLAIM Stage 1	-	-	-	-	.753	.674	.667	.670
<b>CLAIM (Our Work)</b>	-	-	-	-	<b>.757</b>	.821	<b>.731</b>	<b>.727</b>

Table 3: Results of the manipulation detection task on LEGALCON.  $P$ ,  $R$ , ACC, and  $F_1$  stand for binary precision, binary recall, accuracy, and  $F_1$  -score, respectively.

Experiment Setting	Llama-3.1 8B				Mistral 7B			
	$P$	$R$	ACC	$F_1$	$P$	$R$	ACC	$F_1$
Zero-shot prompting	.476	.481	.481	.467	.512	.340	.340	.340
Few-shot prompting	.484	.449	.449	.454	.557	.481	.481	.489
CLAIM Stage 1	-	-	-	-	.419	.526	.526	.464
<b>CLAIM (Our Work)</b>	-	-	-	-	<b>.608</b>	<b>.609</b>	<b>.609</b>	<b>.602</b>

Table 4: Results of the primary manipulator identification task LEGALCON.  $P$ ,  $R$ , ACC, and  $F_1$  stand for binary precision, binary recall, accuracy, and  $F_1$  -score, respectively.

Experiment Setting	Llama-3.1 8B					Mistral 7B				
	$P$	$R$	ACC	$F_1$	$Jc$	$P$	$R$	ACC	$F_1$	$Jc$
Zero-shot prompting	.1899	.3379	.2436	.2082	.3106	.1715	.3988	.0385	.1387	.1265
Few-shot prompting	.1515	.4198	.1346	.1915	.2271	.2392	.3394	.2179	.2118	.3145
CLAIM Stage 1	-	-	-	-	-	.2582	.4201	.2115	<b>.2452</b>	.3028
<b>CLAIM (Our Work)</b>	-	-	-	-	-	<b>.2639</b>	<b>.4300</b>	<b>.2564</b>	.2354	<b>.3618</b>

Table 5: Results of the manipulation technique identification task on LEGALCON.  $P$ ,  $R$ , ACC,  $F_1$  and  $Jc$  stand for binary precision, binary recall, accuracy,  $F_1$ -score, and Jaccard coefficient, respectively.

with the intents of each speaker generated in Stage 1 as input. These intent representations provide additional reasoning context, allowing agents to align manipulation judgments with inferred speaker goals. Then the meta agent summarizes and compiles the final result. This framework enables more robust reasoning and improves interpretability, as each decision is traceable to an agent’s role and output.

## 5 Experiments

### 5.1 Experimental Settings

We conducted experiments on three tasks using the LegalCon dataset to assess the performance of CLAIM and SoTA models in analyzing manipulation in courtroom dialogues. These tasks include: Manipulation Detection, Primary Manipulator Classification, and Manipulation Technique Classification. For the experimental data, we randomly split the dataset into 70% for training, 15% for validation, and 15% for testing. We compared two models, Mistral-7B (Jiang et al., 2023) and Llama 3.1 8B (Grattafiori et al., 2024), across

four experimental settings: zero-shot prompting, few-shot prompting, CLAIM Stage 1 alone and CLAIM.

In the zero-shot prompting, courtroom dialogues were presented directly to the models with instructions to detect whether manipulation occurred. In the few-shot prompting, we provided each model with two non-manipulative and three manipulative courtroom conversations as in-context examples along with the task prompt. The format for both zero-shot and few-shot prompting are outlined in the Appendix. Additionally, we experimented with CLAIM Stage 1 alone, where speaker intents were inferred from the dialogues using CoT prompting. Manipulation analysis was then performed based solely on these inferred intents, and corresponding results were calculated. In our proposed framework CLAIM, we applied our two-stage Intent-Driven with a Multi-Agent framework, where the inferred intents of speakers guided specialized agents to analyze manipulation. The agents were powered by a Mistral-7B model fine-tuned using QLoRA, a memory-efficient PEFT method. The fine-tuning was performed with a learning rate of  $1e-4$  to op-

timize the model for legal-domain reasoning and manipulation detection. All experiments were conducted on an MSI GeForce RTX 3060 GPU. Both models were tested at temperatures of 0.4 and 0.6, and the models performed most consistently and accurately at a temperature of 0.4.

## 5.2 Experimental Results

**Table 3** presents the experimental results for manipulation detection, comparing CLAIM against baseline models using zero-shot and few-shot prompting as well as with CLAIM Stage 1. The results indicate that CLAIM outperforms the baseline models, achieving higher accuracy. This demonstrates the effectiveness of our framework in detecting manipulation more reliably than traditional prompting techniques. **Table 4** presents the results for primary manipulator identification, a challenging and inherently subjective task. The findings indicate that all models face difficulties in accurately identifying the manipulator. Further analysis reveals that this challenge arises from the models frequently misattributing manipulative intent. Despite this, CLAIM demonstrates a notable improvement over baseline methods. **Table 5** presents the results for identifying manipulation techniques employed by the primary manipulator. The models do not perform well and exhibit relatively low accuracy. Since this is a multi-label classification task, correctly identifying all the techniques is challenging. Traditional accuracy metrics may not fully capture performance. To address this, we used Jaccard Similarity Coefficient, which is used to calculate the overlap between predicted and actual manipulation techniques as shown in equation (1).

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

where  $A$  is the set of true manipulation techniques for a given instance, and  $B$  is the set of predicted techniques.

CLAIM achieved the highest Jaccard score of 0.3618. However, due to the subjective nature of manipulation detection, distinguishing certain techniques remains challenging and open to debate.

## 6 Conclusion and Future Work

This study introduces LegalCon, a dataset of courtroom conversations aimed at detecting and analyzing manipulation. Alongside this, we propose CLAIM, a two-stage Intent-driven Multi-Agent

framework, to enhance the detection and analysis of manipulation in courtroom conversations. Extensive experiments showed that our method consistently outperformed baseline models on various prompting techniques. However, the models struggled to accurately identify specific manipulative techniques, revealing a critical limitation. These findings highlight the inherently subjective and nuanced nature of manipulation. With LegalCon and CLAIM we hope to address a critical gap in NLP research at the intersection of law and manipulative language. Since legal decisions are lasting and influential we hope that this lays a necessary foundation for further work in this field.

Future work could focus on expanding the LegalCon dataset to include different types of cases and create a more comprehensive dataset. Multi-lingual transcripts can be incorporated to enhance diversity and enable cross-cultural analysis of manipulative language. Multi-modal frameworks can be explored to yield deeper insights into manipulation dynamics. Integrating these advancements into real-world legal settings is a valuable opportunity and moving forward, efforts should focus on the responsible deployment of such models to support legal professionals and promote fairness in courtrooms. Given that manipulation detection in the legal domain is a relatively underexplored area, further research in this field could provide valuable insights and open up new avenues for improving legal processes and contributing to the advancement of the application of technology in the legal domain.

## 7 Limitations

While our proposed framework demonstrates promising results, there are several limitations to consider:

**Subjectivity in Manipulation Detection:** Manipulation, by nature, is subjective, which makes it challenging for models to accurately identify manipulative behavior. Since there are no well-defined standard limits distinguishing between manipulative and non-manipulative behaviors, especially in arguments and debates, it remains particularly complex.

**Dataset Annotation Challenges:** Despite our efforts to ensure high-quality annotations, labeling manipulation, especially for specific techniques remains subjective. While the annotators made

efforts to minimize bias, human interpretation is influenced by personal perspectives. This subjectivity in labeling may affect the consistency and reliability of the dataset, which in turn could impact the model’s training and overall performance.

**Limited Generalizability of the LegalCon Dataset:** LegalCon dataset is limited in scope, covering only a specific set of case types. Hence, the model may not generalize well to other legal contexts or jurisdictions.

**Limited Generalizability of CLAIM framework:** The CLAIM framework was developed for the LegalCon dataset, and optimized particularly to address challenges posed by longer courtroom conversations. However, it may struggle to generalize to shorter dialogues outside of courtroom settings. Additionally, the framework’s complexity might be excessive for such tasks, making it less suitable for simpler or more informal interactions.

**Computational Constraints:** Fine-tuning LLMs requires significant computational resources but due to hardware limitations, we were restricted in terms of batch size and the number of training epochs. The fine-tuning process was conducted on a single MSI GeForce RTX 3060, which limited our ability to experiment with larger models.

## 8 Ethics Statement

All data used in the LegalCon dataset was sourced from publicly available transcripts, including court proceedings and legally staged courtroom television shows. We ensured that no personally identifiable information (PII) was retained. Speaker names were anonymized and replaced with generic role labels such as "Plaintiff", "Defendant", "Plaintiff’s Lawyer", "Defendant’s Lawyer" and "Judge" to protect identities and maintain legal neutrality.

## References

- Jose Fernández Aguado. 2015. Psychological manipulation, hypnosis, and suggestion. *International Journal of Cultic Studies*, 6.
- Michelle Aldridge and June Luchjenbroers. 2007. Linguistic manipulations in legal discourse: Framing questions and ‘smuggling’ information. *The International Journal of Speech, Language and the Law*, 14(1):85–107.
- Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and et al. 2024. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.
- Paternity Court. 2013. [Paternity court](#).
- The People’s Court. 2014. [The people’s court](#).
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.
- Alexander Fischer. 2022. Then again, what is manipulation? a broader view of a much-maligned concept. *Philosophical Explorations*, 25(2):170–188.
- Victor Gold. 1987. Psychology manipulation in the courtroom. *Neb. L. Rev.*, 66:562.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Kung-Hsiang Huang, Hou Pong Chan, Kathleen McKeown, and Heng Ji. 2023. Manitweet: A new benchmark for identifying manipulation of news on social media. *arXiv preprint arXiv:2305.14225*.
- Korir Nancy Jeptoo and Chengjie Sun. 2024. Enhancing fake news detection with large language models through multi-agent debates. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 474–486. Springer.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chaplot. 2023. Diego de las casas. *Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed*, pages 50–72.
- Laurie C Kadoch. 2000. Seduced by narrative: Persuasion in the courtroom. *Drake L. Rev.*, 49:71.
- Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024. Large language model agent for fake news detection. *arXiv preprint arXiv:2405.01593*.
- Christopher J Lively, Laura Fallon, Brent Snook, and Weyam Fahmy. 2020. Seeking or controlling the truth? an examination of courtroom questioning practices by canadian lawyers. *Psychology, Crime & Law*, 26(4):343–366.
- Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2024. Detecting conversational mental manipulation with intent-aware prompting. *arXiv preprint arXiv:2412.08414*.
- Weicheng Ma, Hefan Zhang, Ivory Yang, Shiyu Ji, Joice Chen, Farnoosh Hashemi, Shubham Mohole, Ethan Gearey, Michael Macy, Saeed Hassanpour, and et al. 2025. Communication is all you need: Persuasion dataset construction via multi-llm communication. *arXiv preprint arXiv:2502.08896*.



Banks McDowell. 1991. The lawyer as manipulator: Is this a useful model for legal education and practice. *Washburn LJ*, 31:506.

Oyez. 2020. [Oyez: U.s. supreme court media](#).

Sahand Sabour, June M Liu, Siyang Liu, Chris Z Yao, Shiyao Cui, Xuanming Zhang, Wen Zhang, Yaru Cao, Advait Bhat, Jian Guan, and et al. 2025. Human decision-making is susceptible to ai-driven manipulation. *arXiv preprint arXiv:2502.07663*.

Donald E Vinson. 1982. Juries: Perception and the decision-making process. *Trial*, 18:52–55.

Yuxin Wang, Ivory Yang, Saeed Hassanpour, and Soroush Vosoughi. 2024. Mentalmanip: A dataset for fine-grained analysis of mental manipulation in conversations. *arXiv preprint arXiv:2405.16584*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Support Court with Judge Vonda B. 2018. [Support court with judge vonda b](#).

Seth William Wood. 2012. *Courtroom Discourse as Verbal Performance: Describing the Unique Sociolinguistic Situation of the American Trial Courtroom*. Brigham Young University.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*.

Ivory Yang, Xiaobo Guo, Sean Xie, and Soroush Vosoughi. 2024. Enhanced detection of conversational mental manipulation through advanced prompting techniques. *arXiv preprint arXiv:2408.07676*.

Yizhou Zhang, Karishma Sharma, Lun Du, and Yan Liu. 2024. Toward mitigating misinformation and social media manipulation in llm era. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1302–1305.

## A Appendix

### A.1 Labeling Schema for LegalCon

Definitions of the 11 manipulative techniques selected and used for labeling LegalCon listed in [Figure 5](#) are:

1. **Gaslighting:** A form of psychological manipulation where a person makes someone doubt their perceptions or sanity by denying the truth or altering reality.
2. **Guilt tripping:** A manipulative tactic where someone tries to make another feel guilty to control their behavior. It often involves exaggerating the impact of their actions or making them feel responsible for things not their fault.
3. **Persuasion:** Influencing someone’s beliefs or actions through reasoning or appealing to their interests.
4. **Evasion:** The act of avoiding a question, responsibility, or engagement, while manipulation involves influencing or controlling someone or something unfairly to one’s advantage.
5. **Framing the narrative:** Selectively highlighting certain aspects of a story to influence an audience’s perception and understanding.
6. **Dismissal:** Ignoring other people’s concerns or questions with the aim to monopolize information and control other people’s choices and decisions.
7. **Character Attack:** Deliberate and sustained effort to damage a person’s reputation, often through manipulation.
8. **Deflection:** Avoiding addressing true feelings or actions by shifting focus onto someone or something else. Deflection may also be used to evade responsibility or to place blame on others, thereby avoiding accountability.
9. **Minimization:** Downplaying or trivializing events, emotions, or experiences to reduce their perceived importance. Often used to invalidate feelings or diminish the impact of harmful behavior.
10. **Emotional appeal:** Attempting to influence others by exploiting emotions instead of using logic or evidence. Often relies on misleading or sentimental language to provoke fear, guilt, or sympathy and bypass rational judgment.
11. **Playing the victim:** Exaggerating or fabricating an event, experience, or emotion to portray themselves as a victim in the situation when in reality they are not a victim.

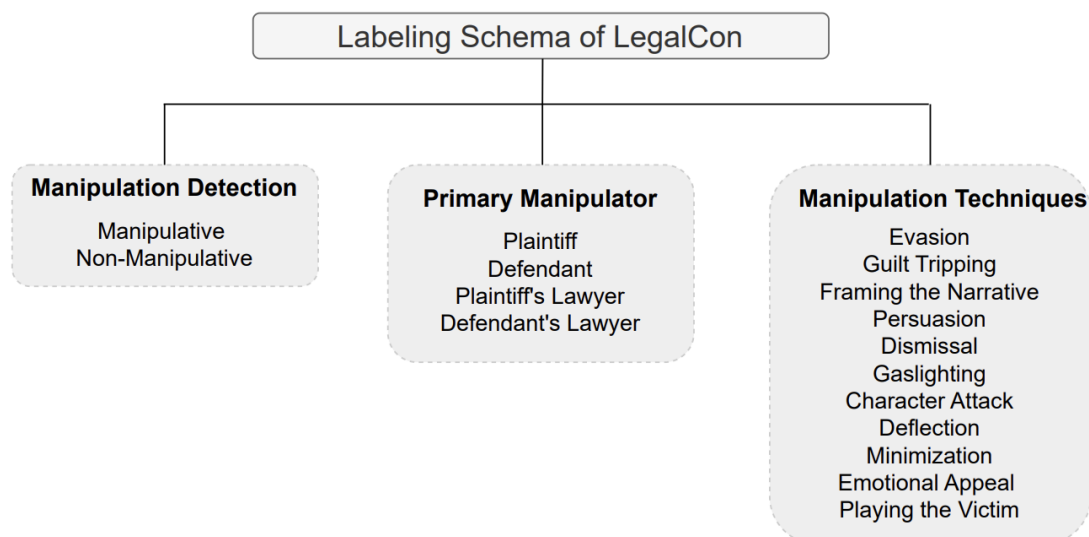


Figure 5: Labeling schema showing 11 manipulative techniques used for annotation in LegalCon.

## A.2 Example Results of CLAIM framework

### CLAIM Stage 1 Results

**Plaintiff's Intent:** The Plaintiff is claiming that she was not paid for her services, was provided uncertified training, and cannot get her training certificate. She is suing the Defendant, for every penny of the \$3,375 she is owed.

**Defendant's Intent:** The Defendant is stating that she has a counterclaim for \$5,000 for breach of contract, as the Plaintiff, did not complete the required 200 hours of training and left a hole in her schedule, owing \$1,040.

### CLAIM Stage 2 Results

#### Detector Agent Result:

Yes

#### Analyzer Agent Results:

Primary Manipulator - Plaintiff

Manipulative Techniques - Emotional appeal, Framing the narrative, Character attack

#### Evidence Agent Results:

Primary Manipulator - Plaintiff

Manipulative Techniques - Emotional appeal, Framing the narrative, Character attack

#### Meta Agent Results:

Manipulation Present - Yes

Primary Manipulator - Plaintiff

Techniques Used - Emotional appeal, Framing the narrative, Character attack

### A.3 Prompt Template in Experiments

#### Zero-shot Prompting

You are a manipulation detection expert. Your task is to analyze this dialogue and determine if there is manipulation present.

**### Instructions:**

1. Determine whether any manipulation is present. Answer with only 'Yes' or 'No'. Do not add anything else.
2. If manipulation is present, then identify the primary manipulator. Answer using only the speaker label. Don't add anything else.
3. Identify the manipulation techniques used by them, selecting only from the allowed list. Provide the techniques using only those exact terms. Do not add anything else.

<insert dialogue>

#### Few-shot Prompting

You are a manipulation detection expert. Your task is to analyze this dialogue and determine if there is manipulation present. Here are five examples:

Example 1:

<insert manipulative\_dialogue1>

<insert manipulative\_answer1>

Example 2:

<insert manipulative\_dialogue2>

<insert manipulative\_answer2>

Example 3:

<insert manipulative\_dialogue3>

<insert manipulative\_answer3>

Example 4:

<insert nonmanipulative\_dialogue1>

<insert nonmanipulative\_answer1>

Example 5:

<insert nonmanipulative\_dialogue2>

<insert nonmanipulative\_answer2>

**### Instructions:**

1. Determine whether any manipulation is present. Answer with only 'Yes' or 'No'. Do not add anything else.
2. If manipulation is present then identify the primary manipulator. Answer using only the speaker label. Don't add anything else.
3. Identify the manipulation techniques used by them, selecting only from the allowed list. Provide the techniques using only those exact terms. Do not add anything else.

<insert dialogue>

### CLAIM Stage 1 Prompt

You are reading a transcript from a courtroom conversation.

1. Carefully read the dialogue.
2. Think step-by-step about what the plaintiff's and defendant's statements suggest.
3. Reason about the plaintiff's and defendant's goals or motives behind their words.
4. Summarize the plaintiff's and defendant's intent in a sentence.

### CLAIM Stage 2 Prompts

#### **Detector Agent:**

You are a manipulation detection expert. Your task is to analyze the dialogue and the corresponding intents to determine whether manipulation is present.

#### **### Instructions:**

1. Read the dialogue carefully.
2. Analyze it in the context of the provided intents.
3. Determine whether any manipulation is present. Answer with only 'Yes' or 'No'. Do not add anything else.

#### **Analyzer Agent:**

You are responsible for identifying manipulation analysis within a courtroom dialogue using both the dialogue and the inferred speaker intents.

#### **### Instructions:**

1. Identify the primary manipulator. Answer using only the speaker label. Do not add anything else.
2. Identify the manipulation techniques used by them, selecting only from the allowed list. Provide the techniques using only those exact terms. Do not add anything else.

#### **Evidence Agent:**

You are tasked with validating the manipulation analysis based on the dialogue.

#### **### Instructions:**

1. Review whether the identified primary manipulator and manipulative techniques are correct.
2. If incorrect, update them. Answer only with the updated result.