LLMs in front of the Supreme Court: Can LLMs understand and reason Justice and Advocate?

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) excel at legal document retrieval and case summarization, their strategic theory of mind (ToM) reasoning in legal tasks remains largely uninvestigated. This paper examines LLMs of different roles in U.S. Supreme Court oral arguments, where *Justices* challenge *Advocates* through adversarial and strategically framed questioning before making final decisions. We introduce `LegalToM`, the first benchmark designed to evaluate LLMs ToM reasoning in legal contexts, using U.S. Supreme Court oral arguments as a natural testbed. In `LegalToM`, a *Justice agent* predicts future lines of inquiry and produces legally coherent, strategically structured questions. An *Advocate agent* interprets judicial intentions and generates persuasive responses that adapt to the evolving context. We evaluate them in four tasks: understanding intentions of justices, predicting subsequent questions from justices, and generating contextually adaptive responses for both justices and advocates. Evaluations of a diverse range of open-source and proprietary LLMs show that while larger models and extended context windows yield consistent improvements, even the strongest systems fall short of expert human performance. `LegalToM` provides a rigorous testbed for exploring the emergence and boundaries of multi-agent ToM reasoning and human-AI interactions in professional legal processes. Our project is available at `https://anonymous.4open.science/r/oral_argument-48DD/`.

1 Introduction

Large language models (LLMs) are increasingly applied in the legal domain for contract review, document drafting, and legal retrieval. However, their capability to engage in complex legal reasoning dialogues, such as U.S. Supreme Court oral arguments (Wolfson, 2001; Johnson et al., 2006), remains largely unexplored. In these complex legal settings, justices and advocates interact dynamically to clarify legal issues and challenge one another's reasoning, making it an ideal testbed for evaluating the limits of LLM reasoning.

Supreme Court oral arguments pose unique challenges for LLMs. Questions from the Justices are strategically crafted and often contain implicit intentions or hypothetical traps. Moreover, the dialogues are long, adversarial, and highly context-dependent, requiring reasoning about the beliefs and goals of other participants. These properties go beyond surface-level text understanding and naturally align with the core elements of Theory of Mind (ToM) (Premack and Woodruff, 1978), including the ability to infer others' hidden intentions, predict their future actions, and generate contextually adaptive responses.

To systematically study these challenges, we introduce `LegalToM`, a new benchmark that formalizes Supreme Court oral arguments as a two-agent environment. As shown in Figure 1, the advocate agent must infer a justice's hidden intent and produce persuasive responses, while the justice agent predicts subsequent questions and generates coherent, strategically framed prompts. This formulation captures both legal reasoning and ToM perspective taking, offering a realistic multi-agent testbed for LLMs.

`LegalToM` is constructed from real Supreme Court oral argument transcripts (Oyez, Inc., 2025), comprising 31 cases and 3,884 evaluation samples annotated across three difficulty levels and contextual settings. We evaluate a diverse set of open-source models, including Qwen2.5_1.5B, Qwen2.5_7B, Qwen2.5_70B (Yang et al., 2024), Qwen3_4B, Qwen3_30B (Yang et al., 2025), LLaMA3.3_70B (Meta, 2024), Gemma3_4B (Team et al., 2025), Phi-4_14B (Abdin et al., 2024), SaulLM-7B (Colombo et al., 2024), as well as proprietary models such as GPT-4o (Hurst et al., 2024) and OpenAI's reasoning

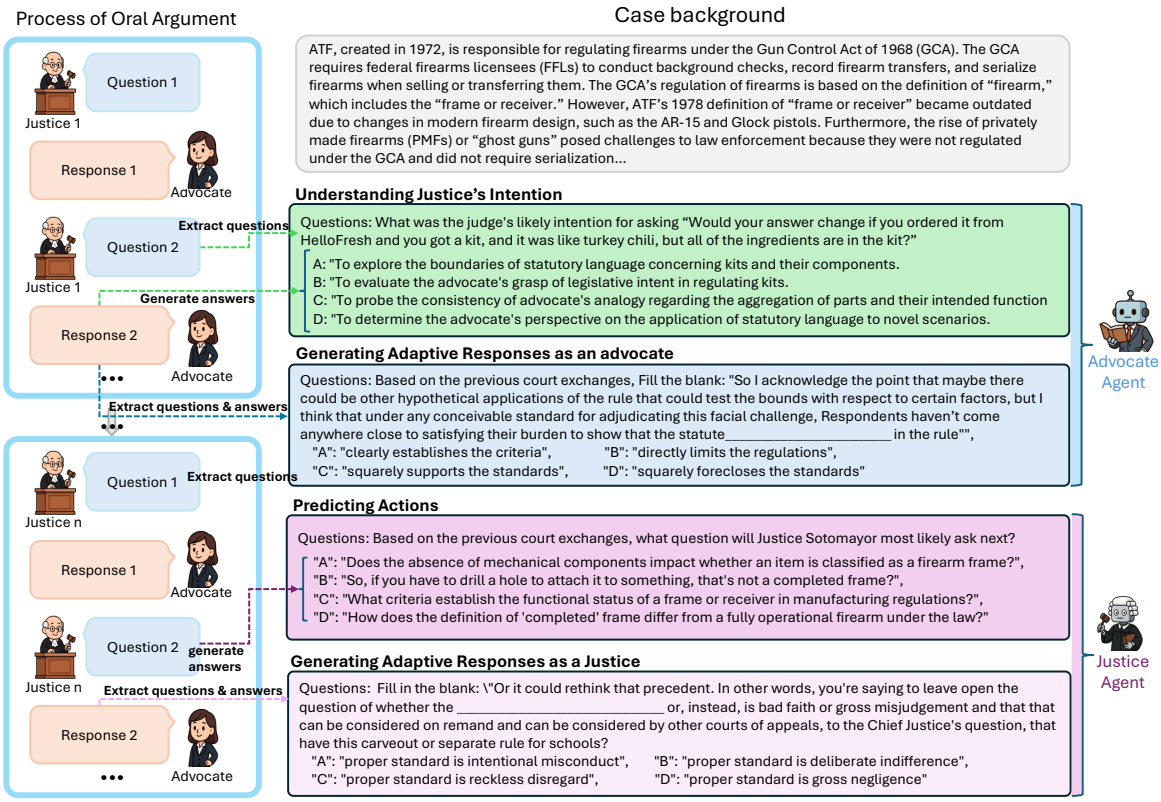


Figure 1: In Oral Arguments, Justices pose strategically framed and challenging questions. Advocates are required to infer the motivation underlying a Justice's question and to construct a reasoned response in defense of their legal position.

model o3.

Our evaluation yields several key insights: (1) Scaling Performance: While performance scales with model size, with larger models capturing nuanced interactions in multi-turn arguments more effectively than smaller models, all architectures struggle with evolving long-range coherence. (2) Contextual Dynamics: Although richer conversational context generally improves results, tracking shifts in legal dialogues remains a significant challenge. (3) Reasoning Capabilities: Reasoning-oriented models, such as o3, demonstrate superior proficiency in inferring the intentions of justices and advocates, whereas non-reasoning models frequently overlook underlying beliefs. (4) Inherent Limitations: The fact that shifts in a justice's intention within extended passages continue to mislead predictions highlights a fundamental deficit in the strategic and social reasoning capabilities of current LLMs.

Our contributions are threefold:

- We formalize Supreme Court oral arguments as a multi-agent legal reasoning environment. Our environment captures the strategic and adversarial nature of real legal interactions.
- We define four ToM-inspired tasks to evaluate LLMs' capacity for perspective-taking and strategic reasoning, including intention inference, question prediction, adaptive response generation for justices and advocates.
- We present `LegalToM`, a benchmark built from authentic U.S. Supreme Court oral arguments, and conduct a comprehensive analysis across eleven LLMs, revealing persistent limitations in ToM-like reasoning.

2 Related Work

LLM about Legal Problems Preliminary research has explored LLMs understanding and reasoning of legal concepts. Benchmarks such as LegalBench (Guha et al., 2023), CaseHOLD (Arvin, 2025), LexGLUE (Chalkidis et al., 2021), LEXTREME (Niklaus et al., 2023), LexEval (Li et al., 2024), and SafeLawBench (Cao et al., 2025) evaluate LLMs across various dimensions, including legal knowledge, reasoning, ethics, multilinguality, and safety. Specialized legal LLMs such as LEGAL-BERT (Chalkidis et al., 2020), ChatLaw (Cui et al., 2023), LawLLM (Shu et al., 2024), and

SaulLM-7B (Colombo et al., 2024) have further advanced domain-specific understanding. Beyond single-agent reasoning, multi-agent legal frameworks such as Debate-Feedback (Chen et al., 2025) and Reflective (Bo et al., 2024) enable agents to provide diverse perspectives through iterative debate and reflection, thereby improving legal argumentation. However, such multi-agent environments do not fully reflect the dynamic, context-dependent, and adversarial nature of real legal interactions, making it difficult to comprehensively evaluate models' abilities.

Benchmarking LLMs' ToM Reasoning Accurately evaluating the ToM capabilities of LLMs requires benchmarks that go beyond simple belief or intention inference. Early datasets such as ToMi (Le et al., 2019) and ToM-b (Nematzadeh et al., 2018) assess reasoning about beliefs and intentions in multi-character narratives, while HiToM (He et al., 2023) introduces higher-order ToM reasoning. Recent studies (Kosinski, 2023, 2024) suggest that models like GPT-4 exhibit emerging ToM-like behaviors comparable to young children, though their generalization to realistic, strategic, and socially complex settings remains limited. Subsequent benchmarks (ToMChallenges (Ma et al., 2023), BigToM (Gandhi et al., 2023), OpenToM (Xu et al., 2024), NegotiationToM (Chan et al., 2024), ToMBench (Chen et al., 2024), and EPITOME (Jones et al., 2023)) introduce richer and more ecologically valid ToM evaluations, incorporating causal templates, negotiation scenarios, and automated grading systems. However, these benchmarks still rely on short, synthetic narratives and fail to capture the strategic dynamics, contextual dependencies, and adversarial reasoning found in real-world social interactions.

3 Oral Argument

3.1 Background of Oral Argument

U.S. Supreme Court oral arguments (Bright, 1986; Wolfson, 2001) play a critical role in the judicial process. As the highest judicial authority, the Court resolves cases with significant legal and societal implications, and its rulings establish binding precedent for lower courts, shaping the interpretation and application of law nationwide. These oral arguments provide a uniquely challenging setting for evaluating LLMs' advanced Theory-of-Mind reasoning capabilities. During arguments, justices carefully craft their questions, while advocates must respond strategically to defend their positions. This dynamic interaction offers an ideal environment to test whether LLMs can uncover latent intentions, predict cognitive state progression, and generate contextually adaptive responses.

U.S. Supreme Court oral arguments follow a structured procedure, where justices pose questions and advocates provide responses. The session usually opens with a statement from the Chief Justice or the presiding justice, formally marking the start of the oral argument. Counsel for the applicants begins by presenting their legal position and supporting rationale. The justices then pose questions intended to probe the arguments, evidence, and legal foundations advanced by counsel. The respondents' counsel subsequently delivers their arguments, which are similarly examined through questions posed by the justices. The advocate must reason about the hidden intention behind the justice's questions and construct a response that defends their position. Figure 1 illustrates this interaction flow. Motivated by this structure, we adopt a ToM formulation to assess whether LLMs can reason about a justice's latent intention, anticipate subsequent actions, and generate contextually appropriate responses in adversarial legal dialogues.

3.2 Three Core Challenges of Oral Argument

Understanding Hidden Intentions LLMs are typically used to process the explicit meaning of a sentence and generate a relevant factual response. However, ToM requires the model to go further; it must reason about the speaker's beliefs, intentions, and knowledge, none of which are explicitly stated in the utterance. In Supreme Court oral arguments, justices rarely ask about intentions explicitly; instead, they pose strategically challenging questions or hypotheticals to test whether an advocate's legal argument remains valid under difficult scenarios or could lead to problematic precedents. Consequently, LLMs must infer the justice's underlying intention rather than respond simply to the literal form of the question.

Predicting Actions To evaluate whether LLMs can function as a professional justice agent, we examine their ability to generate legally precise and strategically probing questions. The LLM must not only formulate questions that test the advocate agent's legal reasoning, but also anticipate how its prompts will be interpreted and responded to. This requires reasoning about the advocate's be-

liefs, intentions, and potential follow-up arguments. By simulating a justice agent in this way, LLMs demonstrate the capacity to perform contextually adaptive questioning, reflecting professional judgment and advanced ToM reasoning ability.

Generating Adaptive Responses The ultimate test of ToM capability lies in generating persuasive and strategically sound responses in high-stakes interactions. We frame the test of LLMs' ToM capabilities in Supreme Court oral arguments in terms of two agent roles: the advocate agent and the justice agent. The advocate agent must produce legally sound, logically consistent, and strategically effective responses, while the justice agent generates coherent and targeted questions that anticipate the advocate's likely replies. the justice agent generates coherent and targeted questions that anticipate the advocate's likely replies.

4 Methodology of LegalToM

4.1 Problem Formulation

We categorize the challenges of oral arguments into three core tasks: understanding intentions, predicting actions, and generating adaptive responses for both the advocate and justice agents. However, since these challenges arise dynamically over the course of long, evolving dialogues, traditional benchmarks are not well suited for their evaluation. To address this, we formalize oral arguments as a multi-agent environment, ensuring that the evolving context and interactive nature of the debate are properly captured.

We formalize U.S. Supreme Court Oral Arguments as a *multi-agent ToM environment*. At each turn t, the dialogue involves two primary roles:

- **Justice** (J): poses strategically framed questions a_t^J, guided by hidden intentions θ_t^J (e.g., probing the generality of a legal principle, testing consistency, and exposing weaknesses).
- **Advocate** (A): responds in real time, aiming to clarify, defend, and persuade by generating the utterance a_t^A

At each step t, $a_t^J \sim \pi_J(\cdot \mid s_t, \theta_t^J)$ and $a_t^A \sim \pi_A(\cdot \mid s_t, a_t^J)$, where a_t^J is the justice's utterance, a_t^A is the advocate's response, π_J is the justice's questioning strategy conditioned on hidden intention θ_t^J, and π_A is the advocate's responding strategy.

The advocate does not have direct access to the justice's latent intention θ_t^J, and must infer it from the dialogue context and the justice's question. Effective advocacy therefore requires reasoning the hidden intention θ_t^J. In addition, the advocate needs to anticipate the justice's likely next question a_{t+1}^J based on the current dialogue. Finally, the advocate must generate a response a_t^A that is both contextually appropriate and persuasive.

4.2 Task Design

We formalize three evaluation tasks using LLM predictions. In the **Intention Understanding** task, the model predicts the justice's hidden intention given the dialogue context s_t and the justice's question a_t^J:

$$\hat{\theta}_t^J = \text{LLM}(s_t, a_t^J),$$

where $\hat{\theta}_t^J$ is the predicted intention generated by the LLM and compared against the labeled ground truth. The justice's true intention θ_t^J is inherently unobservable. In addition, in oral arguments, the advocate (typically the Solicitor General or other highly experienced attorneys) represents the highest level of human ability to interpret a justice's intention. We therefore extract the justice's intention from the advocate's responses and use it as ground truth $\tilde{\theta}_t^J$.

In the **Action Prediction** task, the model anticipates the justice's next question based on the dialogue context, the current justice question, and the advocate's response:

$$\hat{a}_{t+1}^J = \text{LLM}(s_t, a_t^J, a_t^A),$$

where $\hat{a}_{t+1}^J$ is the predicted next question and is evaluated against the actual next question a_{t+1}^J in the transcript.

In the **Adaptive Response Generation** task, the model produces the advocate's response conditioned on the dialogue context and the current justice question:

$$\hat{a}_t^A = \text{LLM}(s_t, a_t^J),$$

where $\hat{a}_t^A$ is the LLM-generated response compared to the human advocate's reference response.

For the justice agent task, the model fills in missing portions of the dialogue to generate legally precise and coherent prompts that reflect proper legal reasoning:

$$\hat{a}_{t+1}^J = \text{LLM}(s_t, a_t^A),$$

where $\hat{a}^J_{t+1}$ represents the LLM-generated justice prompt, formulated based on the dialogue context and the advocate's current response.

Together, these tasks cover the full cycle of ToM reasoning in adversarial dialogue: (i) inferring hidden intentions θ^J_t, (ii) predicting future actions a^J_{t+1}, and (iii) generating adaptive responses a^A_t and a^J_{t+1}. This design provides a structured yet realistic benchmark for probing LLMs' multi-agent reasoning capabilities in high-stakes legal settings.

4.3 ToM Tasks Data Curation from Oral Arguments

Each Supreme Court oral argument case typically lasts one to two hours, resulting in long and complex transcripts that require reasoning over extended contexts. Our goal is to transform real-world oral argument data into a multi-agent setting, where LLM justice agents pose questions and the LLM advocate agent responds based on authentic transcripts. This formulation allows us to simulate realistic courtroom interactions while evaluating the model's ability to reason, predict, and generate contextually appropriate responses. As illustrated in Figure 1, our task construction follows a structured pipeline that converts raw oral argument transcripts into task-specific multiple-choice evaluation instances. We use GPT-4o (Hurst et al., 2024) to assist with content generation during this transformation process.

Task-Specific Data Generation To support the construction of task-specific evaluation benchmarks, we first preprocess the raw case materials by converting case PDFs into Markdown format, enabling structured and efficient downstream processing. The detailed conversion pipeline is provided in the appendix A.1.

After preprocessing, we generate evaluation tasks targeting key challenges for LLM-based advocates: (1) inferring a justice's hidden intention, (2) predicting the justice's next question, and (3) producing contextually adaptive responses for both the advocate and the justice.

For **Understanding Hidden Intentions**, we construct multiple-choice questions that ask for the justice's likely intention. Since true intentions are not directly observable, we assume that expert advocates can reliably infer them from context. Accordingly, the advocate's inferred response is treated as the correct option, while distractors are designed to be plausible alternatives. These distractors mirror the structure and legal framing of the correct answer but are intentionally modified to be incorrect. Concretely, a distractor may adopt a similar legal tone or argumentative role while addressing a different legal issue, misapplying a relevant doctrine, or being inconsistent with the specific context of the dialogue. To validate this assumption, we sample a subset of the data and ask human legal experts to review the LLM-extracted advocate intentions, confirming the correctness of the generated labels. All prompts used for generating questions and answers are provided in Appendix A.3.

For **Predicting Actions**, the model is required to anticipate the justice's next question based on prior dialogue and context. The ground truth answer is the actual subsequent question from the transcript, requiring the model to emulate human-like legal reasoning. Incorrect and misleading options are generated to increase task difficulty. Details of the prompt design are included in Appendix A.3.

For **Generating Adaptive Responses**, we convert expert responses into multiple-choice items by selecting key phrases or legal terms as correct answers. We use LLMs to generate distractor options for evaluation. This tests whether models can identify critical information and produce legally precise, context-aware responses. In addition, we construct two types of response generation tasks: one for advocates and one for justices to assess whether LLMs can successfully function as either side in oral arguments. The advocate generation task tests whether the model can generate persuasive and contextually appropriate replies to justices' questions, while the justice generation task evaluates whether the model can formulate coherent, strategically probing questions or comments that reflect judicial reasoning. The prompts used to generate the questions and corresponding answers are provided in the last part of the appendix A.3.

4.4 Experts Verification

To ensure the reliability of our evaluation, we consulted legal experts to verify the accuracy of LLM-generated content on a subset of the dataset, especially concerning intention extraction. The experts additionally affirmed that our tasks and questions were well-defined and meaningful within the legal context (Seeing Appendix C).

4.5 Evaluation

To evaluate LLMs' abilities on Understanding Hidden Intentions and Predicting Actions, we adopt

Model Names	Understanding Intention			Predicting Actions			Generating Responses for Advocate			Generating Responses for Justice		
	BG	BG+CC	BG+CC+PC	BG	BG+CC	BG+CC+PC	BG	BG+CC	BG+CC+PC	BG	BG+CC	BG+CC+PC
Closed-source Models												
o3(Reasoning)	**84.83**	**85.48**	82.56	**75.36**	88.40	**91.13**	**81.77**	**83.09**	**82.71**	**73.45**	**73.88**	**75.20**
Gpt-4o	82.89	80.86	81.83	60.63	70.27	88.66	69.80	70.72	72.61	57.14	56.13	62.14
Open-source Models												
Qwen2.5_1.5b_instruct	68.21	69.26	65.53	59.71	65.19	60.76	53.81	56.22	59.13	49.34	53.51	53.70
Qwen2.5_7b_instruct	77.45	77.21	78.02	50.20	77.44	75.88	60.10	62.34	63.59	53.77	58.16	59.30
Qwen2.5_72b_instruct	82.32	83.62	**85.89**	45.76	77.44	80.05	**71.86**	**73.44**	**75.31**	**63.30**	**64.23**	**66.51**
Qwen3_4b_instruct	73.15	73.15	75.75	42.11	66.10	64.41	56.85	58.52	62.22	47.80	51.29	52.45
Qwen3_30b_instruct	83.29	83.37	84.27	60.63	83.70	85.01	65.32	66.87	67.94	57.75	60.85	60.98
Gemma3_4b_instruct	64.23	64.23	70.80	29.34	49.28	52.02	46.91	48.86	52.28	40.89	44.75	46.94
Phi-4_14b	82.16	83.13	84.02	38.98	73.53	70.14	66.08	66.49	66.60	60.31	63.79	63.89
Llama3.3_70b_instruct	**84.83**	**85.00**	84.91	**64.93**	**88.79**	**88.79**	67.53	70.12	72.30	62.05	62.91	64.88
Saul_7b_instruct (law)	66.34	66.91	60.02	33.25	51.63	47.72	39.79	38.28	35.27	34.53	36.24	33.02

Table 1: Performance of reasoning, closed-source, and open-source models under three context conditions. **BG** provides the case background, **BG+CC** further includes the current conversation, and **BG+CC+PC** incorporates the preceding three conversations. We evaluate Understanding Intention, Action Prediction, and Response Generation using accuracy (%). **Bold-underlined** indicates the strongest among all models; **bold** indicates the strongest among non-reasoning models.

multiple-choice questions. This setup renders evaluation straightforward, as performance can be directly quantified using accuracy. Since all of our questions are in multiple-choice format, we only need to calculate the accuracy to evaluate how different models perform on our tasks.

5 Experiments

The evaluation comprises four tasks, each probing a distinct aspect of ToM reasoning in legal discourse. The first task focuses on reasoning about hidden intentions, requiring the model to infer the true motivations behind a justice's questions. The second task involves predicting next actions, where the model anticipates the justice's subsequent question based on the preceding dialogue context. The third and fourth tasks examine generating adaptive responses in which the models act as advocates and justices to produce high quality, contextually appropriate replies.

5.1 Experimental Setup

Baseline Models We evaluate GPT-o3 and GPT-4o (Hurst et al., 2024) as a representative closed-source model. GPT-o3 denotes the reasoning model. For open-source models, we assess a range of Instruct models spanning 1.5B to 72B parameters, including Qwen2.5-1.5B-Instruct, Qwen2.5-7B-Instruct, Qwen2.5-72B-Instruct (Yang et al., 2024), Qwen3-4B-Instruct, Qwen3-30B-Instruct (Yang et al., 2025), Gemma3-4B-Instruct (Team et al., 2025), Phi-4 (Abdin et al., 2024), LLaMA3.3-70B-Instruct (Dubey et al., 2024), and Saul-7B-Instruct (Colombo et al., 2024). Among these, Saul-7B-Instruct is specifically fine-tuned on legal texts, optimizing its performance for legal terminology.

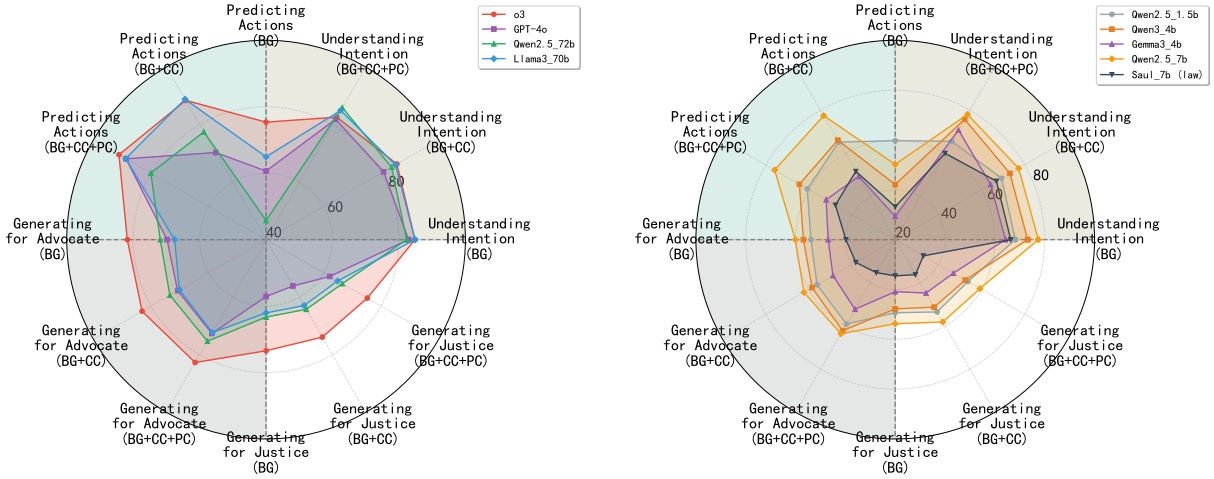
Oral Arguments Dataset We selected 31 oral argument cases covering a diverse range of legal scenarios, including education, criminal and procedural matters, immigration and citizenship, environmental and administrative regulation, labor and employment, commercial and civil disputes, and constitutional and public rights. These cases provide a rich testbed for evaluating both legal reasoning and the reasoning capabilities of LLMs.

From these 31 cases, we generated a dataset comprising 3,884 question-answer pairs across three tasks. The Understanding Hidden Intentions task contains 1,233 questions, Predicting Next Actions includes 767 multiple-choice questions, and Generating Adaptive Responses comprises 1,884 questions, divided between the advocate (970) and the justice (914) roles.

5.2 Experimental Results

This section presents the results of our experiments, evaluating the performance of various LLMs on four distinct tasks. To facilitate analysis, We segmented the long dialogue into a series of smaller conversations, each anchored by the names of justices, to facilitate a comparative evaluation of model performance under two distinct conversational context configurations: a limited condition providing only the immediate background and current turn, and a richer condition including three preceding conversational turns. In Table 1, we divide the context into different types: the background is denoted as **BG**, the current conversation as **CC**, and the preceding three conversations as **PC**.

As shown in the radar chart (Figure 2a), the rea-



(a) Top-tier models comparison

(b) Small models comparison

Figure 2: Performance comparison across all tasks with different contexts (BG, BG+CC, BG+CC+PC).

soning model o3 achieves the best overall performance across all four tasks. We observe that the reasoning model o3 achieves the best overall performance across all four tasks. Among the non-reasoning models, Qwen2.5-72B attains the highest scores in Understanding Intention and Generating Responses for advocate, Llama3-70B performs best in Predicting Actions, and GPT-4o achieves the top score in Generating Responses for justice.

Figure 2b shows that smaller models perform less effectively, with Saul achieving the lowest scores. Although Saul is fine-tuned on legal-domain data, its lower performance indicates that our benchmark primarily evaluates models' ToM reasoning under long-context scenarios rather than domain-specific knowledge.

5.3 Difficulty Classification

We categorized the tasks by difficulty (see Table 2), enabling our benchmark to better establish difficulty levels and evaluate models of varying capabilities. We categorize question difficulty based on the accuracy achieved by different models. Under the hard tasks, different models achieve accuracies ranging from 37% to 62%, providing a set of highly challenging problems for LLMs. Detailed results of three difficulties are presented in Tables 8, 9, and 10.

5.4 Analysis

Conversational Context As shown in Table 1, our experiments reveal a consistent trend: providing models with richer conversational context, including preceding turns, improves their performance. We observe that model performance generally improves as more context is provided, especially for larger models. However, smaller models, such as Qwen2.5-1.5B and Saul-7B, sometimes exhibit degraded performance when additional context is introduced. This decline may be due to their limited capacity to effectively encode and utilize long-range contextual information, which can lead to confusion or interference from irrelevant details, thereby reducing overall task accuracy.

Task	Easy	Medium	Hard	Total
Advocate Keywords	324	323	323	970
Justice Keywords	307	305	305	917
Intention Analysis	369	554	310	1,233
Prediction Task	230	345	192	767
Total	1,230	1,527	1030	3,887

Table 2: Task difficulty statistics

Effect of Model Size Our analysis on the oral arguments dataset reveals that the Qwen2.5 series follows a clear scaling law. Performance consistently improves with model size. Seeing Figure 3, this strong trend indicates that larger parameter counts and greater overall capacity are crucial for capturing the complex, nuanced information required by these tasks. Notably, smaller models, such as Qwen2.5-1.5B, may struggle with longer contextual dependencies, highlighting the importance of both scale and context utilization in multi-turn reasoning.

Reasoning Model As a reasoning model, o3 more accurately captures judges' and advocates'

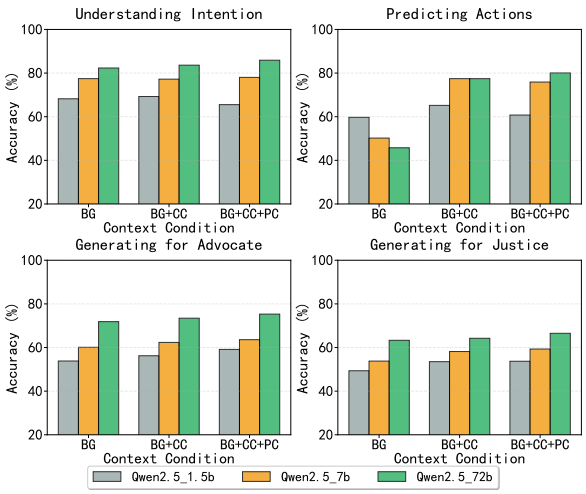


Figure 3: Qwen2.5 models follow a scaling law: performance consistently improves as model size increases from 1.5B to 72B, with larger models benefiting more from richer context.

subtle beliefs and intentions in long-context dialogues, enabling it to make more accurate predictions about their likely next moves. For example, in the item shown in Table 6, nearly all non-reasoning models failed, whereas o3 correctly inferred that *Justice Sotomayor had just pointed out that "intent is not even an issue here,"* and anticipated that her next question would refocus the argument on the matter for which the Court granted certiorari (see Table 4 for details).

In contrast, non-reasoning models such as Qwen2.5-72B did not effectively capture the justice's beliefs regarding the previous question. Qwen2.5-72B focused *primarily on the definitions and implications of "bad faith" and "intent to discriminate," and on the need to further clarify how the standards for intent apply differently to various forms of relief, such as damages versus injunctive relief* (see Table 5). Because it failed to account for Justice Sotomayor's underlying concern about the specific legal question, Qwen2.5-72B produced an incorrect prediction.

Questions All Agents Fail In `LegalToM`, several highly challenging cases remained unsolved by all models, including the reasoning model o3. These failures occurred when models could not track shifts in a justice's intention during long explanatory passages, instead over-focusing on surface details, which led to a misunderstanding of the hidden intention. As shown in Table 7, although Justice Sotomayor devoted a lengthy discussion to the relationship between *mens rea and intent*, she began by stating, *"Motive, in — intent — we don't care about motive,"* indicating that motive and intent were not her primary concern. Nevertheless, even the most capable o3 model was misled by her extended exposition and concluded that her aim was to *"Evaluate the advocate's interpretation of the statutory language regarding intent requirements."*. This example illustrates the challenges LLMs face in reasoning over long, nuanced legal discourse and accurately inferring hidden intentions.

6 Conclusion

We introduce a novel multi-agent framework to evaluate LLMs' ToM reasoning in the complex, adversarial setting of U.S. Supreme Court oral arguments, formalizing three core tasks: understanding hidden intentions, predicting subsequent questions, and generating contextually adaptive responses. Using this testbed, we find that performance scales with model size, larger models better capture multi-turn interactions, and reasoning-oriented models such as o3 more effectively infer justices' and advocates' intentions. Nevertheless, all models struggle with evolving intentions and long-range coherence, and shifts in a justice's intention within extended passages reveal fundamental limitations in strategic and social reasoning, highlighting challenges for developing context-aware, multi-agent legal reasoning LLMs.

7 Limitations

While our study demonstrates the benefits of providing richer conversational context in multi-turn legal dialogues, several limitations remain. First, our dataset, although drawn from real oral argument dialogues, is still limited and may not fully capture the diversity and complexity of real-world legal proceedings. Importantly, even though the justices and advocates are highly professional, errors can still occur during their exchanges. Since we take the justices' and advocates' dialogues as the ground-truth labels for our tasks, this may introduce occasional inaccuracies. Second, despite improvements in handling long contexts, models continue to face challenges in capturing deeper reasoning and nuanced ToM aspects, such as inferring judges' beliefs, intentions, and strategic reasoning in complex interactions.

8 Ethical Considerations

This research complies with the ACL Ethics Policy. All data used in this study are drawn from publicly available U.S. Supreme Court oral argument transcripts. A subset of the data, including LLM-extracted intentions, was manually verified by legal experts to ensure accuracy. All data contain no private or personally identifiable information. The dialogues analyzed represent official judicial proceedings and are used solely for academic research purposes. Our work aims to advance understanding of language models' reasoning and Theory of Mind capabilities in legal contexts, without attempting to replace or simulate actual judicial decision-making. No data manipulation, fabrication, or misrepresentation was involved. We also ensured that model outputs and analyses were reviewed for fairness and interpretability to avoid potential bias or misuse in real-world legal applications.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Chuck Arvin. 2025. Identifying legal holdings with llms: A systematic study of performance, scale, and memorization. *arXiv preprint arXiv:2505.02172*.

Xiaohe Bo, Zeyu Zhang, Quanyu Dai, Xueyang Feng, Lei Wang, Rui Li, Xu Chen, and Ji-Rong Wen. 2024. Reflective multi-agent collaboration based on large language models. *Advances in Neural Information Processing Systems*, 37:138595–138631.

Myron H Bright. 1986. The power of the spoken word: In defense of oral argument. *Iowa L. Rev.*, 72:35.

Chuxue Cao, Han Zhu, Jiaming Ji, Qichao Sun, Zhenghao Zhu, Yinyu Wu, Juntao Dai, Yaodong Yang, Sirui Han, and Yike Guo. 2025. Safelawbench: Towards safe alignment of large language models. *arXiv preprint arXiv:2506.06636*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024. Negotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding. *arXiv preprint arXiv:2404.13627*.

Xi Chen, Mao Mao, Shuo Li, and Haotian Shangguan. 2025. Debate-feedback: A multi-agent framework for efficient legal judgment prediction. *arXiv preprint arXiv:2504.05358*.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and 1 others. 2024. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and 1 others. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.

Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Timothy R Johnson, Paul J Wahlbeck, and James F Spriggs. 2006. The influence of oral arguments on the us supreme court. *American Political Science Review*, 100(1):99–113.

Cameron Robert Jones, Sean Trott, and Ben Bergen. 2023. Epitome: Experimental protocol inventory for theory of mind evaluation. In *First Workshop on Theory of Mind in Communicating Agents*.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *Advances in Neural Information Processing Systems*, 37:25061–25094.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. 2023. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. *arXiv preprint arXiv:2305.15068*.

AI Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 20:2024.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L Griffiths. 2018. Evaluating theory of mind in question answering. *arXiv preprint arXiv:1808.09352*.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. *arXiv preprint arXiv:2301.13126*.

Oyez, Inc. 2025. Oyez: U.s. supreme court. The Oyez Project.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the us legal system. In *Proceedings of the 33rd ACM International Conference on information and knowledge management*, pages 4882–4889.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Warren D Wolfson. 2001. Oral argument: Doe it matter. *Ind. L. Rev.*, 35:451.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv preprint arXiv:2402.06044*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

A Further Details on Generating Questions

A.1 Method for Preprocessing

We obtain the oral argument transcripts (Oyez, Inc., 2025) in PDF format and preprocess them by converting the PDFs into Markdown format. This preprocessing segments the conversations between individual justices and advocates into structured blocks, as illustrated in Listing 1.

Listing 1: Example of segmented oral argument dialogue in Markdown format

```
## Justice Thomas
### Justice Thomas:
Under your approach, what would that correct
    standard look like and how would it be
    applied here?
### Nathaniel A.G. Zelinsky:
Justice Thomas, we think the standard is the "
    totality of the circumstances" standard that
     this Court articulated in Graham and Garner
    , Scott, and Plumhoff.
...
## Justice Kavanaugh
### Justice Kavanaugh:
Was it reasonable to --for the officer to jump
    on the side of the car?
### Nathaniel A.G. Zelinsky:
So, Just -- Justice Kavanaugh, we don't think it
     was in this particular case, but that's
    precisely the issue that the lower courts
    couldn't evaluate because they applied this
    legal amnesia and only look at the fact that
     the officer was on the moving vehicle.
...
```

A.2 Method for Generating Questions

For the first task, **Understanding Hidden Intentions**, we construct multiple-choice questions that ask for the justice's intention behind a given question, for example: *"What was the justice's likely intent in asking this question?"* followed by the verbatim text of the justice's question. For generating correct and distract options in multiple-choice questions, there is no definitive ground truth that can conclusively determine a justice' intentions. In the dialogue, justices and advocates merely discuss the specifics of a case, and the hidden intention behind a justice's question remains concealed, making it impossible to observe directly. Nevertheless, both the justices and the advocates participating in oral arguments are highly specialized experts in the legal domain. The justices are members of the Supreme Court, and the advocates are either representatives of the Solicitor General's office or senior attorneys. Advocates engage in thorough preparation before oral arguments and bring substantial legal expertise to the proceedings. Based on this, our approach is as follows: we assume that the advocate can infer the justice's intention.

For the second task, **Predicting Actions**, the question is formulated as: *"Based on the previous court exchanges, what question will the justice most likely ask next?"* The Justice's next question can serve as the ground truth, so the correct answer is the next question extracted directly from the transcript. LLMs must integrate prior judicial questions, advocates' responses, and the broader legal context. This task goes beyond surface-level text generation, aiming instead to emulate human-like legal reasoning. The model is required to anticipate the legal or factual issues that a justice is most likely to focus on next.

For the final task, **Generating Adaptive Responses**, we assume that the advocate and justice are sufficiently expert to provide professional and reasonable answers based on human knowledge. We then convert the advocate's and justice's responses into multiple-choice questions by selecting key phrases or legal terms as the correct answers and creating plausible distractors. This setup allows us to evaluate whether models can identify and select critical information, articulate core arguments, and use appropriate legal terminology in context, rather than assessing the overall coherence of long form responses.

A.3 Prompts for Generating Questions

To generate questions for assessing understanding of intentions, we employ the following prompt:

```
You are an expert in legal dialogue analysis.
    Your task is to evaluate
whether a dialogue segment contains substantive
    judicial interactions worth analyzing.

BE VERY STRICT: Only approve dialogues that
    reveal significant judicial reasoning.

IMMEDIATELY REJECT if the dialogue contains:
- Simple "Yes/No" answers or basic confirmations
- Pure procedural exchanges ("Thank you, Your
    Honor")
- Brief interruptions or clarifications
- Administrative housekeeping matters
- Single question without substantial response
- Repetitive or redundant content

ONLY APPROVE if the dialogue demonstrates:
1. Complex judicial questioning with strategic
    purpose
2. Substantial advocate responses revealing
    judicial intent
3. Clear evidence of deep legal reasoning or
    analysis
```

```
4. Multiple meaningful exchanges showing judicial thought process
5. Significant substantive legal content worth studying

Return a JSON object with:
{
  "generate_questions": true/false,
  "complexity_score": 1-10,
  "reason": "explanation for decision",
  "key_interactions": ["list of important judicial questions"]
}

"You are a SUPREME COURT CLERK creating IMPOSSIBLY DIFFICULT questions that would challenge the finest legal minds. "
"Your task is to select the 3 most MICROSCOPICALLY DECEPTIVE incorrect options that would fool Supreme Court justices themselves. "
"ELITE SELECTION CRITERIA - Choose options that are:\n"
"1. INDISTINGUISHABLE from correct answers even to constitutional law professors\n"
"2. Require GENIUS-LEVEL understanding of judicial psychology, precedent, and micro-contextual analysis to eliminate\n"
"3. Use IDENTICALLY sophisticated legal terminology, sentence structure, and analytical depth as correct answer\n"
"4. Represent judicial strategies so nuanced they differ by single words or microscopic timing assumptions\n"
"5. Test distinctions so fine that only perfect understanding of this specific justice's psychology can resolve them\n"
"6. Would be considered correct by 95% of federal judges and 90% of constitutional law professors\n"
"7. Differ from the correct answer by virtually undetectable nuances that require PhD-level legal theory to spot\n\n"
"ABSOLUTE REJECTIONS - NEVER select options that:\n"
"- Can be eliminated by anyone short of a Supreme Court justice or top-tier law professor\n"
"- Use even slightly different levels of sophistication, complexity, or terminology\n"
"- Would be obviously wrong to federal appellate judges\n"
"- Lack the surgical precision and microscopic accuracy of the correct answer\n\n"
"Return only the 3 most challenging distractors, separated by commas."
```

To generate questions for Predicting Actions, we employ the following prompt:

```
"You are an expert in legal dialogue analysis. Your task is to evaluate "
"whether a dialogue segment contains substantive judicial interactions worth analyzing.\n\n"

"BE VERY STRICT: Only approve dialogues that reveal significant judicial reasoning.\n\n"

"IMMEDIATELY REJECT if the dialogue contains:\n"
"- Simple \"Yes/No\" answers or basic confirmations\n"
"- Purely procedural exchanges (scheduling, format, etc.)\n"
"- Administrative comments without legal substance\n"
"- Repetitive or redundant exchanges\n"
"- Greetings, thanks, or basic courtesies\n\n"

"APPROVE ONLY if the dialogue contains:\n"
"- Complex legal reasoning or argumentation\n"
"- Judicial probing of legal concepts or precedents\n"
"- Detailed factual or procedural discussions\n"
"- Strategic legal positioning by advocates\n"
"- Constitutional or statutory interpretation\n"
"- Multi-part judicial questioning revealing judicial reasoning\n\n"

"Rate complexity on a scale of 1-10:\n"
"1-3: Simple/procedural (REJECT)\n"
"4-6: Moderate complexity (CONDITIONAL)\n"
"7-10: High complexity (APPROVE)\n\n"

"Return JSON format:\n"
"{\n"
"  \"generate_questions\": true/false,\n"
"  \"complexity_score\": 1-10,\n"
"  \"reason\": \"Brief explanation\",\n"
"  \"key_interactions\": [\"list of important exchanges or concepts\"]\n"
"}"

"You are creating incorrect alternative questions for a multiple choice test about predicting judicial questions. "
"Your goal is to make wrong answers that are EASY to distinguish (target 70% correct rate).\\n\\n"

"The fake questions should be:\\n"
"- Most should be about obviously different legal topics or concepts\\n"
"- Many should be generic questions that could be asked anytime\\n"
"- Some should be about different stages or timing that don't fit the current flow\\n"
"- Include procedural questions that don't match the immediate context\\n"
"- Make the majority clearly distinguishable from the correct answer\\n\\n"

"Make these distractors EASY to distinguish - they should be obviously wrong "
"when you consider what was just discussed and what would logically follow.\\n\\n"

"Return JSON format:\\n"
'{\\n'
'  "possible_questions": ["question1", "question2", ... "question10"]\\n'
'}'
```

To generate Adaptive Response Generation questions for advocate, we employ the following prompt:

```
"You are an expert in legal dialogue analysis. Your task is to evaluate "
"whether a dialogue segment contains substantive
```

```
     judicial interactions worth analyzing.\n\n"

"BE VERY STRICT: Only approve dialogues that
    reveal significant judicial reasoning.\n\n"

"IMMEDIATELY REJECT if the dialogue contains:\n"
"- Simple \"Yes/No\" answers or basic
    confirmations\n"
"- Purely procedural exchanges (scheduling,
    format, etc.)\n"
"- Administrative comments without legal
    substance\n"
"- Repetitive or redundant exchanges\n"
"- Greetings, thanks, or basic courtesies\n\n"

"APPROVE ONLY if the dialogue contains:\n"
"- Complex legal reasoning or argumentation\n"
"- Judicial probing of legal concepts or
    precedents\n"
"- Detailed factual or procedural discussions\n"
"- Strategic legal positioning by advocates\n"
"- Constitutional or statutory interpretation\n"
"- Multi-part judicial questioning revealing
    judicial reasoning\n\n"

"Rate complexity on a scale of 1-10:\n"
"1-3: Simple/procedural (REJECT)\n"
"4-6: Moderate complexity (CONDITIONAL)\n"
"7-10: High complexity (APPROVE)\n\n"

"Return JSON format:\n"
"{\n"
"  \"generate_questions\": true/false,\n"
"  \"complexity_score\": 1-10,\n"
"  \"reason\": \"Brief explanation\",\n"
"  \"key_interactions\": [\"list of important
    exchanges or concepts\"]\n"
"}"
```

To generate Adaptive Response Generation questions for justice, we employ the following prompt:

```
"You are an expert in legal dialogue analysis.
    Your task is to evaluate "
"whether a dialogue segment contains substantive
     judicial interactions worth analyzing.\n\n"

"BE VERY STRICT: Only approve dialogues that
    reveal significant judicial reasoning.\n\n"

"IMMEDIATELY REJECT if the dialogue contains:\n"
"- Simple \"Yes/No\" answers or basic
    confirmations\n"
"- Purely procedural exchanges (scheduling,
    format, etc.)\n"
"- Administrative comments without legal
    substance\n"
"- Repetitive or redundant exchanges\n"
"- Greetings, thanks, or basic courtesies\n\n"

"APPROVE ONLY if the dialogue contains:\n"
"- Complex legal reasoning or argumentation\n"
"- Judicial probing of legal concepts or
    precedents\n"
"- Detailed factual or procedural discussions\n"
"- Strategic legal positioning by advocates\n"
"- Constitutional or statutory interpretation\n"
"- Multi-part judicial questioning revealing
    judicial reasoning\n\n"

"Rate complexity on a scale of 1-10:\n"
"1-3: Simple/procedural (REJECT)\n"
"4-6: Moderate complexity (CONDITIONAL)\n"
"7-10: High complexity (APPROVE)\n\n"

"Return JSON format:\n"
"{\n"
"  \"generate_questions\": true/false,\n"
"  \"complexity_score\": 1-10,\n"
"  \"reason\": \"Brief explanation\",\n"
"  \"key_interactions\": [\"list of important
    exchanges or concepts\"]\n"
"}"
```

We generate multiple misleading incorrect options and then select the three most effective ones to increase the difficulty of the questions.

B Further Details on Answering Questions

To answer questions for understanding of intentions, we employ the following prompt:

```
"You are an expert in legal language analysis
    and advocate statement comprehension. "
"You will be given a legal case background,
    dialogue context, and a multiple-choice
    question "
"about identifying key legal terms, phrases, or
    concepts from advocate statements.\n\n"

"Your task is to:\n"
"1. Carefully analyze the case background and
    dialogue context\n"
"2. Focus on the specific advocate statement
    provided\n"
"3. Identify which legal concept, term, or
    phrase is most relevant to the advocate's
    argument\n"
"4. Consider the legal significance and context
    of each option\n"
"5. Select the option that best captures the key
     legal point being made\n\n"

"Focus on:\n"
"- Core legal concepts and principles\n"
"- Important statutory references and legal
    authorities\n"
"- Procedural requirements and legal standards\n
    "
"- Case law principles and precedents\n"
"- Constitutional provisions and interpretations
    \n"
"- Technical legal terminology essential to the
    argument\n\n"

"Please provide your reasoning and end with your
     final answer in the format \\boxed{{answer
    }}."
```

To answer questions for predicting actions, we employ the following prompt:

```
"You are an expert in judicial questioning
    patterns and legal dialogue analysis. "
```

```
"You will predict what question a Justice is
    most likely to ask next based on the context
     provided.\n\n"

"Your task is to:\n"
"1. Analyze the cumulative context and dialogue
    flow\n"
"2. Consider the Justice's established
    questioning patterns and priorities\n"
"3. Evaluate the logical progression of the
    legal discussion\n"
"4. Predict what information or clarification
    the Justice would most likely seek next\n"
"5. Use your understanding of judicial behavior
    and legal procedure\n"
"6. Select the most likely next question from
    the given options\n\n"

"Focus on:\n"
"- Natural flow of judicial inquiry\n"
"- Unresolved issues that need clarification\n"
"- The Justice's apparent concerns and
    priorities\n"
"- Standard legal questioning protocols\n"
"- Follow-up questions that logically flow from
    the current discussion\n\n"

"Please provide your reasoning and end with your
     final answer in the format \\boxed{{answer
    }}."
```

To answer Adaptive Response Generation questions for advocate, we employ the following prompt:

```
"You are an expert in legal language analysis
    and advocate statement comprehension. "
"You will be given a legal case background,
    dialogue context, and a multiple-choice
    question "
"about identifying key legal terms, phrases, or
    concepts from advocate statements.\n\n"

"Your task is to:\n"
"1. Carefully analyze the case background and
    dialogue context\n"
"2. Focus on the specific advocate statement
    provided\n"
"3. Identify which legal concept, term, or
    phrase is most relevant to the advocate's
    argument\n"
"4. Consider the legal significance and context
    of each option\n"
"5. Select the option that best captures the key
     legal point being made\n\n"

"Focus on:\n"
"- Core legal concepts and principles\n"
"- Important statutory references and legal
    authorities\n"
"- Procedural requirements and legal standards\n
    "
"- Case law principles and precedents\n"
"- Constitutional provisions and interpretations
    \n"
"- Technical legal terminology essential to the
    argument\n\n"

"Please provide your reasoning and end with your
     final answer in the format \\boxed{{answer
    }}."
```

To answer Adaptive Response Generation questions for justice, we employ the following prompt:

```
"You are an expert in legal language analysis
    and judicial questioning comprehension. "
"You will be given a legal case background,
    dialogue context, and a multiple-choice
    question "
"about identifying key legal terms, phrases, or
    concepts from judicial questions.\n\n"

"Your task is to:\n"
"1. Carefully analyze the case background and
    dialogue context\n"
"2. Focus on the specific judicial question
    provided\n"
"3. Identify which legal concept, term, or
    phrase is most relevant to the justice's
    inquiry\n"
"4. Consider the legal significance and context
    of each option\n"
"5. Select the option that best captures the key
     legal point being made\n\n"

"Focus on:\n"
"- Core legal concepts and principles\n"
"- Important statutory references and legal
    authorities\n"
"- Procedural requirements and legal standards\n
    "
"- Case law principles and precedents\n"
"- Constitutional provisions and interpretations
    \n"
"- Technical legal terminology essential to the
    argument\n\n"

"Please provide your reasoning and end with your
     final answer in the format \\boxed{{answer
    }}."
```

C Analysis of Legal Experts

The correct answer is C of the Question 3, and this question is interesting precisely because it exposes the pressure point of Supreme Court advocacy.

Justice Jackson isn't quibbling over wording for sport. She's performing a jurisdiction-defining move. When she asks what the question "presented in this case actually is," she's signaling something fundamental: the Court can only decide what is properly before it. Everything else-policy arguments, doctrinal cleanup, institutional anxiety-is noise unless tethered to that question.

1. Drawing a hard boundary around the Court's role: Supreme Court justices are acutely sensitive to overreach. By drilling into the Question Presented, Justice Jackson is asking whether the Court is being asked to decide whether a uniquely stringent standard applies at all, or merely whether this

particular standard (bad faith or gross misjudgment) qualifies as such. That distinction determines the scope of the ruling—narrow case resolution versus sweeping doctrinal shift.

2. Exposing strategic ambiguity in advocacy: Lisa Blatt attempts to present two readings of the petition. Justice Jackson's questioning highlights that advocates cannot argue narrow when their briefing argues broad. This reveals how credibility and consistency are essential currencies before the Court.

3. Framing everything that follows: Once the Court locks in what the question is, entire lines of argument either survive or collapse—remedies, intent standards, and statutory interpretation all depend on it.

Justice Jackson's question is interesting because it is a gatekeeping move. It forces clarity, constrains overreach, and reinforces that the Supreme Court is a court of limited questions with lasting consequences.

Case Name: A.J.T. v. Osseo Area Schools, Independent School District
Question: question: "What was the judge's likely intent for asking 'presented in this case actually is.'?" A: "To probe the consistency of the advocate's argument with previous case law.", B: "To explore the broader implications of the legal standards being debated.", C: "To clarify the specific legal question that the court needs to address in this case.", D: "To challenge the advocate's interpretation of the legal standards involved." **Answer**: "C

Table 3: Example Analyzed by Legal Experts

Case Name: A.J.T. v. Osseo Area Schools, Independent School District
Model Name: o3
Predicted Answer: A
Explanation: Justice Sotomayor has just pointed out that intent "is not even an issue here" because the family already obtained IDEA relief, and counsel responded, "They want more." A natural follow-up for Justice Sotomayor is to refocus the argument on what the Court actually granted certiorari to decide. She is likely to press counsel to confirm that, notwithstanding whatever additional relief the plaintiffs might desire, the only legal question properly before the Court is whether the Eighth Circuit applied the correct intent (versus heightened-intent) standard under Section 504/ADA. Option A mirrors that clarification: it brackets counsel's "they want more" remark and zeroes in on whether the dispute before the Court is simply the level of intent required. The other options either broaden the discussion too abstractly (B, D) or divert it to remedies under IDEA (C), which is not what she has just highlighted.

Table 4: OpenAI o3 Model Output Example

D Representative Dialogue Cases

We provide a representative oral argument case, as shown in Table 6. The table includes a portion of the preceding context from the argument and the corresponding question generated by our justice agent. This illustrates how our method captures the legal dialogue and generates contextually relevant questions.

E Details of Difficulty Levels

Case Name: A.J.T. v. Osseo Area Schools, Independent School District
Model Name: Qwen2.5-72B
Predicted Answer: B
Explanation: Based on the previous court exchanges and the specific context provided, Justice Sotomayor is most likely to ask:**Option B: Is there a distinction between seeking damages and obtaining injunctive relief under the IDEA?** Reasoning: 1. **Natural Flow of Judicial Inquiry**: Justice Sotomayor has been focusing on the definitions and implications of "bad faith" and "intent to discriminate." She is concerned about the broader application of these concepts across different statutes. 2. **Unresolved Issues**: There is a need to clarify how the standards for intent apply differently to various forms of relief, such as damages versus injunctive relief. 3. **Apparent Concerns and Priorities**: Justice Sotomayor has shown a concern for the procedural and definitional clarity in discrimination cases, particularly in the context of the IDEA. 4. **Standard Legal Questioning Protocols**: It is common for justices to explore the distinctions between different types of legal remedies to ensure that the court's decision is well-grounded and consistent with existing legal principles. Thus, Option B aligns best with the natural progression of the legal discussion and Justice Sotomayor's apparent concerns.

Table 5: Qwen2.5-72B Model Output Example

Court Dialogue

...

Justice Sotomayor: But they didn't pass the policy with antireligion animus.

Lisa S. Blatt: If you – let me just give you another example.

Justice Sotomayor: You're asking – when you're using the words "bad faith," you're talking about animus.

Lisa S. Blatt: No, I'm talking about — and you can — you're in charge, so you can say: Intent to discriminate is the standard. We're not going to use bad faith. We don't like that word...

Justice Sotomayor: Counsel, it would have been nice to have known that we were biting off that big a chunk.

Lisa S. Blatt: I agree. But in terms of what we had to do when you granted cert was look at the text, and then the blue brief said that there is no intent required.

Justice Sotomayor: By the way, intent's not even an issue here because there wasn't an injunction being — or the lack of an injunction — challenged here. They got the injunction under the IDEA, didn't they?

Lisa S. Blatt: They want more.

Question for Justice Agent:

Question: Based on the preceding exchange, what question will Justice Sotomayor most likely ask next?

A. Well, we can put aside whether they want more. But the only thing before us on the decision below is whether it's an intent standard or a heightened standard, correct?
B. In determining the appropriate standard for this case, should we consider whether courts must require proof of intentional discrimination instead of negligence or indifference?
C. Should we focus on the adequacy of the remedies available under the IDEA, rather than changes to the intent standard?
D. How does the interpretation of intent in discrimination cases differ between educational and employment contexts under Title VII?

Correct Answer: A

Table 6: Court Dialogue and Predicting Actions Example

Court Dialogue

...

Justice Sotomayor: I'm not sure where any of these tests come from, because mens rea is generally willfulness, which requires knowing what the law is, but the statute doesn't talk about willfulness. Motive, in – intent – we don't care about motive. We've said that repeatedly in a bunch of different contexts. It's do you know you're doing the act and are you intending to do the act. If it violates the law, you're guilty. Pardon the pun. This is a tort, but you're responsible. Or you do it knowingly, knowing that you're doing the act. So I don't know where the bad faith comes from. I don't know where the gross indifference comes from. I don't know where the deliberate indifference comes from. Have you figured that out?

Question for Advocate Agent:

Question: What was the judge's likely intention for asking 'I'm not sure where any of these tests come from, because mens rea is generally willfulness, which requires knowing what the law is, but the statute doesn't talk about willfulness. Motive, in – intent – we don't care about motive. We've said that repeatedly in a bunch of different contexts. It's do you know you're doing the act and are you intending to do the act. If it violates the law, you're guilty. Pardon the pun. This is a tort, but you're responsible. Or you do it knowingly, knowing that you're doing the act. So I don't know where the bad faith comes from. I don't know where the gross indifference comes from. I don't know where the deliberate indifference comes from. Have you figured that out?'?

A. Clarify the advocate's stance on the interplay between statutory causation requirements and mens rea.
B. Assess the advocate's understanding and articulation of the statutory basis for different mens rea standards within disability discrimination law.
C. Probe the advocate's ability to distinguish between statutory interpretation and judicially created tests.
D. Evaluate the advocate's interpretation of the statutory language regarding intent requirements.

Correct Answer: B

Table 7: Court Dialogue and Understanding Intention Example

Model Names	Understanding Intention			Predicting Actions			Generating Responses for Advocate			Generating Responses for Justice		
	BG	BG+CC	BG+CC+PC	BG	BG+CC	BG+CC+PC	BG	BG+CC	BG+CC+PC	BG	BG+CC	BG+CC+PC
Closed-source Models												
o3 (Reasoning)	**94.82**	93.53	92.35	**77.09**	89.26	**91.23**	**88.43**	**89.31**	**87.33**	**83.47**	**82.61**	**84.27**
GPT-4o	92.06	89.71	90.70	62.72	72.67	88.37	76.66	76.53	77.85	73.04	71.06	75.53
Open-source Models												
Qwen2.5_1.5b_instruct	82.35	83.53	88.53	60.66	66.03	62.72	68.42	70.97	71.54	64.54	66.17	64.89
Qwen2.5_7b_instruct	89.12	89.12	91.76	51.82	78.83	77.41	75.38	75.33	74.38	71.97	72.55	72.77
Qwen2.5_72b_instruct	90.29	92.35	**96.18**	48.50	78.20	80.57	**83.46**	**85.77**	**84.82**	**78.34**	**78.30**	**79.15**
Qwen3_4b_instruct	86.18	86.76	90.29	45.02	67.14	66.19	72.57	71.24	72.79	67.09	68.30	67.45
Qwen3_30b_instruct	94.41	**94.71**	95.00	63.35	84.20	85.47	78.32	78.54	78.32	75.80	78.09	76.81
Gemma3_4b_instruct	76.76	76.76	89.12	32.07	52.13	54.82	62.59	61.86	63.19	57.11	58.51	59.36
Phi-4	91.76	91.76	95.88	42.65	74.25	71.41	78.20	77.99	77.61	76.43	77.66	77.66
Llama3.3_70b_instruct	94.71	**94.71**	**96.18**	**67.14**	**89.42**	**89.10**	79.32	79.70	80.65	77.49	77.02	78.30
Saul_7b_instruct	79.41	81.18	72.35	35.07	52.13	49.45	53.57	48.77	40.99	43.52	42.55	36.60

Table 8: Performance on **EASY** difficulty questions. **BG** provides the case background, **BG+CC** further includes the current conversation, and **BG+CC+PC** incorporates the preceding conversations. We evaluate Understanding Intention, Action Prediction, and Response Generation using accuracy (%). **Bold-underlined** indicates the strongest among all models; **bold** indicates the strongest among open-source models.

Model Names	Understanding Intention			Predicting Actions			Generating Responses for Advocate			Generating Responses for Justice		
	BG	BG+CC	BG+CC+PC	BG	BG+CC	BG+CC+PC	BG	BG+CC	BG+CC+PC	BG	BG+CC	BG+CC+PC
Closed-source Models												
o3 (Reasoning)	86.06	86.49	84.31	**75.88**	89.14	**91.44**	**86.09**	**87.34**	**86.63**	**75.48**	**76.74**	**78.29**
GPT-4o	85.40	83.22	77.94	60.23	70.38	89.22	72.96	75.05	75.50	60.36	58.99	65.25
Open-source Models												
Qwen2.5_1.5b_instruct	67.97	69.93	61.87	58.67	64.46	60.51	52.69	56.13	58.55	49.70	54.75	55.56
Qwen2.5_7b_instruct	81.26	80.83	84.97	49.79	77.86	75.74	63.45	67.29	67.29	54.33	59.60	61.41
Qwen2.5_72b_instruct	86.49	**88.45**	**93.03**	44.85	77.86	80.68	**76.81**	**77.88**	**80.11**	**70.02**	**70.51**	**71.92**
Qwen3_4b_instruct	76.91	77.34	79.96	41.47	65.59	63.61	58.97	61.54	65.17	46.88	52.12	52.93
Qwen3_30b_instruct	**88.45**	**88.45**	89.76	60.51	83.64	84.91	72.44	72.01	71.79	62.58	64.24	65.45
Gemma3_4b_instruct	61.22	61.22	71.46	28.63	49.22	52.19	45.64	48.14	54.09	37.63	42.63	45.86
Phi-4	84.97	86.27	89.11	37.80	73.77	70.38	72.73	72.86	71.19	65.79	68.08	68.48
Llama3.3_70b_instruct	86.71	86.93	88.02	**64.60**	**89.28**	**89.28**	73.47	74.91	77.88	67.81	68.28	69.09
Saul_7b_instruct	67.32	67.54	59.04	32.86	51.48	47.25	36.92	36.25	36.06	33.60	35.96	34.34

Table 9: Performance on **MEDIUM** difficulty questions. **BG** provides the case background, **BG+CC** further includes the current conversation, and **BG+CC+PC** incorporates the preceding conversations. We evaluate Understanding Intention, Action Prediction, and Response Generation using accuracy (%). **Bold-underlined** indicates the strongest among all models; **bold** indicates the strongest among open-source models.

Model Names	Understanding Intention			Predicting Actions			Generating Responses for Advocate			Generating Responses for Justice		
	BG	BG+CC	BG+CC+PC	BG	BG+CC	BG+CC+PC	BG	BG+CC	BG+CC+PC	BG	BG+CC	BG+CC+PC
Closed-source Models												
o3 (Reasoning)	73.14	71.90	68.18	**70.22**	84.93	**88.89**	**73.53**	**74.58**	**75.00**	**62.67**	**62.86**	**65.18**
GPT-4o	61.16	59.09	**77.00**	55.51	63.60	85.58	59.34	60.67	61.52	43.85	42.66	50.00
Open-source Models												
Qwen2.5_1.5b_instruct	52.89	54.55	43.80	55.70	60.48	56.62	40.91	45.90	49.36	37.10	40.28	41.47
Qwen2.5_7b_instruct	52.48	51.65	50.41	43.20	72.61	71.14	43.64	49.00	51.37	40.67	44.44	46.03
Qwen2.5_72b_instruct	64.46	65.70	64.88	38.79	71.32	74.63	**57.45**	**59.74**	**62.30**	**45.44**	**47.22**	**50.60**
Qwen3_4b_instruct	46.28	46.69	46.28	36.76	60.29	58.64	42.38	44.55	50.10	34.33	35.71	38.49
Qwen3_30b_instruct	60.33	59.92	61.16	53.31	79.04	80.51	50.10	54.26	56.63	40.08	43.25	44.64
Gemma3_4b_instruct	51.24	51.24	49.59	24.26	42.46	44.85	36.91	39.89	44.26	31.75	35.12	34.92
Phi-4	66.12	66.12	66.53	32.35	67.28	64.15	49.82	51.73	52.28	42.26	46.63	50.00
Llama3.3_70b_instruct	**71.90**	**73.55**	**68.60**	**59.01**	**85.29**	**84.93**	50.55	55.01	58.83	44.05	46.23	50.00
Saul_7b_instruct	47.52	47.93	43.80	30.51	47.43	43.93	32.00	31.33	31.15	25.00	26.59	24.60

Table 10: Performance on **HARD** difficulty questions. **BG** provides the case background, **BG+CC** further includes the current conversation, and **BG+CC+PC** incorporates the preceding conversations. We evaluate Understanding Intention, Action Prediction, and Response Generation using accuracy (%). **Bold-underlined** indicates the strongest among all models; **bold** indicates the strongest among open-source models.