JURISBENCH-OMNI: A JUDICIAL-PRACTICE-ORIENTED BENCHMARK FOR LEGAL OMNI-MODAL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have achieved significant progress in *multi-modal* tasks; however, they still face challenges in the *omni-modal* understanding of real-world signals. Court hearings constitute one of the most consequential stages of judicial practice and are inherently omni-modal: judges, litigants, and witnesses jointly convey legal meaning through oral arguments, facial expressions, bodily actions, procedural contexts, spatial environments, causal relations, and other symbolic cues (e.g., silence). Much of this meaning is compressed into written judgments. Current legal AI evaluation remains predominantly text-centric or treats audio and video as separated and auxiliary signals appended to general-purpose multi-modal templates and, as a result, fails to assess whether models can comprehend judicial hearings in a human-like manner by integrating speech, actions, and intentions into a coherent, continuous task representation. We introduce JURISBENCH-OMNI, a benchmark for legal omni-modal evaluation constructed directly from real hearing videos and their corresponding judgment documents. The benchmark decomposes judicial competence into four omni-modal dimensions: *role attribution*, *intention understanding*, *evidence recognition*, and *hearing-to-judgment grounding*.

1 INTRODUCTION

Large language models have transformed how text-based tasks are evaluated and deployed, and the same agenda is now being extended to perception across multiple modalities. Real-world communication does not arrive in a single modality. This challenge is most consequential in professional settings whose core working units are themselves omni-modal events. The court hearing is a paradigmatic such event. It is the procedural context in which formal claims are made and answered, in which factual and legal disputes are surfaced and crystallized, and in which the evidentiary record is established for the adjudicator to act on. The legal meaning of what occurs at the bench is carried jointly by what is said, who says it, how it is said, the visual conduct of the participants, and the procedural stage in which the exchange takes place. Much of this content is later transformed into the written judgment, and a faithful reading of the hearing record therefore depends on the joint interpretation of speech, audio, video, and document context.

Despite the central role of the hearing, current operational use of hearing recordings is narrow. The dominant pipeline reduces a hearing to an automatic speech-to-text transcript and treats this transcript as the primary artifact for downstream review. Information carried by speaker identity, prosodic emphasis, and visual conduct is largely discarded once the audio channel is collapsed into text. Practitioners report that audio-only transcripts tend to be undifferentiated records of the entire proceeding, without explicit marking of speaker roles, procedural stages, or the points at which factual and legal disputes are formed. The value such transcripts provide is largely confined to retrospective verification, namely confirming what was said and when, and the efficiency gain for the substantive task of drafting a judgment is therefore limited.

Closing this gap requires treating the hearing as an omni-modal object rather than as a transcript proxy, and treating omni-modal competence as a property that should be measured directly. The AI research community currently lacks both targeted study of judicial omni-modal understanding and a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

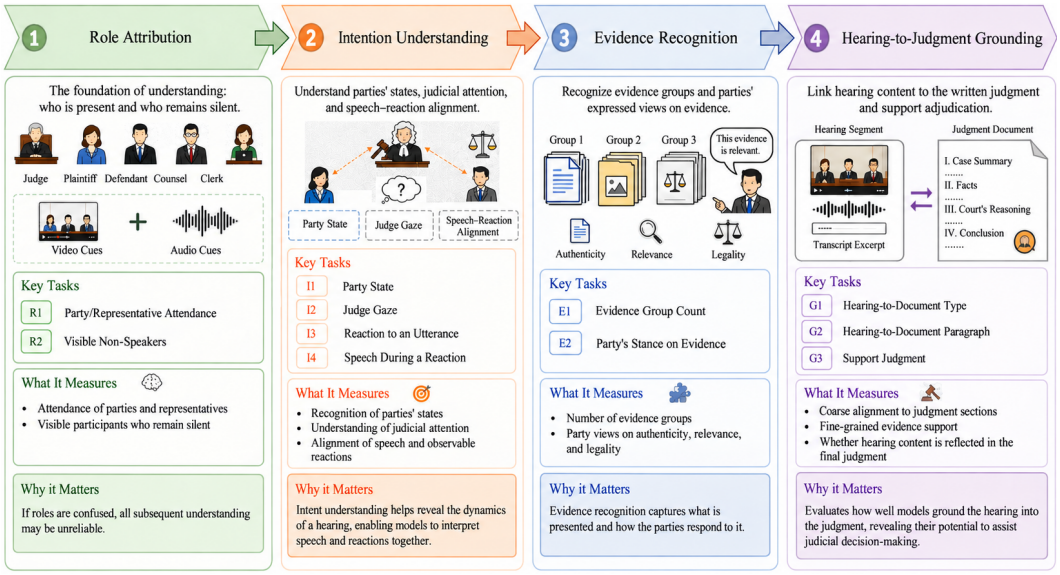

Figure 1: Overview of JURISBENCH-OMNI. The benchmark decomposes judicial omni-modal competence into four complementary capability dimensions, instantiated by 11 tasks across two temporal granularities. Judgment documents serve as reference materials when constructing answers for the hearing-to-judgment grounding tasks.

benchmark that can measure whether a model can interpret hearing content in a manner aligned with adjudicative practice. Existing legal AI benchmarks remain predominantly text-centered (Chalkidis et al., 2022; Guha et al., 2023; Fei et al., 2024), while recent work that introduces visual or video signals into legal tasks treats additional modalities as auxiliary inputs to an end prediction task rather than as the object of evaluation (Kang et al., 2026). General-purpose multimodal benchmarks, in turn, are designed around transferable perceptual and reasoning skills and do not reflect the procedural structure under which judicial meaning is formed.

We introduce JURISBENCH-OMNI, a benchmark for legal omni-modal evaluation built from real court hearings paired with the corresponding judgment documents. The benchmark uses Chinese first-instance civil hearings as its source, primarily because the Chinese judicial system makes a large body of hearing video publicly available through official channels. The capability formulation, however, is not specific to a particular jurisdiction: *role attribution*, *intention understanding*, *evidence recognition*, and *hearing-to-judgment grounding* are core elements of any adversarial adjudicative process. JURISBENCH-OMNI consists of 150 hearing videos spanning 10 civil case categories, paired with their corresponding judgment documents, and a total of 1,650 questions distributed across the four capability dimensions.

The contribution of this work is threefold. First, it formulates legal omni-modal evaluation around the structure of judicial practice rather than around generic multimodal QA, with explicit attention to the role of audio and video information that current audio-only pipelines discard. Second, it provides a benchmark constructed from authentic hearing recordings and judgment documents, organized along four capability dimensions and two temporal granularities, with expert-validated labels. Third, it provides an evaluation protocol that supports modality ablation, capability-wise breakdown, and diagnostic analysis of omni-modal models in adjudication-oriented settings.

2 RELATED WORK

The legal domain still lacks an omni-modal benchmark for evaluating whether a model can read a court hearing, even though a hearing is itself an omni-modal event in which legal meaning is carried jointly by speech, audio, video, the procedural role of the speaker, and the procedural stage in which the exchange occurs. General-purpose multimodal, omni-modal, and vertical-domain benchmarks each address parts of the picture, but none combines them in the legal setting.

Omni-modal benchmarks restore modality itself as an evaluation target. One line focuses on synchronous recognition and reasoning across visual, acoustic, and textual inputs (Li et al., 2024b; Chen et al., 2024a); another examines audio-visual integration in real-world settings, with recent work pushing toward fine-grained temporal annotation, longer videos, bilingual settings, and grounding tasks (Gong et al., 2024; Hong et al., 2025; Geng et al., 2025; Zhou et al., 2025; Zhang et al., 2025). These benchmarks make the modality channel measurable, but evaluate it as a generic capability rather than within any specific professional setting.

General-purpose multimodal benchmarks treat modality as input format rather than as a carrier of meaning. They are typically grouped by the competence they target: perception and alignment (Antol et al., 2015; Goyal et al., 2017); knowledge-mediated understanding (Marino et al., 2019; Lu et al., 2022; Yue et al., 2024); structure-sensitive reasoning over temporal order, distributed evidence, or comparative relations (Lu et al., 2023; Li et al., 2024a; Ben-Ami et al., 2025; Kil et al., 2024); and capability-aggregate evaluation across multiple dimensions (Fu et al., 2023; Liu et al., 2024; Li et al., 2023; Yu et al., 2023). Across these groups, multimodal competence is operationalized as a set of generic, transferable skills, and the institutional setting in which those skills would be deployed is not part of the evaluation.

Vertical-domain benchmarks supply the missing institutional setting, but in domains whose core working units are bimodal. In medicine, single-image VQA datasets target radiology, multi-organ reasoning, and pathology (Lau et al., 2018; Liu et al., 2021; He et al., 2020; Zhang et al., 2023); in finance, FinQA and TAT-QA evaluate numerical reasoning over reports and tables (Chen et al., 2021; Zhu et al., 2021). More recent work broadens coverage and aligns evaluation with professional structure across clinical and financial tasks (Chen et al., 2024b; Hu et al., 2024; Luo et al., 2025; Gan et al., 2025). These benchmarks share a common discipline—tasks correspond to units of professional work and errors are interpreted in domain terms—but they remain bimodal because the core working units in medicine and finance are themselves bimodal artifacts. Law differs: its core working unit, the court hearing, is itself an omni-modal event, and a vertical benchmark for law cannot inherit this bimodal template.

Legal AI benchmarks remain text-centric. In English-language settings, LexGLUE, CaseHOLD, and LegalBench evaluate legal language understanding, holding identification, and expert-designed legal reasoning (Chalkidis et al., 2022; Zheng et al., 2021; Guha et al., 2023); in Chinese civil-law settings, CAIL2018 and LawBench cover judgment prediction and broader legal knowledge (Xiao et al., 2018; Fei et al., 2024). Their inputs and outputs center on statutes, case descriptions, and documents. Recent work by Kang et al. takes a first step toward multimodality by introducing visual and video signals into judgment prediction (Kang et al., 2026), but multimodality there is a means rather than an object of evaluation: whether a model actually reads the hearing in a manner consistent with adjudicative practice is not itself measured.

JurisBench-Omni is designed to make these implicit, partial, or auxiliary elements into the evaluation target itself. It decomposes hearing understanding into role attribution, intention understanding, evidence recognition, and hearing-to-judgment grounding, four complementary capability dimensions grounded in adjudicative practice.

3 Benchmark Design

This section presents the design of JurisBench-Omni. We first decompose judicial omni-modal competence into four capability dimensions and the 11 concrete tasks that instantiate them (Section 3.1). We then describe how the benchmark is built from real court hearings, how questions are constructed, and how models are evaluated (Sections 3.2–3.4). Table 1 summarizes the input type and answer format of every task and serves as the structural reference for the rest of the section.

3.1 Capability Dimensions and Tasks

JurisBench-Omni evaluates whether a model can interpret a court hearing as a structured judicial event rather than as a generic audio-visual stream. Judicial meaning in a hearing is not produced by any single modality in isolation: the same utterance carries different legal weight depending on the speaker's role, the procedural stage, and the interactional intent it expresses, and is ultimately fixed by whether and how it is reflected in the written judgment. JurisBench-Omni therefore decomposes

Table 1: Question design of JURISBENCH-OMNI. Each task takes one of two video granularities as input. For the hearing-to-judgment grounding tasks, the corresponding written judgment is consulted when constructing answers. All questions use single-choice formats with task-specific candidate sets.

Dimension	ID	Task	Input clip	Answer format	# Q.
Role attribution	R1	party/representative attendance	utterance video	1-of-5 single choice	150
	R2	visible non-speakers	utterance video	1-of-5 single choice	150
Intention understanding	I1	party state	utterance video	1-of-4 single choice	150
	I2	judge gaze	utterance video	1-of-4 single choice	150
	I3	reaction to an utterance	utterance video	1-of-4 single choice	150
	I4	speech during a reaction	utterance video	1-of-4 single choice	150
Evidence recognition	E1	evidence group count	multi-turn interaction video	1-of-5 single choice	150
	E2	stance on evidence	multi-turn interaction video	1-of-5 single choice	150
Hearing-to-judgment grounding	G1	hearing-to-document type	utterance video	1-of-8 single choice	150
	G2	hearing-to-document paragraph	utterance video	1-of-5 single choice	150
	G3	support judgment	utterance video	1-of-3 single choice	150
Total		11 tasks × 150 cases	—	—	**1,650**

judicial omni-modal competence into four complementary capability dimensions, instantiated by 11 tasks in total (one task per case for each task type, yielding 11 questions per case).

- ***Role attribution*.** Identifying who is present and who is speaking is the entry point of judicial interpretation. A hearing includes both parties and their legal representatives, and participants visible in the courtroom may remain silent during a given exchange. Role attribution therefore requires distinguishing attendance from speaking activity.
  - **R1, party/representative attendance:** identify the attendance of the plaintiff and defendant parties and their legal representatives in the target clip (United States District Court for the Northern District of Florida, n.d.).
  - **R2, visible non-speakers:** identify which participants are visible in the frame but do not speak during the target clip. This distinction is studied in audio-visual active speaker detection, including visible faces labeled as non-speaking (Roth et al., 2020).
- ***Intention understanding*.** Beyond what is literally said, a hearing carries the participants' intentions through a party's conduct, the addressee structure of the bench's inquiry, and the reactions accompanying an exchange. Interpreting these interactions requires attending to observable behavior and aligning it with the speech occurring at the same moment.
  - **I1, party state:** identify the state or attitude of the party at a given moment, primarily from non-verbal conduct visible only in audio and video (Hoffman & Weiner, 2013).
  - **I2, judge gaze:** identify the direction of the judge's gaze while speaking, recovering who is being addressed and asked to respond (Auer, 2018).
  - **I3, reaction to an utterance:** identify a specified party's reaction when a given utterance is spoken, aligning the verbal cue with the contemporaneous response. Related work models listener facial responses conditioned on timestamped words (Ng et al., 2023).
  - **I4, speech during a reaction:** identify what is being said when a specified party exhibits a given reaction, aligning the observed response with the contemporaneous speech. This tests temporal grounding, motivated by work on the temporal coordination of speech, gaze, and listener responses (Bavelas et al., 2002).
- ***Evidence recognition*.** Understanding the evidentiary content of a hearing requires tracking the groups of evidence presented and recognizing the parties' expressed positions on their authenticity, relevance, and legality. These tasks use multi-turn clips to provide the context in which evidence is introduced and examined.
  - **E1, evidence group count:** identify the number of groups of evidence appearing in the target segment. This targets the tracking of evidence across its presentation and examination, central stages of Chinese civil hearings (Zhang, 2023).
  - **E2, stance on evidence:** identify a specified party's expressed position on the authenticity, relevance, and legality of a specified group of evidence in the target segment. Parties' ex-

pressed positions on these attributes are documented in an analysis of Chinese civil courtroom interaction (Xu & Ge, 2023).

- ***Hearing-to-judgment grounding.*** A hearing is consequential because much of its content is later carried into the written judgment. Grounding traces this transformation and is closest to model-assisted judgment drafting; failures here—fabricating support or missing genuine support—break the chain of accountability between proceeding and outcome.
  - **G1, hearing-to-document type:** identify the type of judgment-document section a hearing span corresponds to (coarse alignment) (Strong, 2015).
  - **G2, hearing-to-document paragraph:** identify the specific paragraph of the judgment document that a hearing span supports (fine-grained alignment) (Varsava, 2021).
  - **G3, support judgment:** judge whether the content of a hearing span is carried into the final judgment at all, targeting the negative case (Oldfather, 2008).

The four dimensions—role attribution, intention understanding, evidence recognition, and hearing-to-judgment grounding—provide complementary views of judicial hearing understanding. This capability-wise organization enables fine-grained diagnosis that aggregate scores do not provide. Each instance is drawn from a *case bundle* (hearing video and judgment document of the same case), is associated with an explicit modality configuration, and specifies its temporal granularity (utterance or multi-turn interaction), so that the input scope of every task is fully determined.

3.2 Data Sources, Filtering, and Segmentation

Data sources and matching. Hearing videos are obtained from the China Court Hearing Public Network, the official platform through which Chinese courts publish live-streamed and archived courtroom recordings (Supreme People's Court of the PRC, n.d.a). Judgment documents are obtained from the China Judgments Online portal, the central repository of publicly accessible Chinese judgments (Supreme People's Court of the PRC, n.d.b). Cases are matched by case number, which appears in both the hearing record and the judgment header.



Figure 2: Case selection pipeline. Review criteria and final dataset statistics are in Appendix A.

Quality filtering. Two failure modes account for the majority of exclusions at this stage. The first is insufficient visual detail for conduct- and gaze-related judgments. Source recordings at 720p or higher are preferred; lower-resolution recordings are retained when manual review finds the relevant visual cues interpretable. The second is severe audio loss caused by courtroom microphone configurations, which leaves portions of party speech largely inaudible. ASR-based screening followed by manual review excludes recordings with severe or pervasive audio loss, while isolated degradation is tolerated if the overall transcript remains interpretable. Audio quality accounts for the larger share of attrition. Review criteria and measured input statistics are reported in Appendix A.

Two-granularity segmentation. For each retained case, we use the audio-derived transcript as the temporal anchor and segment the hearing video at two granularities, summarized in Table 2. Each granularity has its own boundary-detection rule, its own characteristic length, and the set of tasks for which it serves as the input clip; together they let the same hearing supply local cues and multi-turn interactions without re-collecting data.

Table 2: The two temporal granularities used by JURISBENCH-OMNI. Durations and turn counts summarize the associated task inputs. Evidence recognition uses nominal five-minute windows, with actual durations varying according to clip boundaries; the remaining tasks use utterance-level clips. Measured distributions and boundary rules are described in Appendix A.

Granularity	Duration	Mean turns	Used by
Utterance	44.19 s	1	R1, R2; I1, I2, I3, I4; G1, G2, G3
Multi-turn	4.84 min	15.00	E1, E2

3.3 QUESTION DESIGN AND EVALUATION METRICS

Question design. For each of the 11 tasks defined in Section 3.1, the input granularity that the task substantively requires is used as the input clip (Table 1): role attribution, intention understanding, and hearing-to-judgment grounding tasks take utterance-level video (R1, R2, I1, I2, I3, I4, G1, G2, G3); evidence recognition tasks take approximately five-minute multi-turn segments (E1, E2). For the hearing-to-judgment grounding tasks, annotators consult the corresponding written judgment when constructing and verifying answers. The full judgment is not supplied as an additional model input; G2 provides candidate judgment paragraphs in the answer options. All questions use single-choice formats with task-specific candidate sets. Task definitions and example options are provided in Appendix B; answer construction and verification procedures are described in Appendix A.5.

Annotation. The annotators are nine law students and doctoral candidates, all of whom have completed core coursework in civil procedure and evidence law. Candidate answers and distractors were prepared by a two-member Construction Group. For R1, R2, I3, I4, E1, E2, G1, G2, and G3, each question was independently reviewed by two members of the Verification Group. I1 and I2 were generated and reviewed within the Construction Group. Disagreements were resolved by the Construction Group member responsible for the question. Items flagged as irresolvable by multiple verifiers were removed from the benchmark. Annotator demographics, per-task guidelines, and verification and adjudication procedures are described in Appendices A.5–A.7.

Evaluation metrics. Because all questions use single-choice formats, the primary per-task metric is plain accuracy. We report two complementary aggregate metrics. Mean accuracy summarizes performance across all eleven tasks and, together with per-task results, characterizes capabilities that may support specific forms of judicial assistance. However, reliable understanding of a hearing requires competence across role attribution, intention understanding, evidence recognition, and hearing-to-judgment grounding. Strong performance in one dimension cannot fully compensate for weakness in another. We therefore additionally report a *composite score* defined as the product of the four dimension mean accuracies, which penalizes such weaknesses more strongly than an arithmetic average:

$$\mathrm{CScore} \;=\; \mathrm{Acc}_R \cdot \mathrm{Acc}_I \cdot \mathrm{Acc}_E \cdot \mathrm{Acc}_G, \qquad \mathrm{Acc}_d \in [0,1].$$

Each Acc_d is the arithmetic mean of the task accuracies in dimension d. The composite score uses unrounded accuracies; it is a summary of cross-dimensional competence, not an observed success rate for complete judicial proceedings.

3.4 EVALUATED MODELS

Evaluation requires foundation models that can simultaneously process hearing video clips, embedded audio, and accompanying text within a single inference call. We selected models accessible at the time of evaluation for compatibility with these input requirements and the single-choice response format. The model-selection rationale is described in Appendix C. Two systems meeting these requirements are used as the main evaluation targets: `gemini-3.1-pro-preview` and `qwen-3.5-omni-flash`. For both models, inference is conducted in a zero-shot setting using a unified prompt template per task type, with no in-context examples drawn from the benchmark. The prompt structure, a representative example, and the inference and scoring protocol are described in Appendix C.

4 Experiments and Analysis

4.1 Experimental Setup

We evaluate the two omni-modal foundation models that natively process hearing video clips with embedded audio together with text in a single inference call: `gemini-3.1-pro-preview` and `qwen-3.5-omni-flash`. All inference is zero-shot under the V (video with embedded audio) configuration with a unified prompt template per task type. Both models are evaluated on all eleven subtasks. Throughout this section, parser failures (i.e., outputs from which no option letter can be extracted) are charged as *wrong*, so all reported accuracies are lower bounds on what the models could achieve under more lenient parsing. We report per-(model, task) accuracy and summarize performance across the four dimensions by the *CScore* defined in Section 3.3. Dimension-level summaries use the arithmetic mean of constituent task accuracies.

4.2 Results Analysis

Per-task accuracy. Table 3 reports per-task accuracy with parser failures counted as wrong, alongside the Product (%) score ($100 \times$ CScore) and mean accuracy across all eleven tasks. `qwen-3.5-omni-flash` achieves a mean accuracy of 44.67% across the eleven tasks, compared with 39.88% for `gemini-3.1-pro-preview`. It leads on seven tasks, trails on three (I1, I2, G1), and ties on G2 at 60.00%. The largest gap is on I4: `qwen-3.5-omni-flash` reaches 43.33%, compared with 20.00% for `gemini-3.1-pro-preview`, a difference of 23.33 percentage points. Its advantages on E1, E2, and G3 are 9.33, 10.00, and 11.33 points, respectively. The strongest result for either model is 64.00% on I2, while both remain near the 20% random-choice reference on the role tasks.

Table 3: Main results on JurisBench-Omni (V configuration; accuracy in %, parser failures counted as wrong). **Bold** marks column-wise best results, including ties. Cells below 50% are red; those at or above 50% are green, with darker green indicating higher values. **Prod. (%)** is $100 \times$ CScore, the product of the four dimension mean accuracies (Section 3.3); **Mean (%)** averages all eleven tasks. Aggregates use unrounded accuracies.

	Role		Intention				Evidence		Ground			Prod. (%)	Mean (%)
Model	R1	R2	I1	I2	I3	I4	E1	E2	G1	G2	G3		
Random choice	20.0	20.0	25.0	25.0	25.0	25.0	20.0	20.0	12.5	20.0	33.3	0.22	22.35
`gemini-3.1-pro-preview`	20.7	16.7	**48.0**	**64.0**	51.3	20.0	36.0	39.3	**48.7**	**60.0**	34.0	1.53	39.88
`qwen-3.5-omni-flash`	**21.3**	**24.0**	43.3	59.3	**55.3**	**43.3**	**45.3**	**49.3**	44.7	**60.0**	**45.3**	**2.70**	**44.67**

Capability shape across the four dimensions. Figure 3 plots per-task accuracy over the eleven subtasks, grouped by the four JurisBench-Omni dimensions (Role: R1, R2; Intention: I1, I2, I3, I4; Evidence: E1, E2; Ground: G1, G2, G3). Both polygons contract toward the centre on the Role arc and extend toward I2 and G2. The clearest difference appears at I4, where `gemini-3.1-pro-preview` drops to 20.00% despite reaching 51.33% on I3. For `qwen-3.5-omni-flash`, the corresponding I3 and I4 accuracies are 55.33% and 43.33%. This contrast shows that the two directions of speech–reaction alignment have different observed difficulty: identifying a reaction from a verbal cue and identifying concurrent speech from a reaction are distinct evaluation targets. The gap is also visible on the Evidence arc, where `qwen-3.5-omni-flash` leads on both evidence-group counting and recognition of a party's expressed position on evidence.

Granularity is not difficulty. A useful cross-cut to read alongside the dimension view is the temporal granularity at which each subtask is administered (Figure 4). The nine utterance-level tasks (R1, R2, I1, I2, I3, I4, G1, G2, G3) contribute 1,350 questions; the two approximately five-minute multi-turn tasks (E1, E2) contribute 300. `gemini-3.1-pro-preview` scores 40.37% on utterance-level questions and 37.67% on multi-turn questions, whereas `qwen-3.5-omni-flash`

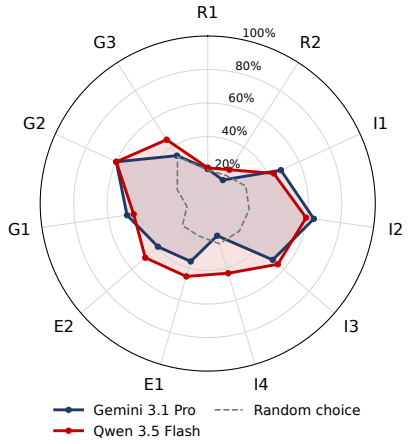

Figure 3: Per-task accuracy across the eleven subtasks. The dashed polygon marks task-specific random-choice references.

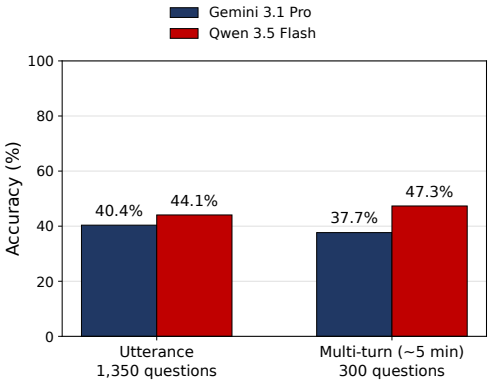

Figure 4: Accuracy pooled within each temporal granularity. Groups contain different tasks; the comparison does not isolate clip length.

scores 44.07% and 47.33%, respectively. The models therefore exhibit opposite orderings across the two granularities. Because the groups differ in task content and answer-option counts, these results do not establish an effect of duration alone. In particular, the multi-turn group measures evidence recognition, so its stronger separation between models also reflects the capability being tested.

Composite performance: the weakest-dimension penalty. `qwen-3.5-omni-flash` leads in all four dimensions: Role (22.67% versus 18.67%), Intention (50.33% versus 45.83%), Evidence (47.33% versus 37.67%), and Ground (50.00% versus 47.56%). Multiplying these four dimension mean accuracies yields Product scores of 2.70% and 1.53%, respectively, compared with 0.22% for random choice. CScore summarizes dimension-level mean accuracies multiplicatively; it is distinct from the observed fraction of cases on which all tasks are answered correctly. Together, these results suggest partial competence and potential for narrowly scoped assistance under human supervision, but do not establish reliability for autonomous judicial workflows. The following case-level analysis tests joint correctness.

Cumulative performance across dimensions. Figure 5 tracks joint correctness over the same 150 cases in the fixed order Role → Intention → Evidence → Ground. A case passes a dimension only when every task in that dimension is correct, and remains in the cumulative count only if it also passes all preceding dimensions. After Role, 6/150 cases (4.0%) remain for `gemini-3.1-pro-preview` and 9/150 (6.0%) for `qwen-3.5-omni-flash`. Requiring all four Intention tasks to be correct reduces both counts to zero; the cumulative rates consequently remain zero after Evidence and Ground. Thus, neither model answers all eleven tasks correctly for any case. The curve is non-increasing by construction: it measures joint correctness under a strict all-task criterion, rather than demonstrating causal error propagation between independently answered questions. First-failure attribution is reported in Appendix D.

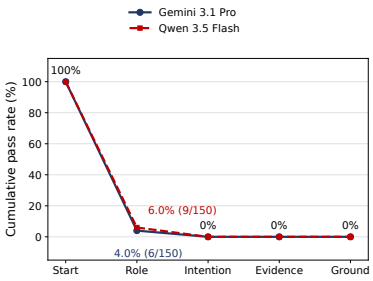

Figure 5: Cumulative case pass rates in the fixed R→I→E→G order. All tasks in the current and preceding dimensions must be correct; the denominator remains 150 cases.

Input-modality ablations and supplementary evaluation. Figure 6 compares four input ablations for `qwen-3.5-omni-flash`: options only, question and options, silent video with question and options, and audio with question and options. Each uses 150 questions per task, retaining unsuccessful requests in the denominator. Mean accuracy is 21.09%, 24.73%, 31.94%, and 35.21%, respectively,

versus 44.67% with full audio-video input. Silent video preserves I2 accuracy (59.33%), while audio-only G1 (56.67%) exceeds the full-input result (44.67%), showing task-dependent modality effects. We also report full-input `qwen-3.5-omni-plus` on 50 questions per task (550 total), with a mean of 47.82%. This smaller evaluation is reported separately and does not establish a matched-sample advantage over Flash.

Qwen 3.5 Omni: Flash ablations + Plus supplement Accuracy (%)

	Role		Intention				Evidence		Ground			
	R1	R2	I1	I2	I3	I4	E1	E2	G1	G2	G3	Mean
Options only	17.3	18.0	31.3	8.7	28.7	23.3	29.3	26.0	4.7	35.3	9.3	21.1
Question + options	13.3	16.7	13.3	2.7	46.0	28.7	24.7	17.3	11.3	42.0	56.0	24.7
Video only (silent)	19.3	17.3	41.3	59.3	57.3	24.7	19.3	18.0	18.0	37.3	39.3	31.9
Audio only	24.0	22.0	26.7	40.0	41.3	26.0	28.0	18.7	56.7	60.0	44.0	35.2
Audio + video (main)	**21.3**	**24.0**	**43.3**	**59.3**	**55.3**	**43.3**	**45.3**	**49.3**	**44.7**	**60.0**	**45.3**	**44.7**
Plus full (50/task)	26.0	34.0	48.0	66.0	58.0	44.0	40.0	48.0	54.0	58.0	50.0	47.8

Figure 6: Flash input ablations (150 questions/task) and full-input Plus (50/task). All conditions except options only include the question and options. Colors show accuracy (%); the red outline marks Flash full input.

Interpreting task-specific chance levels. The tasks have different candidate-set sizes, so a common 50% threshold would obscure their reference levels. For `gemini-3.1-pro-preview`, R2 (16.67%) and I4 (20.00%) fall below their respective uniform random-choice references of 20% and 25%, while R1 (20.67%) and G3 (34.00%) are only about 0.67 percentage points above theirs. `qwen-3.5-omni-flash` is numerically above the reference on all eleven tasks, but its margins on R1 and R2 remain only 1.33 and 4.00 points. By contrast, both models reach 60.00% on G2, 40.00 points above its 20% reference. These comparisons describe the observed accuracy margins; establishing statistical significance or attributing a low score to answer-letter bias would require further analysis.

5 Conclusion and Future Work

JurisBench-Omni is motivated by the gap between how court hearings are currently used in practice, where audio-only transcription discards a substantial portion of the information that judges rely on, and the omni-modal nature of hearings as judicial events. By decomposing competence into role attribution, intention understanding, evidence recognition, and hearing-to-judgment grounding, JurisBench-Omni supports capability-wise analysis, modality ablation, and structured error attribution that aggregate end-task scores do not provide.

The current version of JurisBench-Omni has several limitations that suggest directions for future work. The benchmark is restricted to first-instance civil hearings in the Chinese judicial system, and extending coverage to criminal and administrative hearings, as well as to other civil-law and common-law jurisdictions, would broaden its applicability. The four capability dimensions are objective in form, which is suitable for stable cross-model comparison, but does not directly evaluate the quality of generated reasoning or draft judgment text; future versions can incorporate controlled open-ended subsets aligned with the same case bundles. The benchmark also currently treats annotation labels as fixed gold answers; integrating measures of judicial uncertainty, including disagreement among reviewers, would allow evaluation to better reflect the calibration requirements of high-stakes legal use. Finally, while JurisBench-Omni is intended for capability evaluation rather than autonomous deployment, the long-term goal of this line of work is to support tools that help judges use the full hearing record more effectively, and we view careful benchmark-driven evaluation as a prerequisite for any such use.

AI USE STATEMENT

We used large language models to assist with benchmark construction, specifically the generation of candidate answers and distractor options. The associated human review and verification procedures are described in Appendix A: candidate answers and distractors were verified against the hearing materials and corresponding judgment documents where applicable. Disagreements and uncertain items were handled through the annotation and adjudication procedures described in Appendix A.5. The foundation models evaluated in our experiments are identified in Section 3.4, with inference details in Appendix C.

We also used generative AI tools to assist with language editing and LaTeX formatting. The authors take responsibility for reviewing AI-assisted work and for the final content of the paper, including its claims, results, and any materials prepared with AI assistance.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Peter Auer. Gaze, addressee selection and turn-taking in three-party interaction. In Geert Brône and Bert Oben (eds.), *Eye-tracking in Interaction: Studies on the Role of Eye Gaze in Dialogue*, volume 10 of *Advances in Interaction Studies*, pp. 197–232. John Benjamins, 2018. doi: 10.1075/ais.10.09aue.

Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, 52(3):566–580, 2002. doi: 10.1111/j.1460-2466.2002.tb02562.x.

Dan Ben-Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. HERBench: A benchmark for multi-evidence integration in video question answering. *arXiv preprint arXiv:2512.14870*, 2025.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, Zifeng Wang, Yandong Li, Pranav Shyam, Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, et al. OmnixR: Evaluating omni-modality language models on reasoning across modalities. *arXiv preprint arXiv:2410.12219*, 2024a.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J. Seibel, Junjun He, and Yu Qiao. GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. In *Advances in Neural Information Processing Systems*, 2024b.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.300.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Ziliang Gan, Dong Zhang, Haohan Li, Yang Wu, Xueyuan Lin, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, Rongjunchen Zhang, and Yong Dai. MME-Finance: A multimodal finance benchmark for expert-level understanding and reasoning. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2025.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, and Shiliang Zhang. FunASR: A fundamental end-to-end speech recognition toolkit. In *Proceedings of Interspeech 2023*, pp. 1593–1597, 2023. doi: 10.21437/Interspeech.2023-1428.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. LongVALE: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18959–18969, 2025.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. AV-Odyssey bench: Can your multimodal LLMs really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shangsheng Gao, Spencer Williams, Sunny Gandhi, Tomer Zur, Varun J. Iyer, and Zehua Li. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2023.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Janet Lee Hoffman and Andrew Weiner. The juror as audience: The impact of non-verbal communication at trial, 2013. Oregon State Bar Litigation Journal.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. *arXiv preprint arXiv:2502.04326*, 2025.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Zhaolu Kang, Junhao Gong, Qingxi Chen, Hao Zhang, Jiaxin Liu, Rong Fu, Zhiyuan Feng, Yuan Wang, Simon Fong, and Kaiyue Zhou. Multimodal multi-agent empowered legal judgment prediction. *arXiv preprint arXiv:2601.12815*, 2026.

Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Chao. MLLM-COMPBENCH: A comparative reasoning benchmark for multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2024.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. OmniBench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024b.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, Sirui Han, Ming Zhang, and Yike Guo. FinMME: Benchmark dataset for financial multi-modal reasoning evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10083–10093, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Ng_Can_Language_Models_Learn_to_Listen_ICCV_2023_paper.html.

Chad M. Oldfather. Writing, cognition, and the nature of the judicial function. *Georgetown Law Journal*, 96:1283, 2008.

Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew C. Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, and Caroline Pantofaru. AVA-ActiveSpeaker: An audio-visual dataset for active speaker detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. URL https://research.google/pubs/ava-activespeaker-an-audio-visual-dataset-for-active-speaker-detection/.

S. I. Strong. Writing reasoned decisions and opinions: A guide for novice, experienced, and foreign judges. *Journal of Dispute Resolution*, pp. 93, 2015.

Supreme People's Court of the PRC. China court hearing public network. https://tingshen.court.gov.cn/, n.d.a.

Supreme People's Court of the PRC. China judgments online. https://wenshu.court.gov.cn/, n.d.b.

United States District Court for the Northern District of Florida. Role of the judge and other courtroom participants, n.d. URL https://www.flnd.uscourts.gov/role-judge-and-other-courtroom-participants. Accessed: 2026-05-06.

Nina Varsava. Professional irresponsibility and judicial opinions. *Houston Law Review*, 59:103, 2021.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. CAIL2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*, 2018.

Jun Xu and Lei Ge. Displaying epistemic stance through same-turn self-repair in chinese civil courtroom interaction. *Frontiers in Psychology*, 14:1285759, 2023. doi: 10.3389/fpsyg.2023.1285759.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Chenyang Zhang. Trial. In *Win in Chinese Courts: Practice Guide to Civil Litigation in China*, pp. 63–76. Springer, Singapore, 2023. doi: 10.1007/978-981-99-3342-6_5.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Yiman Zhang, Ziheng Luo, Qiangyu Yan, Wei He, Borui Jiang, Xinghao Chen, and Kai Han. OmniEval: A benchmark for evaluating omni-modal models with visual, auditory, and textual inputs. *arXiv preprint arXiv:2506.20960*, 2025.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pp. 159–168, 2021.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-Omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 3277–3287, 2021.

A Data Construction Details

A.1 Source Channels and Matching Procedure

Hearing videos were collected from the China Court Hearing Public Network (https://tingshen.court.gov.cn/), the official platform through which Chinese courts publish live-streamed and archived courtroom recordings. Judgment documents were collected from the China Judgments Online platform (https://wenshu.court.gov.cn/), the central repository for publicly accessible written judgments issued by Chinese courts. Case matching was performed by aligning the case number, which is unique to each judicial proceeding and appears in both the hearing record and the judgment document header. Initial candidate collection and matching were conducted through manual web browsing of both platforms rather than automated crawling, as neither platform provides a bulk download interface. From approximately 1,600 candidate hearing videos surveyed, 602 could be matched to a retrievable judgment document on China Judgments Online, yielding a matching rate of approximately 38%.

A.2 Quality Filtering

Matched cases were subjected to two sequential quality filters before entering the benchmark.

Visual resolution filter. The preferred lower bound on source-video resolution was 720p, chosen to support the conduct- and gaze-related judgments required by the Intention understanding tasks. Lower-resolution recordings were retained as exceptions when manual review found the relevant visual details interpretable. Across the 150 cases, the utterance-task inputs have 121 cases at 720p or higher, 28 at 576p, and one at 450p. Evidence-task clips also include lower-resolution inputs. Table 5 reports the measured resolution distribution across all eleven tasks; the source-video preference is not a minimum resolution requirement for every task input.

Audio quality filter. Audio quality was assessed in two stages. All matched recordings were first transcribed using FunASR (Gao et al., 2023), an automatic speech recognition system. The resulting transcripts were then submitted to a large language model for an initial quality assessment, which flagged transcripts exhibiting high rates of unintelligible segments, incomplete utterances, or anomalous character distributions attributable to microphone-induced signal loss. Cases flagged at this stage were subsequently reviewed manually, and recordings with severe or pervasive audio loss were excluded. Partial audio degradation affecting only isolated turns was tolerated provided the overall transcript remained interpretable.

Filter outcome. Of the 602 matched cases, 197 passed both quality filters and were retained as candidates for the final benchmark, corresponding to an overall retention rate of approximately 33% relative to the matched pool, or roughly 1 in 8 of the original candidate videos. The primary cause of attrition at the quality stage was audio rather than visual: while low-resolution recordings were straightforwardly identifiable, inadequate microphone configurations—particularly in older or lower-tier courtrooms—accounted for the majority of exclusions.

A.3 Final Benchmark Statistics

From the 197 quality-filtered cases, 150 were selected for the final benchmark. Cases were organized into 10 civil dispute categories. Category selection prioritized case types for which sufficient hearing videos and matching judgment documents were jointly available on the two platforms. Case types involving family, marital, or inheritance matters were excluded because hearings in these categories are frequently withheld from public streaming due to privacy considerations, and the available volume after matching and quality filtering was insufficient to support a balanced category. The 10 selected categories and their case counts are listed in Table 4.

Category counts vary because the initial candidate pool for each category was fixed during the browsing phase, and categories with lower video availability or higher attrition at the quality filter stage yielded fewer usable cases.

Table 4: Distribution of cases across the 10 civil dispute categories in JURISBENCH-OMNI.

Category	Cases
Sale contract disputes	18
Insurance disputes	18
Property service contract disputes	16
Property rights protection disputes	16
Lease contract disputes	16
Product liability disputes	16
Financial loan contract disputes	15
Labor disputes	13
Motor vehicle accident disputes	13
Private lending disputes	9
Total	**150**

A.4 CLIP STATISTICS AND GRANULARITY DEFINITIONS

Clip statistics are computed over the task-linked video inputs. The dataset contains 150 distinct case identifiers shared by all eleven tasks, with one question, one answer file, and one selected video input per case and task. Counts are therefore question-weighted: a source clip reused by two tasks contributes two video inputs, and the total of 1,650 does not imply 1,650 distinct excerpts. Durations and frame dimensions are read from the MP4 files; turn counts are the numbers of dialogue entries in their paired ASR metadata.

Table 5: Measured resolution distribution across the 1,650 task-linked video inputs. Reused clips are counted once for each question that uses them.

Resolution	Inputs	Proportion
1920×1080	941	57.03%
1440×900	32	1.94%
1440×810	32	1.94%
1280×720	302	18.30%
1024×576	268	16.24%
854×480	5	0.30%
800×450	11	0.67%
704×576	33	2.00%
640×360	5	0.30%
426×240	20	1.21%
384×240	1	0.06%
Total	**1,650**	**100%**

The two task-level temporal granularities are utterance clips and multi-turn interaction clips. R1, R2, I1, I2, I3, I4, G1, G2, and G3 use the utterance granularity, while E1 and E2 use the multi-turn granularity. This assignment specifies the task context rather than imposing a hard duration limit or a minimum number of ASR dialogue entries.

Utterance level. An utterance boundary is defined by a speaker change in the FunASR diarization output. Because a full hearing contains a large number of speaker turns, utterance-level clips are not drawn exhaustively. Instead, for each case, the transcript is divided into non-overlapping windows of 10 consecutive turns, and a large language model selects 3 turns from each window based on transcript content quality, prioritizing turns that are complete, intelligible, and substantively informative. Clip boundaries are aligned to the ASR timestamps. The 1,350 inputs for R1, R2, I1, I2, I3, I4, G1, G2, and G3 each contain one ASR dialogue entry. Their measured mean duration is 44.19 seconds, with a median of 23.00 seconds. An extended single turn is still classified as an utterance, not as a procedural-stage segment.

Multi-turn interaction level. Both E1 and E2 are classified as multi-turn interaction tasks. Their clips are selected around evidence presentation and examination, using nominal five-minute windows anchored to the ASR transcript. Boundaries follow complete speaker turns, so actual durations can exceed five minutes or be shorter. Across all 300 evidence-task inputs, the measured mean duration is 290.59 seconds (4.84 minutes), with a median of 307.01 seconds. Their ASR metadata contain a mean of 15.00 dialogue entries.

Task assignment and clip statistics. The benchmark therefore groups 1,350 questions under the utterance granularity and 300 questions under the multi-turn granularity. Table 6 reports measured statistics for all inputs within each task, including shorter clips.

Table 6: Measured clip duration and ASR turn count by task. Each task has 150 questions. Duration summaries are in seconds and include every task input.

Task	Mean duration	Median	Mean turns
R1	58.80	43.02	1.00
R2	52.19	49.02	1.00
I1	13.86	10.52	1.00
I2	14.27	9.51	1.00
I3	39.80	24.51	1.00
I4	79.58	57.00	1.00
E1	257.74	303.02	14.21
E2	323.44	312.04	15.78
G1	40.16	17.50	1.00
G2	57.67	16.50	1.00
G3	41.33	25.00	1.00

Data consistency checks. The task folders contain the expected number of candidate options for every question, and every gold option letter is present in its corresponding option set. All 1,650 MP4 inputs contain both video and audio streams.

A.5 Annotator Recruitment and Roles

A total of 9 annotators were recruited from law students and doctoral candidates at Peking University School of Law, all of whom had completed core coursework in civil procedure and evidence law. All annotators participated voluntarily and will be acknowledged in the paper. Demographic information is provided in Table 7.

Annotators were divided into two functionally independent groups. The Construction Group consisted of 2 annotators, who were responsible for generating candidate answers and distractors for each task based on the case bundle materials. The Verification Group consisted of the remaining 7 annotators, who independently reviewed the outputs of the Construction Group for correctness and consistency.

Table 7: Demographic information of annotators.

Annotator ID	Age	Gender	Location
Annotator 1	29	Male	Beijing
Annotator 2	24	Male	Beijing
Annotator 3	32	Male	Beijing
Annotator 4	18	Female	Beijing
Annotator 5	23	Male	Beijing
Annotator 6	23	Male	Beijing
Annotator 7	28	Female	Beijing
Annotator 8	26	Female	Beijing
Annotator 9	22	Male	Beijing

A.6 PER-TASK WRITTEN GUIDELINES

Each of the 11 tasks is a single-choice question. For tasks R1, R2, I3, I4, E1, E2, G1, G2, and G3, the Construction Group generates a candidate answer and distractor options; the Verification Group reviews whether the answer is correct by watching the corresponding video clip and consulting the ASR transcript where necessary. For tasks I1 and I2, answers are generated and reviewed by the Construction Group directly, as these tasks require sustained attention to specific visual details that are more efficiently handled within a single workflow.

R1 Watch the target utterance clip. Identify the people attending on the plaintiff and defendant sides, distinguishing the parties themselves from their authorized representatives. Check which of the five options best describes the attendance configuration, including joint attendance where applicable. Attendance does not require speaking. Revise the answer or options if they conflict with the audio-visual evidence and record the change.

R2 Watch the target utterance clip. Identify the procedural participants who are visible but do not speak during the clip, excluding spectators. Check whether the selected option describes that combination of roles. This remains a single-choice task even though an option can name several roles. Revise if incorrect and record the change.

I1 This task is completed by the Construction Group. The annotator locates the target person in the frame using the position and clothing description provided in the prompt, observes the specified time window, and judges whether the person produces verbal expression, physical conduct, or both. Physical conduct includes head movements and body movements but excludes minimal postural adjustments such as blinking or breathing.

I2 This task is completed by the Construction Group. The annotator locates the judge in the frame, observes the judge's head orientation and eye direction during the target utterance, and selects the option that best describes the predominant gaze direction. If the judge is not visible on screen during the utterance, select the corresponding option.

I3 Locate the phrase specified in the question and the designated participant on the plaintiff or defendant side. Check the participant's observable reaction at the time indicated by the question and select the matching option. Use the temporal relation specified in the prompt, including "after" where applicable, rather than inferring a reaction from the phrase alone. Revise the answer or distractors as necessary and record any changes.

I4 Locate the specified reaction or movement of the designated participant. Listen to the speech occurring at that moment and check which of the four options matches its content. The selected speech must coincide with the reaction, rather than merely occur elsewhere in the same clip. Revise if incorrect and record the change.

E1 Watch the evidence-related clip and count the groups of evidence presented within it. Distinguish a group from individual documents or items within that group, and do not count repeated references to the same group as new groups. Select one of the five count options. Check the answer against the presentation in the clip and record any revisions.

E2 Locate the evidence specified in the question and the examining party's expressed position on its authenticity, relevance, and legality. Select the option that matches that position, including which attributes are disputed or accepted. The target is the party's stated opinion, not the annotator's independent assessment of the evidence's legal validity. Revise if incorrect and record the change.

G1 Watch the target clip and consult the judgment document. Check whether the labeled document section correctly corresponds to the primary content of the clip. Revise if incorrect and record the change.

G2 Watch the target clip and read the full judgment document. Check whether the labeled paragraph is the one most directly supported by the hearing exchange, and whether the distractor paragraphs are

drawn from the same document. Revise the answer if a more directly corresponding paragraph exists and record the change.

G3 Watch the target clip and read the full judgment document. Check whether the labeled degree of reflection accurately describes how much of the clip's content appears in the document. Revise if incorrect and record the change.

A.7 Agreement and Resolution

Verification was conducted at the case level. Each verifier worked through the case folders sequentially, using the hearing video (MP4) as the primary reference and the ASR transcript as a supplementary aid. For the tasks assigned to the Verification Group, the verifier assessed whether the candidate answer and distractors were correct and consistent with the video content, and made revisions where necessary. Each such question was independently reviewed by two verifiers; in cases where the two verifiers reached different conclusions, the discrepancy was referred to the Construction Group member responsible for that question, whose judgment was treated as final. I1 and I2 followed the Construction Group review procedure described above.

Upon completing each case, the verifier updated the case status to `verified`. All modifications were logged in a separate text file recording the case identifier, the original content, and the revised content for every item that was changed. Verifiers were instructed to provide a response for every question without exception; where a judgment was uncertain, verifiers were asked to record their best-effort interpretation of the video and flag the item in the log file for subsequent review. Items flagged by multiple verifiers as irresolvable were removed from the benchmark.

B TASK DEFINITIONS AND EXAMPLES

For each of the four capability dimensions and 11 tasks, we provide the task definition followed by a representative example translated from the Chinese prompt and its answer file. Examples use case 001, except E2, which uses case 002; these are dataset folder identifiers rather than positions in a consecutive list. Each example includes the question stem, the candidate options, and the gold answer.



Figure 7: Real Trial Video Images

B.1 ROLE ATTRIBUTION

R1 Given an utterance-level hearing video clip, identify the attendance configuration on the plaintiff and defendant sides, distinguishing the parties themselves from their authorized representatives. Select the best matching configuration from five options.

Example. *What is the attendance configuration on the plaintiff and defendant sides in this hearing? Select the best matching option.*

A. The plaintiff attends in person; the defendant attends in person.

B. The plaintiff attends in person; the defendant is represented by an authorized representative.

C. The plaintiff is represented by an authorized representative; the defendant attends in person.

D. Both sides are represented by authorized representatives.

E. Both the plaintiff and the defendant attend together with their respective representatives.

Gold answer: E

R2 Given an utterance-level hearing video clip, identify the combination of procedural participants who are visible but do not speak, excluding spectators. Each of the five options specifies a combination of roles, and only one option is selected.

Example. *Apart from the speaker, which roles are visible in this hearing clip but do not speak? Exclude spectators and select the best matching option.*

A. Judge and court clerk
B. Court clerk, plaintiff's representative, and third party
C. Plaintiff's representative and third party
D. Defendant's representative and court clerk
E. Plaintiff's representative and defendant's representative

Gold answer: B

B.2 Intention understanding

I1 Given a hearing video clip, identify whether the designated party produces verbal expression, physical conduct, or both during the specified time window. Physical conduct includes head movements and body movements.

Example. *Observe the first 10 seconds of the video and focus on the leftmost person in the frame. Which of the following best describes the state of that person during this period?*

A. Speaking with noticeable physical conduct
B. Speaking without noticeable physical conduct
C. Not speaking but with noticeable physical conduct
D. Not speaking and without noticeable physical conduct

Gold answer: C

I2 Given a hearing video clip, identify the direction of the judge's gaze while speaking during the specified time window.

Example. *Observe the first 10 seconds of the video and focus on the judge's gaze direction while speaking. Which of the following best describes where the judge is looking?*

A. Looking toward the plaintiff side
B. Looking toward the defendant side
C. Looking at own materials or screen
D. Judge not visible on screen; gaze direction cannot be determined

Gold answer: C

I3 Given an utterance-level hearing clip and a specified verbal cue, identify the designated participant's observable reaction at the time indicated by the question. The four options describe alternative reactions; the task requires aligning the verbal cue with visual conduct.

Example. *After the phrase "We will now begin the court investigation" is spoken in the video, what does the man seated at the far right of the plaintiff's table do?*

A. Holds materials in one hand and reads them
B. Lowers his head and bends forward toward the table
C. Continues looking straight ahead
D. Touches his ear with one hand

Gold answer: A

I4 Given an utterance-level hearing clip and a specified reaction of a participant, identify the speech occurring at that moment. The four options specify alternative speech contents. This reverses the direction of alignment in I3: visual conduct supplies the cue, and speech supplies the target.

Example. *When the woman on the defendant's side scratches her head and adjusts her hair in the clip, what is being said at that moment?*

A. The defendant's representative is saying, "prove that the leased premises in this case are . . . "

B. The defendant's representative is saying, "is [a measure] by an administrative authority against a party in default . . . "

C. The defendant's representative is saying, "regarding the third claim . . . "

D. The defendant's representative is saying, "require the defendant to pay late fees . . . "

Gold answer: B

B.3 EVIDENCE RECOGNITION

E1 Given an evidence-related hearing clip, identify how many groups of evidence are presented in it. The five options are counts from zero to four. A group can contain several documents or individual items, so the number of groups need not equal the number of items.

Example. *How many groups of evidence are presented in this hearing clip? Select the best matching option.*

A. 0

B. 1

C. 2

D. 3

E. 4

Gold answer: E

E2 Given an evidence-related hearing clip and a specified piece or group of evidence, identify the examining party's expressed position on its authenticity, relevance, and legality. Select one of five descriptions of that position. The model is not asked to independently determine whether the evidence meets these legal criteria.

Example (case 002). *Which aspect of the surveillance-video evidence does the examining party dispute in this hearing clip? Select the best matching option.*

A. Authenticity is disputed.

B. Authenticity and relevance are disputed.

C. Legality is disputed.

D. No objection is raised to authenticity, legality, or relevance.

E. Authenticity and legality are disputed.

Gold answer: D

B.4 HEARING-TO-JUDGMENT GROUNDING

G1 Given a hearing video clip, identify the section of the judgment document to which it most directly corresponds.

Example. *Which section of the judgment document does the following hearing clip correspond to?*

A. Basic information of the parties
B. Plaintiff's claims
C. Defendant's claims
D. Disputed issues
E. Facts as found by the court
F. Reasoning for the judgment
G. Operative decision
H. Not corresponding to any of the above sections

Gold answer: C

G2 Given a hearing video clip, identify the specific paragraph of the judgment document that it most directly supports. The candidate options are paragraphs drawn from the judgment document in the case bundle; the remaining paragraphs serve as distractors.

Example. *Which paragraph of the judgment document does the following hearing clip correspond to?*

A. During the hearing, the defendant acknowledged the facts alleged by the plaintiff regarding the outstanding fees.
B. Plaintiff: Yunnan Tin Longma General Equipment Co., Ltd., Kunming Branch.
C. Third party Tin Longma Company and third party Tin Real Estate Company stated in their defense that they recognized the claims put forward by the plaintiff.
D. The defendant Liefeng Sports Company stated in its defense: the defendant is willing to vacate the premises; since the plaintiff cut off water and electricity to the premises after June 30, 2018, preventing normal use, the defendant should only be liable for rent, water fees, electricity fees, and cleaning fees up to June 30, 2018; the claim for late performance penalties has no factual or legal basis and should be dismissed. Third party Tin Longma Company and third party Tin Real Estate Company recognized the claims put forward by the plaintiff.
E. The hearing clip does not correspond to any paragraph in the judgment document.

Gold answer: D

G3 Given a hearing video clip, assess the degree to which its content is reflected in the final judgment document.

Example. *To what extent does the content of the following hearing clip appear in the final judgment document?*

A. The content is largely reflected in the judgment document.
B. Only part of the content is reflected in the judgment document.
C. The content is not reflected in the judgment document at all.

Gold answer: B

C Model Inference Details

C.1 Model Selection Rationale

We selected multimodal foundation models accessible at the time of evaluation for compatibility with the benchmark's inputs and response format. The input requirements imposed by JurisBench-Omni are: (1) hearing video input, (2) access to the embedded audio, (3) joint audio-video-text processing within a single inference call, and (4) support for the selected utterance and evidence-related clips, including clips that extend beyond the nominal five-minute window. A system must also return a response from which a single option letter can be extracted.

The selection therefore considers the combination of input modalities, supported clip length, and compatibility with the single-choice evaluation protocol. Image-and-text access alone does not cover the required setting, and visual video input alone does not establish access to speech or other audio cues. These requirements concern the interfaces used for this evaluation rather than a general ranking of model families.

The two models retained for the main evaluation are `gemini-3.1-pro-preview` and `qwen-3.5-omni-flash`. Both receive the same task question and candidate options with the corresponding hearing clip. Comparisons in the main results use these two models; other model variants are not pooled into their scores.

C.2 Question Prompts

All prompts were written in Chinese, consistent with the language of the hearing materials. Each question is provided separately with its own assistant instruction, question text, candidate options, and an output format instruction requiring the model to return a JSON object containing only the selected option letter with no additional explanation. In the main evaluation, both models receive the hearing video clip with embedded audio and the same textual question and options. For G1, G2, and G3, the judgment document is a reference for answer construction and verification, rather than an additional full-document model input. G2 includes candidate judgment paragraphs in its options.

A representative question from task **R1** is shown below, with an English translation provided for readability.

Table 8: A representative question from task R1, translated from Chinese.

Component	**Content**
Assistant instruction	You are a professional court hearing video analysis assistant. Observe the attendance configuration on the plaintiff and defendant sides in the supplied video.
Output instruction	Output a JSON object containing only the selected option letter (A/B/C/D/E). Do not output any other explanation.
Question	What is the attendance configuration on the plaintiff and defendant sides in this hearing? Select the best matching option.
Options	A: Plaintiff in person; defendant in person. B: Plaintiff in person; defendant's representative. C: Plaintiff's representative; defendant in person. D: Representatives of both sides. E: Each party together with their respective representatives.

I1, I2, I3, and I4 have four options; R1, R2, E1, E2, and G2 have five; G1 has eight; and G3 has three. A question remains single-choice even when an option describes multiple people or several objections to evidence.

C.3 Inference and Scoring Protocol

Both models were evaluated in a zero-shot setting with no in-context examples drawn from the benchmark. The input is the selected hearing clip with its audio stream; the stored clip dimensions and durations are summarized in Appendix A. Models were instructed to output a JSON object containing only the selected option letter, with no explanatory text. Each task contains one question per case, and accuracy uses all 150 questions as the denominator, including unsuccessful responses. A recovered response is scored only when a single option can be extracted unambiguously; remaining parsing and interface failures count as incorrect.

D CASE-LEVEL FIRST-FAILURE ANALYSIS

Where does the first failure occur? Figure 8 assigns each case to its first dimension with at least one incorrect answer in the fixed Role → Intention → Evidence → Ground order. For `gemini-3.1-pro-preview`, 144 cases (96.0%) first fail at Role and the remaining 6 (4.0%) at Intention. For `qwen-3.5-omni-flash`, the corresponding counts are 141 (94.0%) and 9 (6.0%). Neither model has a first failure at Evidence or Ground, because no case passes both preceding dimensions. These zeros do not indicate error-free evidence recognition or grounding. The attribution depends on the reporting order and complements the per-task results by showing where cases first cease to satisfy the cumulative criterion.

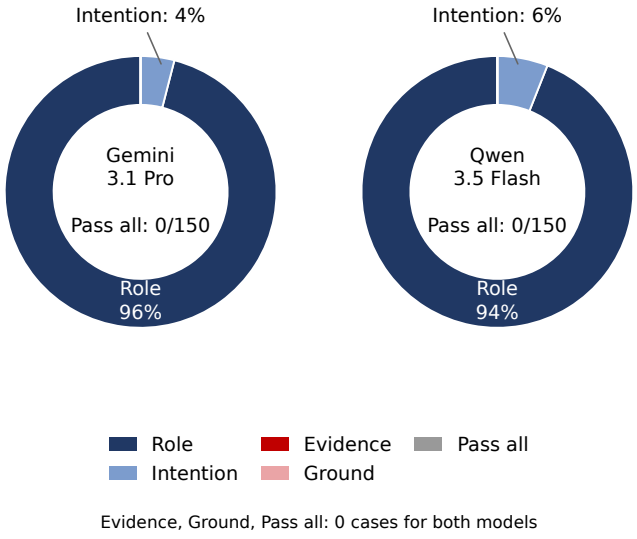

Figure 8: First-failure distribution over 150 cases per model in the fixed R→I→E→G order. A dimension fails if any constituent task is wrong. No case passes all four dimensions.

E Ethical Considerations and Responsible Use

E.1 Data Governance and Access

The hearing videos used in this benchmark were collected from the China Court Hearing Public Network, and the judgment documents were collected from China Judgments Online. Both platforms publish judicial materials that have been made legally accessible to the public in accordance with Chinese court transparency regulations. The collection, organization, and use of these materials comply with the public access rules of both platforms. The benchmark dataset will be released under a controlled access policy: applicants must specify their research purpose and will be granted access only upon review and approval.

E.2 Privacy Treatment

The hearing videos published on the China Court Hearing Public Network have been reviewed and approved for public release by the courts concerned. The judges, parties, and counsel who appear in these recordings participate in their official procedural capacities in a publicly streamed setting. No additional anonymization of individual participants has been applied to the benchmark materials, as the source recordings are themselves legally public. The controlled access policy described above is intended to ensure that the dataset is used solely for academic research purposes and to prevent misuse of the personal information that may be incidentally present in the recordings.

E.3 Jurisdictional Scope

This benchmark is constructed entirely from civil hearing recordings of courts in mainland China. The case types, procedural rules, linguistic conventions, and judgment document formats all reflect the specific characteristics of the Chinese civil litigation system. Evaluation results obtained on this benchmark should not be extrapolated directly to other jurisdictions or to other types of judicial proceedings.

E.4 Responsible Use

JurisBench-Omni is designed to evaluate the capacity of omni-modal foundation models to interpret judicial hearing content, and not to certify the suitability of any model for deployment in actual judicial processes. High performance on this benchmark does not constitute evidence that a model is capable of assisting or replacing human adjudication. Judicial decision-making involves legal accountability, procedural legitimacy, and professional ethics in ways that are not captured by benchmark performance. Any use of model outputs directly in judicial decision-making carries substantial legal and ethical risks. Researchers are advised to treat benchmark results as academic capability indicators rather than as evidence of deployment readiness. Furthermore, this benchmark is not intended to support the use of models for evaluating or profiling the individual conduct of specific parties, judges, or counsel appearing in the hearing recordings.