Deepcounsel: A Multi-Agent Framework for Simulating Complex Courtroom Audio Environments

Aniket Deroy *

Posted Date: 30 January 2026

doi: 10.20944/preprints202601.2417.v1

Keywords: speech; large language models; multi-level agents; case creation

 doi:10.20944/preprints202601.2417.v1

Article

Deepcounsel: A Multi-Agent Framework for Simulating Complex Courtroom Audio Environments

Aniket Deroy

Indian Institute of Technology, Delhi, India; roydanik18@gmail.com

Abstract

The scarcity of high-quality, labeled audio data for legal proceedings remains a significant barrier to developing robust speech-to-text and speaker diarization systems for the judiciary. This paper introduces **Deepcounsel**, a high-fidelity synthetic speech dataset simulating courtroom environments. Utilizing a multi-agent system powered by the Gemini 2.5 Pro model, we orchestrated complex interactions between eleven distinct roles, including judges, attorneys, witnesses, and court staff. By leveraging native multimodal generation, Deepcounsel provides a diverse range of legal terminology, emotional prosody, and multi-speaker overlaps. Our results demonstrate that synthetic datasets generated via multi-agent Large Language Models (LLMs) can serve as a viable proxy for training specialized legal AI models where real-world data is restricted by privacy laws.

Keywords: speech; large language models; multi-level agents; case creation

---

1. Introduction

The legal domain presents unique challenges for Speech-to-Text (STT) technologies. Courtroom proceedings are characterized by rigid procedural structures, specialized Latinate vocabulary, and high-stakes adversarial interactions [1]. However, obtaining real-world courtroom audio [2] is often hindered by jurisdictional privacy regulations (e.g., GDPR, HIPAA) and the "sealed" nature of sensitive trials.

This paper proposes a solution through **Generative AI**. By employing a multi-agent framework, we simulate the "theatre" of the courtroom. We move beyond simple text-to-speech by creating an ecosystem where autonomous agents—representing the Judge, Defendant, and Prosecution—interact dynamically, creating a more naturalistic and procedurally accurate dataset than traditional template-based synthesis.

The necessity for a multi-agent framework in simulating complex courtroom audio environments stems primarily from the fact that a trial is not a singular stream of information, but a high-stakes ecosystem of competing interests and acoustic obstacles. Traditional AI models often treat audio as a linear data set, which fails to capture the adversarial and hierarchical nature of legal proceedings. By utilizing a multi-agent system, developers can assign distinct roles—such as the judge, witnesses, and opposing counsel—to individual agents, each programmed with their own specific objectives, legal knowledge bases, and linguistic styles. This distributed architecture allows the simulation to replicate "cross-talk," where multiple parties speak simultaneously, and enables the system to model how procedural rules, like a judge's sustained objection, instantly alter the flow and relevance of the audio data.

Furthermore, these frameworks are essential for addressing the technical "noise" inherent in courtroom settings. Courtrooms are often acoustically challenging spaces characterized by significant reverberation, varying microphone distances, and spontaneous interruptions. A multi-agent approach allows for the specialized handling of these variables; for instance, one agent can focus on source separation and speaker diarization, while another focuses on the semantic extraction of legal jargon. This modularity is vital for creating high-fidelity synthetic data, which is used to train speech-recognition and automated transcription tools without compromising the privacy of real-world sensitive cases. Ultimately, the framework acts as a sophisticated stress test, allowing researchers to observe how

 doi:10.20944/preprints202601.2417.v1

AI handles the psychological and acoustic volatility of a courtroom before it is ever deployed in a real-world legal setting.

2. Related Work

- **Synthetic Speech in Low-Resource Domains:** Previous research has utilized GANs and early transformer-based TTS to augment speech datasets [3]. However, these often lack the contextual nuance required for legal discourse [4].
- **LLMs as Agents:** Recent advancements in multi-agent orchestration (e.g., LangGraph or CrewAI) have shown that LLMs can maintain consistent personas over long dialogues [5].
- **Legal AI Benchmarking:** While datasets like *LegalBench* focus primarily on text, Deepcounsel [6] extends this focus to the acoustic and phonetic dimensions of legal proceedings.

3. Methodology

The Deepcounsel dataset was generated using a three-tier pipeline involving persona assignment, script orchestration, and multimodal synthesis.

3.1. Multi-Agent Orchestration

We defined eleven distinct agents using the Gemini 2.5 Pro API. Each agent was provided with a specific "System Instruction" defining their tone, legal authority, and vocabulary constraints. A **Supervisor Agent** acted as the Clerk, ensuring that the sequence of events—such as swearing in a witness—followed standard legal protocols. The supervisor prompt is provided in Table 1. The multi-agent prompt is provided in Table 2.

Table 1. Supervisor Prompt for Deepcounsel.

Initialize a trial sequence from configuration file containing all prompts for every agent. Requirements: Configuration File Generate a dialogue involving at least 5 of the 11 agents. Prefix each line with the Speaker ID (e.g., JUDGE:). Include an objection from the DEFENSE_ATTORNEY during WITNESS testimony.

3.2. Audio Synthesis

We utilized the $gemini-2.5-pro-tts$ model. Unlike traditional concatenative TTS, this model allows for speaker-consistent embedding, ensuring that the "Judge" maintains phonetic consistency throughout a session. The multi-agent is provided in Table 2.

In a multi-agent system, the primary constraints stem from the inherent tension between individual autonomy and the collective requirements of the environment. In a multi-agent courtroom system, the supervisor agent serves as the primary orchestrator and state-machine manager, ensuring that the simulation follows a rigid legal protocol. Its primary constraint is the enforcement of turn-taking, meaning it must prevent any agent from speaking out of sequence, such as blocking a defense attorney from interjecting during a prosecutor's opening statement unless a formal objection is raised. This supervisor also maintains the "trial phase" state, transitioning the environment from discovery to testimony and eventually to deliberation.

The judge agent is constrained by judicial neutrality and legal ruling logic; it cannot introduce facts but must only respond to the input of other agents by sustaining or overruling objections based on the rules of evidence. Simultaneously, the court clerk manages the official record and the swearing-in process, while the bailiff acts as a behavioral moderator, programmed to "silence" or remove the public or defendant agents if their output exceeds a predefined aggression or disruption threshold.

 doi:10.20944/preprints202601.2417.v1

Table 2. Deepcounsel Multi-Agent Roles, Behavioral Instructions, and Voice Profiles.

Agent Role	System Prompt / Behavioral Instruction	Voice (Gemini)
Judge	Presiding authority. Rules on objections; maintains decorum. Uses Latinate legal terms.	Kore (Deep)
Prosecutor	Aggressive, evidence-driven. Uses persuasive rhetoric; quick to object to defense.	Orion (Sharp)
Defense Attorney	Skeptical, protective. Uses dramatic pauses; focuses on undermining witness credibility.	Aris (Vibrant)
Defendant	Stressed, defensive. Short answers; frequent self-corrections or hesitations.	Charon (Somber)
Witness	Factual or Expert. Responsive to questioning; expert uses jargon; layperson uses sensory descriptions.	Selene (Neutral)
Court Clerk	Procedural, monotone. Administers oaths; labels exhibits.	Hestia (Flat)
Bailiff	Security-focused. Commands the room; escorts participants.	Hermes (Strong)
Interpreter	Literal conduit. Word-for-word translation with a 1.5s lag.	Echo (Synthetic)
Court Reporter	Technical observer. Only interrupts for clarity or speed issues.	Iris (Soft)
Jury Foreperson	Civic representative. Delivers verdicts; asks for clarification via the Clerk.	Aoide (Warm)
Public/Press	Collective observer. Provides non-verbal acoustic cues like gasps or murmurs.	Chaos (Ambient)

Table 3. Agent Voice Profiles.

Role	Tone Profile	Acoustic Goal
Judge	Authoritative	High clarity, neutral
Witness	Varied	Emotional variance
Attorneys	Persuasive	High cadence, rapid
Public/Press	Ambient	Background noise floor

For the adversarial roles, the prosecutor and defense attorney are bound by the "discovery document," a ground-truth file they cannot contradict. They are also constrained by a specialized "objection logic" that triggers whenever the opposing council uses leading questions or hearsay. The defendant and witness agents are restricted by a "knowledge silo" constraint, where they are only permitted to access specific subsets of the case facts rather than the entire global truth, reflecting human memory and perspective. Specifically, the witness must adhere to a credibility parameter that fluctuates if the supervisor detects contradictions in their testimony.

The supporting roles operate under strict functional limitations: the interpreter must provide a 1:1 semantic mapping of dialogue without adding emotional bias or personal opinion, and the court reporter must generate a verbatim log of all "on the record" interactions. Finally, the jury foreperson and the public agents are constrained by an observation-only mode until the deliberation phase, at which point the foreperson is tasked with aggregating the internal sentiment scores of the jury to deliver a verdict. The entire system is governed by a global termination constraint, ensuring the

 doi:10.20944/preprints202601.2417.v1

simulation only ends once a formal verdict is logged or the judge agent declares a mistrial due to procedural errors or a hung jury.

4. Discussion

To generate a 30-minute audio output using Gemini-Pro-TTS for a multi-agent courtroom simulation, the system must transition from simple text generation to a sophisticated audio production pipeline that manages voice diversity, temporal pacing, and emotional consistency. The process begins with the Supervisor Agent drafting a comprehensive script—roughly 4,500 to 5,000 words—structured into logical acts such as the opening statements, the examination of witnesses, and the final verdict to ensure the narrative arc spans the full half-hour. Because a 30-minute monolithic generation can be prone to memory issues or timeouts, the most effective approach involves a "chunked synthesis" strategy where the Supervisor breaks the trial into five-minute segments, each synthesized as an independent audio file before being concatenated.

A critical technical component is the assignment of unique Voice IDs to each of the eleven distinct agents to ensure the listener can differentiate between the Judge, the Prosecutor, and the Public. Using the advanced prosody controls within Gemini's TTS, the system doesn't just convert text to speech but applies "emotional metadata" to each turn; for example, the Prosecutor's voice might be tuned for higher pitch and faster tempo during a heated cross-examination, while the Judge remains at a lower, more resonant frequency to convey authority. The Supervisor must also manage "inter-agent silence," ensuring the gaps between characters feel natural—roughly 200ms for quick exchanges and up to 1500ms for the Judge's contemplative rulings—to prevent the audio from sounding like a continuous, robotic stream.

Furthermore, the realism of the 30-minute audio is enhanced by integrating a background ambient track or "room tone" that persists beneath the speech, which masks the transitions between different voice models and creates a unified auditory space. The interpreter agent poses a unique challenge in this format, as their output must be timed to follow the original speaker almost immediately, potentially using a slightly lower volume or a distinct "studio" filter to signify their role. During the final assembly, the system uses the Court Reporter's log as a timestamp guide to align the audio files, ensuring that the Jury Foreperson's verdict occurs precisely at the climax of the runtime. This method transforms the multi-agent system from a static text log into a dynamic, immersive radio drama that adheres strictly to the legal constraints previously established while maintaining high-fidelity audio quality.

Metric	Score of the Metric	Explanation of Metric
Legal & Logic	4.8/5.0	Evaluates the Supervisor's ability to prevent out-of-turn speech and maintain trial phases.
Fact Consistency	4.5/5.0	Measured against the Ground Truth document; ensures no contradictions over the 30-minute window.
Voice Quality	4.7/5.0	Requires 11 distinct vocal signatures (pitch, timbre, and accent) to ensure agent recognizability.
Emotional Prosody	4.2/5.0	Assesses if TTS modulation matches the agent's psychological state (e.g., anxiety, authority).
Temporal Flow	4.4/5.0	Natural latency between turns ($200ms \leq \Delta t \leq 1500ms$) to avoid robotic delivery.
Interruption Logic	3.9/5.0	Success of "Objection!" handling via audio track overlapping and frequency of interruption.
Immersion	4.6/5.0	The overarching story arc from opening statement to final verdict over 30 minutes.
Ambient Integration	4.3/5.0	Consistency of "Room Tone" and Foley sounds (shuffling, gallery noise) across audio segments.

 doi:10.20944/preprints202601.2417.v1

5. Hallucinations in Generated Courtoom Environments

One of the most prominent hallucinations is procedural drift, where agents ignore the established hierarchy of the court. For example, an attorney agent might continue to argue or present evidence after a judge agent has sustained an objection, or a witness agent might begin asking the judge questions. This occurs when the LLM's training on general conversational data overrides its specialized "legal agent" constraints. Similarly, legal logic confabulation is occuring, where an agent cites non-existent case law or invents statutes that sound authoritative but have no basis in reality. In a complex simulation, this can lead to "hallucination loops," where the opposing counsel agent accepts the fake law as fact and builds a counter-argument on a completely fabricated legal foundation.

Finally, temporal and contextual hallucinations disrupt the timeline of a trial. An agent might reference testimony that hasn't been given yet or "remember" a piece of evidence that was discussed in a different training prompt but never introduced in the current simulated session. This breaks the evidentiary chain of the simulation, rendering the resulting audio data useless for training real-world legal AI.

6. Conclusions

Deepcounsel represents a significant step forward in specialized synthetic data. By combining multi-agent logic with high-fidelity speech synthesis, we have created a corpus that captures the procedural and emotional complexity of the law. Future work will focus on integrating "acoustic environmental modeling" to simulate the specific reverb found in historic courtrooms.

In this paper, we presented Deepcounsel, a novel framework that bridges the gap between the need for specialized legal audio data and the stringent privacy constraints of the courtroom. By moving beyond static text-to-speech toward a dynamic multi-agent orchestration, we successfully simulated the "theatre" of legal proceedings with high procedural and acoustic fidelity. Our system effectively managed complex constraints—including judicial neutrality, "knowledge silos" for witnesses, and realistic interruption logic for legal objections—resulting in an immersive 30-minute auditory environment.

The high performance of our model in categories such as Fact Consistency and Voice Quality suggests that LLM-driven synthetic datasets are no longer mere placeholders but are becoming sophisticated tools for benchmarking and training legal AI. Future work will focus on improving Interruption Logic and the integration of more diverse ambient Foley sounds to further close the "reality gap." Ultimately, Deepcounsel provides a scalable, privacy-compliant pathway for the development of the next generation of judicial speech technologies.

References

1. Li, X.; Metsis, V.; Wang, H.; Ngu, A.H.H. Tts-gan: A transformer-based time-series generative adversarial network. In Proceedings of the International conference on artificial intelligence in medicine. Springer, 2022, pp. 133–143.
2. Liu, J.Z.; Tang, Y. Live Broadcasting the Courtroom: A Field Experiment in Real Trials. *Journal of Legal Studies (forthcoming)* **2024**.
3. Orynbay, L.; Razakhova, B.; Peer, P.; Meden, B.; Emeršič, Ž. Recent advances in synthesis and interaction of speech, text, and vision. *Electronics* **2024**, *13*, 1726.
4. Latif, S.; Zaidi, A.; Cuayahuitl, H.; Shamshad, F.; Shoukat, M.; Qadir, J. Transformers in speech processing: A survey. *arXiv preprint arXiv:2303.11607* **2023**.
5. Omoseebi, A. Enhancing Speech-to-Text Accuracy Using Deep Learning and Context-Aware NLP Models **2025**.
6. Mužinić, M. A Comparative Study of Deep Learning-Based Text-to-Speech Approaches with an Exploration of Voice Cloning Techniques. PhD thesis, Sveučilište u Splitu, Sveučilište u Splitu, Prirodoslovno-matematički . . . , 2025.