Multimodal AI for Law Students: A Practical Guide

Seth J. Chandler, with help from AI

August 16, 2025

1 What This Is About

Legal work has always involved more than text. You analyze contracts alongside diagrams, review deposition videos with transcripts, and examine photographs with witness statements. Until recently, AI could only handle the text parts. Now it can process images, audio, video, and text together, which changes how legal work gets done.

This guide explains what these systems actually do, which ones exist today, and how you'll likely use them in practice.

2 How Multimodal AI Actually Works

Traditional AI systems could only read text. Modern systems convert different types of information—text, images, audio, video—into the same mathematical format so they can be analyzed together. Think of it as translation: the system converts a contract clause, a signature image, and an audio recording of "we signed the agreement" into the same numerical language, then looks for connections between them.

The process works in four steps. First, specialized components convert each type of input into numerical representations. Text gets broken into pieces and converted to numbers that capture meaning. Images get divided into patches and converted similarly. Audio gets analyzed for patterns in pitch and rhythm. Second, the system combines these representations so different types of information can interact. Third, the system is trained on millions of examples to learn how text, images, and audio relate to each other. Fourth, it can answer questions, find connections, or generate new content based on what it learned.

The key limitation is that these systems work best when inputs are clear. Blurry images, poor audio quality, or badly scanned documents will produce unreliable results. Always verify what the system tells you by checking the original sources.

3 What's Available in 2025

OpenAI's GPT systems handle text and images and include voice interaction through ChatGPT. You can speak to the system while showing it documents or images, and it responds by voice. This works well for studying

and practice simulations.

Google's Gemini processes text, images, audio, video, and documents natively. Gemini Live provides voice interaction in multiple languages and can reference what's on your camera or screen. Google also offers Veo for generating short videos, useful for creating demonstrative evidence.

Anthropic's Claude excels at handling very long documents with images. If you need to analyze hundreds of pages of depositions plus exhibits in one session, Claude's large context window makes this practical.

Other models You should understand that most of the popular AI models available today include some multimodal capabilities, though the combinations vary. Text plus images is now standard– Text plus audio is increasingly common, with many models offering speech recognition and synthesis. Video understanding is emerging but still limited to shorter clips. Truly integrated multimodal systems that seamlessly process text, images, audio, and video simultaneously remain relatively rare and cutting-edge.

When evaluating models on platforms like Hugging Face, you can quickly identify multimodal capabilities through several indicators. Model names often reveal their scope: anything with "vision," "CLIP," "BLIP," or "LLaVA" typically handles images and text, while "Whisper" indicates audio processing. The model description page is most reliable—look for phrases like "processes text and images" or task descriptions such as "image captioning" and "visual question answering." Check the example code snippets: multimodal models require both text and image inputs, while text-only models just need text. In the repository files, multimodal models typically include multiple configuration files (like separate vision and text configs) and both tokenizers and image processors. If you see upload buttons for images or audio in the model's demo interface, rather than just a text box, that's a clear sign of multimodal capability. When in doubt, the model card's task tags and supported formats will definitively tell you what inputs the model

accepts.

Specialized legal tools are emerging rapidly. Speech-to-text services now handle multiple speakers and background noise well. Document processing tools can extract text from handwritten notes and preserve the location of each word on the page. Discovery platforms integrate different file types so you can search across emails, images, and audio recordings simultaneously.

4 Discovery and Document Review

Modern discovery produces mixed file types: native documents, scanned pages, screenshots, photos, and video clips. Traditional text-only systems miss most of this content.

Current multimodal systems run optical character recognition on scanned documents, extract text from handwritten notes where possible, and treat images and videos as searchable content. This means you can find a key email that references a diagram, locate the diagram itself, and identify related audio recordings in one search.

Discovery platforms like Reveal and Brainspace now process these different file types together. You train the system by manually reviewing a small sample of mixed documents, marking which ones are relevant. The AI learns from your examples and automatically prioritizes similar content across all file types. Instead of reviewing text documents separately from images and audio, you see related materials together regardless of format.

The practical benefit is that you spend less time hunting through differ-

ent file types and more time understanding how evidence connects across formats.

5 Deposition Analysis

Transcripts miss important information. They don't capture hesitation, tone changes, or visual cues like looking away from the camera or gesturing toward documents.

Several vendors now offer systems that claim to analyze deposition videos alongside transcripts. Companies like Lexitas, DepoIQ, and others market "behavioral analysis" features that purport to assess witness credibility, detect stress points, and identify moments worth human review. These systems typically provide general assessments rather than specific behavioral measurements.

The practical reality is more limited than the marketing suggests. Current systems can identify basic patterns like long pauses in speech or flag sections where audio quality changes, but claims about detecting deception or measuring credibility through body language remain largely unproven. The science underlying behavioral analysis for credibility assessment is itself disputed among experts.

What these tools do provide is automated flagging of potentially interesting moments: sections with unusual speech patterns, long pauses, or changes in vocal tone. Some can sync these observations with transcript timestamps. However, you should treat any "credibility analysis" or "deception detection" claims with significant skepticism.

The practical value lies in triage rather than analysis. Instead of watching hours of video, you get a list of moments that might warrant closer human examination. For right now, treat these AIs as providing leads requiring verification, not as conclusions about witness truthfulness.

6 Visual Evidence Analysis

When cases involve whether a photo matches a blueprint, whether signage appears where it should, or whether equipment was positioned correctly, multimodal systems can align textual claims with specific regions in images.

Instead of manually comparing a contract clause to a diagram for an hour, you ask the system to find the relevant parts of both, highlight discrepancies, and show you exactly where they occur. The system might return: "Clause 3.2 specifies valve position as 'fully open' but Exhibit C shows valve at 45-degree angle, region coordinates 450,230 to 600,380."

You still verify the analysis, but you start with specific locations and claims rather than hunting through documents.

7 Creating Demonstrative Evidence

Video generation tools are improving rapidly but remain limited for professional legal work. Current systems like Google's Veo 3 can generate 8-second clips at 720p resolution, while Runway's Gen-4 focuses on consistency across

scenes but still requires multiple attempts to get usable results. Industry professionals report generating "10, 20, 30, 40 different videos just to get one that moves correctly and doesn't have any weirdness".

These tools currently work better for rough drafts and concept visualization than finished demonstratives. The output quality varies significantly, and you often cannot predict what you'll get from a given prompt. Video models struggle with "prompt adherence"—accurately following specific instructions, making them unreliable for precise legal illustrations.

The technology is advancing quickly, but for now, traditional animation and video production remain more reliable for courtroom use. If you do experiment with AI-generated content, treat it as preliminary material requiring significant human review and likely professional refinement.

Document any AI assistance thoroughly: the specific tools used, input prompts, number of attempts made, and any manual modifications. This documentation will be essential for admissibility.

8 Contract Review and Transactional Work

Multimodal systems can process marked-up contracts with handwritten annotations, translating ink notes and arrows into structured text suggestions. They can check visual layout against content requirements: signature block placement, presence of corporate seals, proper exhibit formatting.

For deals involving multiple languages, these systems can compare translated versions side by side, checking that diagrams, tables, and text align consistently across language versions.

In due diligence, you can treat spreadsheets, architectural plans, emails, and photographs as one integrated dataset. The system can theoretically connect a photo of a building facade with permit documents and financial records, flagging inconsistencies in address, square footage, or structural features that might affect valuation.

9 Voice Mode for Learning and Practice

Voice interaction represents perhaps the most significant development for legal education because so much legal practice remains fundamentally oral. OpenAI's Advanced Voice Mode in ChatGPT and Google's Gemini Live provide real-time spoken conversation with minimal delay, creating opportunities for practice that were previously impossible without human partners. Moreover, oral practice develops skills you cannot get from text alone: thinking quickly under pressure, maintaining composure when challenged, and articulating complex ideas clearly in real-time. These are fundamental lawyering skills that most law students get limited opportunities to develop before graduation.

Technology specifics: OpenAI's system processes speech in real-time and responds with synthesized voice that can interrupt, pause, and adjust tone. Gemini Live works similarly and can reference materials on your screen or camera while speaking. Both systems maintain conversational flow rather than the stop-start pattern of traditional speech-to-text tools.

Oral argument practice: You can argue cases before the AI, which can play different roles—a skeptical judge, opposing counsel, or a panel with varying judicial philosophies. The system can interrupt with questions, press you on weak points, and force you to think on your feet. Unlike written practice, this develops the rhythm and timing essential for courtroom

advocacy.

Client counseling simulations: Practice breaking bad news, explaining complex legal concepts, or handling emotional clients. The AI can roleplay different client personalities—confused, angry, or unrealistic—helping you develop communication skills across various scenarios. This is particularly valuable for students who haven't had extensive client contact.

Negotiation training: Conduct mock negotiations where the AI represents opposing parties with different strategies, risk tolerances, and personalities. You can practice reading verbal cues, adjusting tactics mid-conversation, and handling pressure tactics in real-time.

Deposition preparation: The AI can play hostile witnesses, evasive deponents, or challenging opposing counsel. Practice your questioning technique, learn to follow up on incomplete answers, and develop the persistence needed for effective discovery.

Judicial clerkship preparation: Practice explaining complex cases orally, defending recommendations, and fielding questions about legal reasoning—skills essential for clerk positions that often aren't taught in traditional coursework.

Oral exams: One issue with AI is that it challenges the integrity of written assessments. The school or business may be learning how well you know how to use AI more than how well you know the substance of the law. Oral assessments are one way around this problem. At present, it is hard to have the AI answer questions for you or whisper good answers in your ear in real time. The voice modes of the AIs, however, can help you practice for oral exams. The AI can simulate different examiner styles and press you for clarity and precision.

International and comparative law: Many of these AIs understand languages other than English. You can thus practice legal discussions in these other languages.

10 Programming in Plain English

Current AI systems let you give instructions in ordinary English and receive working code in return. You can say "create a script that searches through deposition transcripts and finds every time someone says 'I don't remember' with page and line numbers" and get functional Python code.

The necessary skill is writing precise instructions and checking results, not learning programming syntax. If the code doesn't work, you take a screenshot of the error and explain what went wrong. The system debugs and fixes the problem.

This capability makes data analysis accessible to lawyers who aren't programmers. You can process large datasets, create visualizations, and test hypotheses by describing what you want in clear English.

11 Practical Limitations and Risks

Visual errors are harder to spot than text mistakes. If a system misidentifies part of a diagram or mislabels a chart axis, the error might not be obvious. Always verify visual analysis against original sources.

Audio transcription still struggles with accents, background noise, and multiple speakers talking simultaneously. Treat automated transcripts as drafts requiring human review.

Generated content like videos or translated documents should be clearly labeled as synthetic. While these tools can create somewhat realistic output, they're making educated guesses, not reproducing facts.

12 What This Means for Your Career

Multimodal AI is here right now. You need to know how to use what is available today. But you may consider what will be available in the near term future. In 2-3 years, when current students begin practice, expect significantly longer video generation (minutes instead of seconds), reliable real-time transcript analysis during depositions, and voice assistants that can seamlessly handle multiple languages and legal dialects. Document review will likely integrate live translation and cross-modal search as standard features. Within 5 years, we'll probably see AI that can attend meetings virtually, take notes across all modalities, and provide real-time research during live proceedings. While early AR glasses like Google Glass and Apple Vision Pro haven't revolutionized daily work, mature AR systems could transform legal practice by overlaying relevant case law, contract provisions, or evidence analysis directly in your field of view during depositions, court appearances, or client meetings. Imagine reviewing a contract while simultaneously seeing relevant precedents, regulatory guidance, and risk assessments floating next to the clauses you're reading, or conducting a deposition with real-time fact-checking and inconsistency detection visible only to you. The key is learning to orchestrate these tools with judgment and precision, understanding their capabilities and limitations, and maintaining the critical thinking that effective legal work requires regardless of the technology

available.