Project Scoping Submission

Iikshana: ADA Compliant Courtroom Visual Aid for Blind Individuals

'Iikshana': To behold with intent / Observe with care

Course: IE7374 38271 Machine Learning Operations
Team Name: Group 16
Team Members:

- Aditya Vasisht vasisht.ad@northeastern.edu
- Akshata Kumble kumble.a@northeastern.edu
- Amit Karanth Gurpur gurpur.a@northeastern.edu
- Rohit Abhijit Kulkarni kulkarni.rohita@northeastern.edu
- Shridhar Sunilkumar Pol pol.s@northeastern.edu
- Suraj Patel Muthe Gowda muthegowda.s@northeastern.edu

1. Introduction

Access to equitable participation in legal proceedings remains a major challenge for blind and low-vision individuals. While existing accommodations enable access to spoken testimony, they often fail to convey the visual and contextual cues that shape courtroom understanding, such as speaker identity, emotional tone, or interactions between participants. These limitations are further amplified for non-English-speaking users, who must navigate both language barriers and missing visual context. As a result, current assistive approaches relying on human describers or basic accessibility tools fall short of delivering real-time, comprehensive situational awareness.

Our product introduces an AI-powered, real-time courtroom audio assistance platform designed to address these gaps. The system captures live courtroom speech and processes it using Google's SOTA translation models in real time. It integrates multilingual speech recognition, speaker identification, and emotion-aware transcription to produce enriched audio output that conveys not only spoken content but also contextual and emotional information. By translating proceedings into the user's preferred language and embedding cues about tone and interaction, the platform aims to improve accessibility. Our product aims to serve as a replacement to a human translator, thereby aims to be ADA Title II compliant.

Ultimately, the goal is to create a more inclusive courtroom experience for blind / low-vision individuals by transforming audio input into a meaningful, context-rich narrative that helps them better understand and participate in legal proceedings.

2. Dataset Information

2.1 Introduction to the dataset

The platform leverages a combination of publicly available speech datasets to train and evaluate emotion recognition and multilingual speech recognition capabilities. Emotion-aware transcription enhances ADA compliance by conveying not just words but emotional context critical for blind users who cannot observe facial expressions or body language during court proceedings.

Emotion Recognition Datasets

1. RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)
2. IEMOCAP (Interactive Emotional Dyadic Motion Capture)
3. CREMA-D (Crowd-sourced Emotional Multimodal Actors)
4. MELD (Multimodal EmotionLines Dataset)

Supplementary Emotion Datasets

1. TESS: 2,800 files from two female speakers for female voice emotion patterns
2. SAVEE: ~480 utterances from British English speakers for accent variation
3. EMO-DB: 535 German utterances for cross-lingual emotion transfer

Multilingual and Dialect Datasets

1. Mozilla Common Voice
2. VoxPopuli
3. MLS (Multilingual LibriSpeech)
4. FLEURS (Few-shot Learning Evaluation of Universal Representations of Speech)
5. LibriSpeech
6. EmoBox

2.2 Data card

Dataset Category	Dataset Name	Primary Purpose	Content Overview	Emotions / Languages	Key Strengths	Main Limitations	Project Use Case
Emotion (Core)	**RAVDESS**	Clean baseline for recognition	1,440 acted speech clips (24 actors)	8 (Happy, Sad, Angry, Fearful, etc.)	High audio quality; balanced classes	Acted emotions; limited diversity	Baseline benchmarking

Dataset Category	Dataset Name	Primary Purpose	Content Overview	Emotions / Languages	Key Strengths	Main Limitations	Project Use Case
	IEMOCAP	Model conversational dynamics	~12 hrs scripted + improvised dialogue	5 (Happy, Sad, Angry, Neutral, Frustrated)	Natural turn-taking; overlapping speech	Small speaker pool	Conversation-level modeling
	CREMA-D	Reduce demographic bias	7,442 clips (91 actors)	6 (Anger, Disgust, Fear, Happy, Neutral, Sad)	Strong demographic coverage	Acted speech	Bias auditing; stratified evaluation
	MELD	Multi-speaker modeling	13,000+ utterances (TV dialogues)	7 (Anger, Disgust, Sad, Joy, Neutral, etc.)	Realistic dialogue flow	Media dramatization	Multi-speaker interaction modeling
Emotion (Supplementary)	**TESS**	Female voice patterns	2,800 clips (2 female speakers)	Various	Specific gender focus	Very low speaker diversity	Supplementary feature validation
	SAVEE	Accent robustness testing	~480 British English utterances	Various	British accent focus	Small dataset size	Accent variation analysis
	EMO-DB	Cross-lingual transfer	535 German utterances	Various	German language focus	Acted; single language	Language generalization testing
Multilingual	**Mozilla Common Voice**	Multilingual ASR & accent coverage	33,000+ hrs; 133 languages	Multilingual	Largest open corpus; rich metadata	Variable audio quality	Primary ASR training/evaluation
	VoxPopuli	Formal speech modeling	400,000+ hrs; 23 EU languages	EU Languages	Structured; institutional speech	Non-conversational	Courtroom-style modeling

Dataset Category	Dataset Name	Primary Purpose	Content Overview	Emotions / Languages	Key Strengths	Main Limitations	Project Use Case
	FLEURS	Low-resource evaluation	Standardized test sets (102 languages)	102 Languages	Designed for few-shot evaluation	Limited to test/eval focus	Haitian Creole, Khmer, Vietnamese assessment
Benchmarking	**LibriSpeech**	English ASR benchmark	1,000 hrs audiobook speech	English	Gold-standard "clean" data	Read speech only	English ASR performance reference
	EmoBox	Unified benchmarking	32 datasets compiled	14 Languages	Standardized evaluation framework	Aggregated data complexity	Cross-dataset consistency checks

Ethical & Legal Note
Emotion recognition outputs are assistive signals only and are never used for legal decision-making, consistent with ADA, DOJ, and NIST guidance.

2.3 Data sources

The system relies on existing, pretrained Google speech and translation models for all runtime functionality. No model training or fine-tuning is performed / required as part of this project. Publicly available research datasets and open multilingual speech corpora are used only for offline testing, benchmarking, bias analysis, and system validation.

These include:

- Emotion recognition datasets such as RAVDESS, IEMOCAP, CREMA-D, MELD, TESS, SAVEE, and EMO-DB, which provide labeled emotional speech across acted and conversational settings.
- Multilingual speech datasets including Mozilla Common Voice, VoxPopuli, Multilingual LibriSpeech (MLS), FLEURS, and LibriSpeech, which cover a wide range of languages, accents, and speaking styles relevant to courtroom environments.

These datasets are employed solely to evaluate system behavior, validate accuracy, measure bias, and test robustness under controlled conditions.

2.3.1 Operational Data

During deployment, the system processes live courtroom audio streams captured on-prem to provide real-time spoken-language translation and audio accessibility.

- Operational audio is processed transiently for inference only.
- No operational audio is stored, reused for training, or transmitted outside the local court network.
- All processing is performed using pretrained Google models running in an on-prem configuration.

This design preserves data sovereignty, court confidentiality, and regulatory compliance, while ensuring that no real court data influences model behavior beyond immediate inference.

2.4 Data rights & privacy

2.4.1 Dataset Licensing and Usage Rights

All external datasets used in development are open-access or research-licensed, permitting non-commercial academic use. No proprietary or restricted datasets are included. Dataset licenses are respected, and data is used solely within the scope allowed by their terms.

2.4.2 Privacy Protection

- No personally identifiable information (PII) from real court proceedings is collected for training or evaluation.
- Development datasets consist of consented actors, volunteers, or public recordings, with no linkage to real legal cases.
- Any demographic metadata (e.g., gender, accent) is used only for bias auditing and stratified evaluation, not for decision-making.

2.4.3 On-Prem Data Handling

- All live audio processing occurs entirely on-premises within the court's local network.
- No raw audio, derived features, or translated output is transmitted to external services or cloud APIs.
- Access to system outputs and logs is restricted to authorized court personnel only.

2.4.4 Compliance Alignment

These practices align with:

- ADA Title II requirements for effective communication,
- DOJ Language Access Guidance on handling sensitive information,
- Court confidentiality norms, and
- NIST AI Risk Management Framework principles for data governance and auditability.

3. Data Planning & Splits

3.1 Data Pipeline Overview

The system processes live courtroom audio and video streams through minimal, deterministic input pipelines before invoking pretrained Google speech, language, and vision models. All processing occurs on-premises, and no post-processing is applied to model-generated outputs. The system is designed to provide assistive audio translation and emotion cues, with all outputs subject to human oversight.

3.1.1 Input Capture

a. Audio Input

Live audio is captured from courtroom microphones via browser-based interfaces (e.g., WebRTC). Audio may originate from heterogeneous devices with varying formats, sampling rates, and noise characteristics.

b. Video Input (Assistive)

A fixed-angle courtroom camera captures video of speakers for assistive facial emotion detection.
Video capture is limited to real-time analysis and is not used for identity recognition or storage.

3.1.2 Input Audio Preprocessing (Deterministic)

Before model inference, all incoming audio undergoes lightweight, non-learning preprocessing to ensure a consistent and stable input format.

This preprocessing includes:

- Conversion to 16 kHz, mono WAV
- Loudness normalization to a consistent reference level
- Optional trimming of obvious leading and trailing silence
- Basic validation to ensure audio integrity and continuity

This stage performs signal hygiene only and does not attempt noise suppression, enhancement, or content modification.

3.2 Purpose

To provide Google speech models with a standardized audio input independent of microphone or browser variability.

1. Speech-to-Text Inference (Audio Path)

The normalized audio stream is passed to Google Speech-to-Text for real-time transcription. The model internally handles acoustic modeling, noise robustness, speaker characteristics, and confidence estimation.

Output:

- Streaming text transcripts
- Timestamps and confidence scores
- Speaker attribution (where supported)

2. **Language Processing and Translation (Text Path)**

All downstream language processing operates on text, not audio.

- Transcripts are passed to Google Translation / Gemini services
- The system does not assume a known input language; automatic language detection is applied
- Legal terminology is enforced through a domain-specific legal glossary
- Confidence indicators are preserved to support interpreter review

Output:

- Translated text aligned with the original transcript
- Speaker and timing metadata

3. **Vision-Based Emotion Detection (Assistive Path)**

In parallel with audio processing, the video stream is passed to a pretrained facial emotion recognition model running on-prem. Processing includes:

- Lightweight frame sampling
- Face detection and alignment
- Facial expression-based emotion classification with confidence score**s**

Output:

- Timestamped emotion labels (e.g., neutral, distressed, angry, fearful)
- Confidence scores
- Explicit designation as assistive emotion cues

This component is used only to supplement audio-based interpretation, particularly in cases where emotional state may not be evident from speech alone.

4. **Audio Output Generation**

Translated or transcribed text is converted to speech using Google Text-to-Speech.

- Generated audio is played back directly to the user
- No signal modification, filtering, or normalization is applied after generation

This ensures that model-generated audio remains faithful to the original output and avoids unintended distortion.

5. **Fusion, Presentation, and Human Oversight**

Audio-based outputs and vision-based emotion cues are presented side-by-side, without automatic fusion or prioritization. Human interpreters and court staff may:

- Monitor real-time audio translations
- Review confidence indicators and emotion cues
- Override or disregard any AI-generated output

All AI outputs are clearly labeled as assistive, and the system does not present any output as authoritative or legally certified.

3.3 Design Principle & Data Splits

Audio is the primary modality. Vision-based emotion detection is a secondary, assistive signal used to surface potential emotional cues that may not be captured through speech alone. All outputs remain subject to human judgment.

Public research datasets are used only for offline evaluation, calibration, bias analysis, and deployment validation. No live courtroom data is used for training or retained after inference. Data splits are designed to support reliable evaluation and governance, not model training.

1. Development Set (20%)

Purpose: System calibration and integration validation. Used to:

- Calibrate confidence thresholds for flagging content to human interpreters
- Develop and validate legal terminology glossaries
- Validate deterministic audio preprocessing (resampling, normalization, silence handling)
- Verify assistive vision-based emotion detection outputs and confidence labeling
- Perform error analysis across accents, languages, and emotional speech patterns

2. Test Set (70%)

Purpose: Primary performance evaluation. Used to report:

- Speech recognition performance (WER, speaker attribution)
- Translation quality and glossary enforcement
- Vision-based emotion classification accuracy and F1 score
- Latency and robustness under varying noise, lighting, and speaker conditions

Includes stratified analysis by language, emotion class, demographics (where available), and audio/video quality, as well as bias auditing to detect disparate performance.

3. Holdout Set (10%)

Purpose: Final deployment simulation. Used once to:

- Validate end-to-end system behavior
- Confirm generalization beyond the test set
- Verify graceful degradation under rare or extreme conditions
- Measure realistic end-to-end latency and interpreter override responsiveness

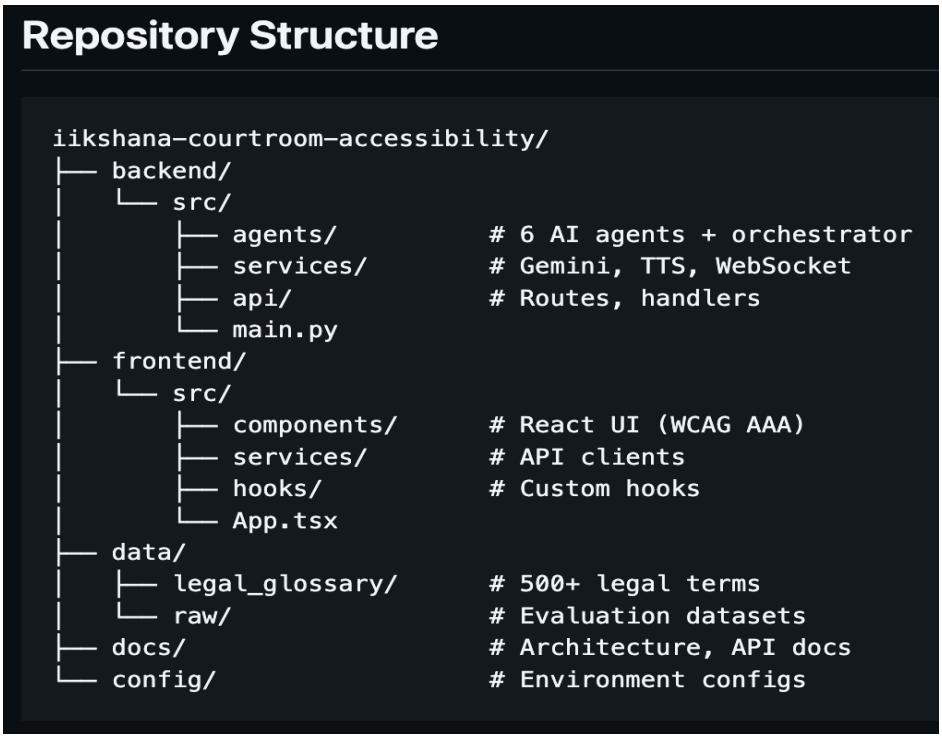
3.4 Stratification Criteria

All splits are stratified to prevent leakage and bias by:

- Speaker identity (no overlap)
- Language
- Emotion class
- Demographics (when available)
- Audio quality (SNR)
- Video conditions (lighting, pose, occlusion)

4. GitHub Repository

Repository - https://github.com/SurajPatelM/iikshana-courtroom-accessibility



5. Project Scope

5.1 Problem Statement

Blind and low-vision individuals face significant barriers when participating in court proceedings. While they can hear spoken testimony, they miss critical visual context: who is speaking, the judge's reactions, witness body language, attorney gestures, evidence displays, and overall courtroom dynamics. Additionally, non-English speaking blind individuals face compounded challenges when proceedings occur in a language they don't understand.

Current accommodations rely heavily on human describers or basic assistive technologies that cannot provide real-time, comprehensive courtroom awareness. This creates an inequitable experience that may impact their ability to fully participate in legal proceedings affecting their lives.

5.2 Project Overview

This project develops an AI-powered real-time courtroom audio assistance platform for blind individuals. The system captures courtroom speech, identifies speakers, detects emotional context, translates content to the user's preferred language, and delivers a comprehensive audio experience that conveys both verbal content and visual dynamics they would otherwise miss.

5.2.1 Core Capabilities

1. Real-Time Speech Transcription & Translation

- Live transcription of all courtroom speech using Google Speech-to-Text (Chirp 3)
- Automatic speaker identification (Judge, Attorney, Witness, Defendant)
- Real-time translation to user's preferred language (supporting 189 languages)
- Audio output via Text-to-Speech in natural, clear voices

2. Emotion-Aware Audio Delivery

- Detection of speaker emotions (anger, fear, sadness, confidence, hesitation)
- Conveying emotional context that blind users cannot observe visually
- Prosody adjustment in synthesized speech to reflect detected emotions
- Confidence indicators for uncertain emotional classifications

3. Courtroom Activity Narration

- Speaker transition announcements ("The witness is now speaking")
- Pause and silence detection with contextual descriptions
- Overlapping speech flagging and resolution
- Session state notifications (recess, adjournment, sidebar)

4. Personalized Audio Experience

- User-selectable target language from 189 supported languages

- Adjustable speech rate and voice preferences
- Volume and clarity optimization for hearing device compatibility
- Real-time audio streaming to personal headphones or earbuds

5.2.2 Target Users

- Blind individuals attending court as defendants, plaintiffs, witnesses, or observers
- Low-vision individuals who cannot read captions or visual displays
- Non-English speaking blind individuals requiring translation
- Court administrators ensuring ADA compliance

5.2.3 Technical Approach

We plan to deploy the platform in two ways. First, by deploying it completely on the edge device, to ensure that we have translation services working even when there is a lack of stable internet connection. The second way employs a cloud based deployment.
The platform deploys on Google Distributed Cloud (GDC) configuration within the courthouse, ensuring all court audio data remains on-premises. The system uses:

- Google Speech-to-Text (Chirp 3): Real-time transcription with speaker diarization
- Google Cloud Translation (Gemini): Context-aware legal translation with custom glossaries
- Google Text-to-Speech: Natural voice synthesis in 380+ voices across 50+ languages
- Emotion Recognition: Custom model using emotion2vec and SenseVoice for speech emotion detection
- Vertex AI: MLOps pipeline for continuous model improvement

5.3 Problems

Problem 1: Inaccessibility of Visual Information in Courtrooms

Blind and visually impaired individuals cannot access critical visual information during courtroom proceedings, including:

- Facial expressions and body language of speakers (judges, attorneys, witnesses)
- Visual evidence presentations (photos, videos, documents displayed on screens)
- Physical demonstrations and gestures
- Document exhibits being referenced
- Courtroom layout and positioning of speakers

This creates a fundamental barrier to equal participation in the legal system, violating the ADA Title II requirements for equal access to judicial proceedings.

Problem 2: Speaker Identification Barriers

In multi-speaker courtroom environments, blind individuals cannot identify:

- Who is currently speaking (judge, prosecutor, defense attorney, witness, jury)
- When speakers change
- Interruptions or simultaneous speech (objections)
- Off-microphone comments or sidebar conversations

Traditional audio-only access provides the words but not the speaker context, making it difficult to follow complex legal arguments and courtroom dynamics.

Problem 3: Missing Emotional and Non-Verbal Context

Courtroom proceedings rely heavily on emotional cues and tone that convey critical information:

- Witness credibility (nervousness, anger, confidence)
- Attorney emphasis and argumentative tone
- Judge's reactions and demeanor
- Emotional weight of testimony (e.g., victim impact statements)

Blind individuals miss these non-verbal cues that sighted participants use to interpret proceedings, leading to incomplete understanding of courtroom events.

Problem 4: Real-Time Information Processing Delays

Current assistive technologies introduce significant delays:

- CART (Communication Access Realtime Translation) has 3-5 second delays
- Screen readers add additional processing time
- Human interpreters may lag behind live proceedings

These delays prevent blind individuals from participating in real-time (e.g., responding to questions, following rapid exchanges during cross-examination).

Problem 5: Legal Compliance Gap with Device Restrictions

Courtrooms have strict security policies that prohibit:

- Personal smartphones and recording devices
- Personal tablets or computers
- Any recording or transmission of courtroom audio/video

This creates a paradox: accessibility technology often requires personal devices, but security requires court-provided devices. Current ADA compliance solutions don't address this constraint, leaving courts unable to provide accessible technology within their security framework.

Problem 6: Cost and Availability of Human Accommodations

Traditional accommodations have significant limitations:

- Human court reporters/CART providers: $300-500 per day, limited availability in rural areas
- Sign language interpreters: Expensive, not helpful for blind individuals (for deaf individuals only)
- Live audio describers: Rare, expensive, require specialized training
- Inconsistent availability: Small courts and rural jurisdictions cannot afford or access human accommodations regularly

Impact Summary:

- 12.2 million blind/visually impaired individuals in the U.S. (NFB, 2023)
- 70% report barriers to accessing legal proceedings (American Foundation for the Blind)
- 40% of courts report inadequate ADA accommodations for blind litigants/witnesses (National Center for State Courts, 2022)
- Legal consequences: Mistrials, appeals, civil rights lawsuits against court systems

5.4 Current solutions

1. Aira - Aira is a professional visual interpreting service that connects blind individuals with highly trained human agents via a smartphone camera. In a courtroom setting, an agent can provide real-time audio descriptions of visual evidence, the layout of the room, and the non-verbal behaviors of witnesses or the jury.

   Strengths:

   - Provides live, nuanced descriptions of visual evidence (photos, physical objects) that AI cannot yet interpret reliably.
   - Offers assistance with navigating complex, non-accessible court e-filing portals and digital evidence management systems.

   Weaknesses:

   - Agents are not "qualified legal interpreters" and cannot provide official linguistic translation or legal advice.
   - Dependence on high-speed courtroom Wi-Fi, which is often restricted or insufficient in older government buildings.

2. Verbit - This solution provides AI-powered, human-verified transcription and real-time captioning specifically for legal environments. It allows blind legal professionals and participants to receive a live text-to-speech stream of courtroom proceedings, ensuring they can follow testimony and legal arguments with the same immediacy as sighted participants. They are ADA Title II compliant.

   Strengths:

   - Delivers "legal-grade" accuracy (99%+) required for official court transcripts and due process.
   - Provides a real-time data stream that integrates with screen readers for immediate audio feedback of spoken testimony.

   Weaknesses:

   - High cost associated with human-in-the-loop verification required for legal standards.
   - Focuses exclusively on spoken word; it cannot describe visual evidence or courtroom gestures.

3. Microsoft Seeing AI - Seeing AI is a mobile application that uses computer vision to describe the environment in real-time. For a blind person in a courtroom, it can read short snippets of text (like a nameplate or room number) and scan longer documents to be read aloud via text-to-speech.

   Strengths:

   - Immediate, local processing of text and documents without the need for a human intermediary.
   - Free to use, making it a highly accessible "reasonable accommodation" for quick visual tasks in court.

Weaknesses:

- AI descriptions of complex legal documents can lack the necessary context for high-stakes litigation.
- Performance can be inconsistent in low-light courtroom environments or with high-gloss evidence photos.

4. Vispero JAWS (Job Access With Speech) - JAWS is the industry-standard screen reader used by blind attorneys and court staff. It is the primary tool for real-time access to digital court records, legal research databases, and word processing. It is the baseline for ensuring a product meets ADA Title II digital accessibility requirements.

   Strengths:

   The most robust tool for navigating complex legal tables, structured court forms, and digital transcripts.

   Supports refreshable Braille displays, allowing a blind person to read legal documents silently during a trial.

   Weaknesses:

   Requires the underlying court software to be perfectly "tagged" and structured for accessibility.

   Has a steep learning curve and high enterprise licensing costs for state government deployment.

5. OneCourt - OneCourt is an emerging tactile technology that converts real-time spatial data into vibrations on a handheld device. While currently marketed for sports, the underlying technology serves as a close alternative for providing blind individuals with real-time spatial awareness of courtroom movements and evidence placement.

   Strengths:

   - Provides a non-auditory way to "feel" the layout and movement within a space in real-time.
   - Reduces "audio fatigue" by allowing the user to track spatial changes through touch rather than constant narration.

   Weaknesses:

   - Currently lacks a dedicated "Legal" or "Courtroom" mode for tracking specific judicial movements.
   - The hardware is specialized and not yet standard equipment in state or local court facilities.

6. Human Interpreters - Courts hire professional interpreters for language translation and occasionally provide human describers who verbally narrate visual courtroom activities for blind attendees.

   Strengths:

- High accuracy for nuanced legal terminology
- Can describe complex visual evidence and non-verbal cues
- Adapts in real-time to unexpected courtroom events

Weaknesses:

- Extremely expensive, limiting availability to high-profile cases
- Shortage of qualified legal interpreters, especially for rare languages
- Human fatigue during long proceedings reduces quality

5.5 Proposed solution: *Iikshana (To behold with intent / Observe with care)*

5.5.1 Solution Overview

Iikshana is a blind-accessible web application powered by Google Gemini 2.0 Flash that transforms courtroom proceedings into a fully accessible experience. A central orchestrator agent coordinates six specialized agents that process audio and visual inputs through two parallel pipelines, delivering emotionally expressive speech with distinctive voices for each speaker.

5.5.2 Agentic Architecture

Central Orchestrator (Gemini 2.0 Flash): Coordinates all specialized agents and manages data flow between pipelines.

Six Specialized Agents:

- Agent 1: Audio Intelligence (transcription, speaker ID, emotion detection)
- Agent 2: Multilingual Translation (language conversion)
- Agent 3: Legal Glossary Guardian (terminology validation)
- Agent 4: Vision Analysis (image captioning - separate pipeline)
- Agent 5: Speech Synthesis Orchestrator (voice assignment, TTS generation)
- Agent 6: Context Manager (conversation history, speaker profiles)

5.5.3 Key Differentiators from Current Solutions

vs. Aira: Eliminates human dependency ($300-500/day), provides 24/7 availability, adds multilingual translation with legal preservation.

vs. Verbit: Extends beyond transcription to include translation, speaker ID, emotion detection, and image captioning at $10-15/session.

vs. Microsoft Seeing AI: Purpose-built for courtrooms with legal terminology preservation and integrated audio-visual processing.

vs. JAWS: Provides active live interpretation; works seamlessly with existing screen readers.

vs. OneCourt: Delivers comprehensive descriptions without specialized hardware.

vs. Human Interpreters: Consistent quality without fatigue, no scheduling constraints, significantly lower cost.

5.5.4 Solution Architecture: Dual Pipeline Design

Pipeline 1: Real-Time Audio Processing (Sequential Flow)
Component 1 → Component 2 → Component 3 → Component 5

Component 1: Audio Intelligence & Multilingual Translation

- Agent 1 (Audio Intelligence): Captures live courtroom audio, transcribes speech to text, identifies speakers by role through contextual analysis, detects emotions from vocal characteristics (pitch, pace, volume, tremor)
- Agent 2 (Multilingual Translation): Translates transcript to target language (Spanish, Mandarin, Vietnamese, etc.) while receiving legal glossary of 500+ terms to preserve
- Agent 3 (Legal Glossary Guardian): Validates no legal terms were mistranslated, automatically corrects violations, adds phonetic pronunciations and brief explanations
- Agent 6 (Context Manager): Maintains conversation history and speaker profiles for improved accuracy

Component 2: Speaker Identification Enhancement

- Refines speaker role assignments using accumulated behavioral patterns
- Maintains persistent speaker-to-role mappings throughout session
- Confidence-based labeling (>85% definitive, 60-85% tentative, <60% generic)

Component 3: Emotion Classification & Communication

- Categorizes detected emotions into 10 states (neutral, angry, fearful, sad, confident, hesitant, agitated, calm, tense, anxious)
- Assigns confidence scores; reports only emotions >75% confidence
- Prepares emotional context for TTS modulation

Component 5: Multi-Voice Speech Synthesis

- Agent 5 (Speech Orchestrator): Assigns persistent distinctive voice to each speaker based on role
- Generates SSML with emotional prosody modulation (adjusts rate, pitch, volume)
- Calls Google Text-to-Speech API with WaveNet/Journey voices
- Streams audio output to user device

Pipeline 2: Visual Evidence Processing (Parallel, Triggered by Image Events)

Component 4: Image Captioning (Separate Pipeline)

Agent 4 (Vision Analysis):

- Triggered when visual evidence is presented (court staff uploads image or camera captures exhibit)
- Multi-stage processing: rapid summary (2-3s) → detailed description (5-10s) → legal relevance (3-5s)
- Uses Gemini 2.0 Flash Vision API for image analysis

- Generates structured captions: short summary, detailed spatial description (left-to-right, top-to-bottom), OCR text extraction, facial expression analysis (if people present), legal significance explanation
- Outputs sent to Agent 5 for conversion to speech

Pipeline Convergence: Unified Output Interface

Component 6: Blind-Accessible Web Interface (PWA)

Receives output from both pipelines and delivers through multiple channels:

From Audio Pipeline (Component 5):

- Emotionally modulated speech with distinctive voices for each speaker
- Explicit verbal announcements of speaker roles
- Real-time streaming via WebSocket

From Visual Pipeline (Component 4 → Agent 5):

- Image caption converted to speech
- Announced when visual evidence is presented
- User can request detailed vs. brief descriptions

Multi-Modal Delivery:

- Primary: Audio output (speech from both pipelines)
- Secondary: Haptic feedback (vibration patterns for speaker changes, objections, judge rulings)
- Tertiary: Optional high-contrast visual display (for partially sighted users)
- Quaternary: Braille display output via WebBluetooth

Interface Features:

- Keyboard-only operation (S: Start/Stop, P: Pause, R: Repeat, E: Explain term, I: Image description, H: Help)
- Voice command support for hands-free control
- ARIA-optimized for screen readers (WCAG 2.1 Level AAA)
- User controls for playback speed, volume, emotion modulation toggle

Technical Implementation

Technology Stack:

- Central Orchestrator: Google Gemini 2.0 Flash (multimodal AI)
- Text-to-Speech: Google Cloud TTS API (WaveNet/Journey voices)
- Frontend: React PWA with TypeScript, Web Audio API, WebSpeech API
- Backend: FastAPI (Python) or Express.js (Node.js) for agent coordination
- Real-time Communication: WebSocket for bidirectional streaming

Processing Flow:

1. Audio captured → Agent 1 processes → Agents 2, 3 enhance → Agent 5 synthesizes → User hears speech
2. Image uploaded → Agent 4 captions → Agent 5 synthesizes → User hears description

3. Both streams merge in Component 6 (PWA interface) for unified delivery

Performance Targets:

- End-to-end latency <1 second for audio pipeline
- Image captioning <10 seconds total (progressive delivery)
- Parallel processing enables simultaneous audio and visual handling

Security & Privacy:

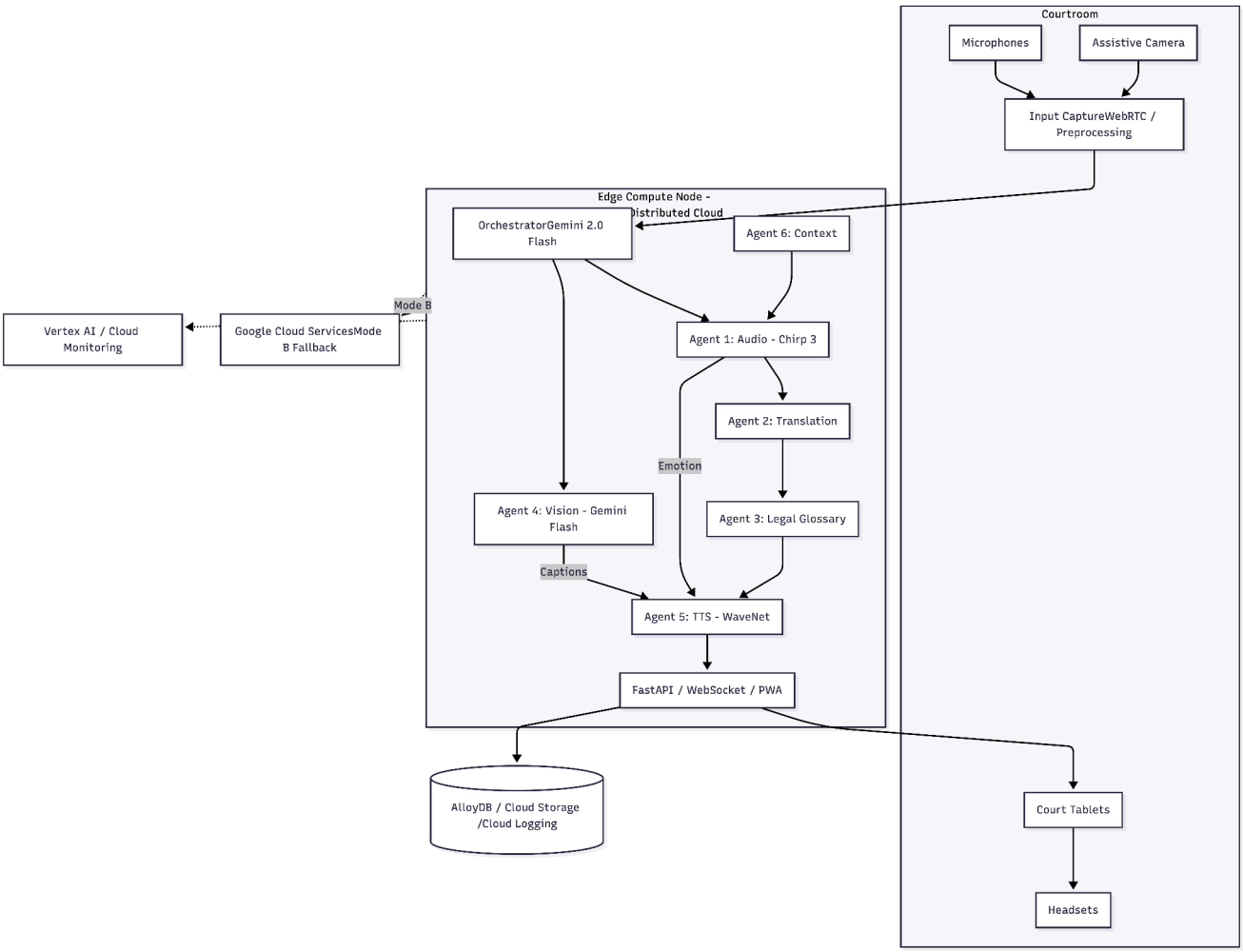
- Real-time-only processing (no audio/transcript storage)
- End-to-end encryption for data transmission
- Privacy-safe logging (hashed data, redaction)
- Operates on court-provided tablets only

This architecture separates the continuous audio processing pipeline from the event-driven image processing pipeline, with both converging at the blind-accessible interface for unified multi-sensory delivery.

6. Current approach - Flow chart & Bottleneck Detection

6.1 Flow Chart

6.1.1 System Architecture Diagram



Overview

This diagram illustrates the end-to-end system architecture of the Iikshana courtroom AI accessibility platform. It shows how real-time courtroom audio and visual inputs are captured, processed through a Gemini-powered multi-agent pipeline on a court-owned edge compute node (Google Distributed Cloud), and delivered as accessible audio output to blind and low-vision court attendees.

Key Components

Courtroom: Microphones capture speech from judges, attorneys, and witnesses. An optional assistive camera captures visual evidence. Court-issued tablets and headsets deliver audio output.

Edge Compute Node (Google Distributed Cloud): The core processing hub containing input capture (WebRTC audio, preprocessing), a Central Orchestrator (Gemini 2.0 Flash), six specialized agents — Agent 1 (Google STT Chirp 3 for transcription, speaker ID, emotion), Agent 2 (Google Cloud Translation), Agent 3 (legal glossary validation), Agent 4 (Gemini 2.0 Flash Vision), Agent 5 (Google Cloud TTS WaveNet), Agent 6 (context management) — and application services (FastAPI, WebSocket, React PWA).

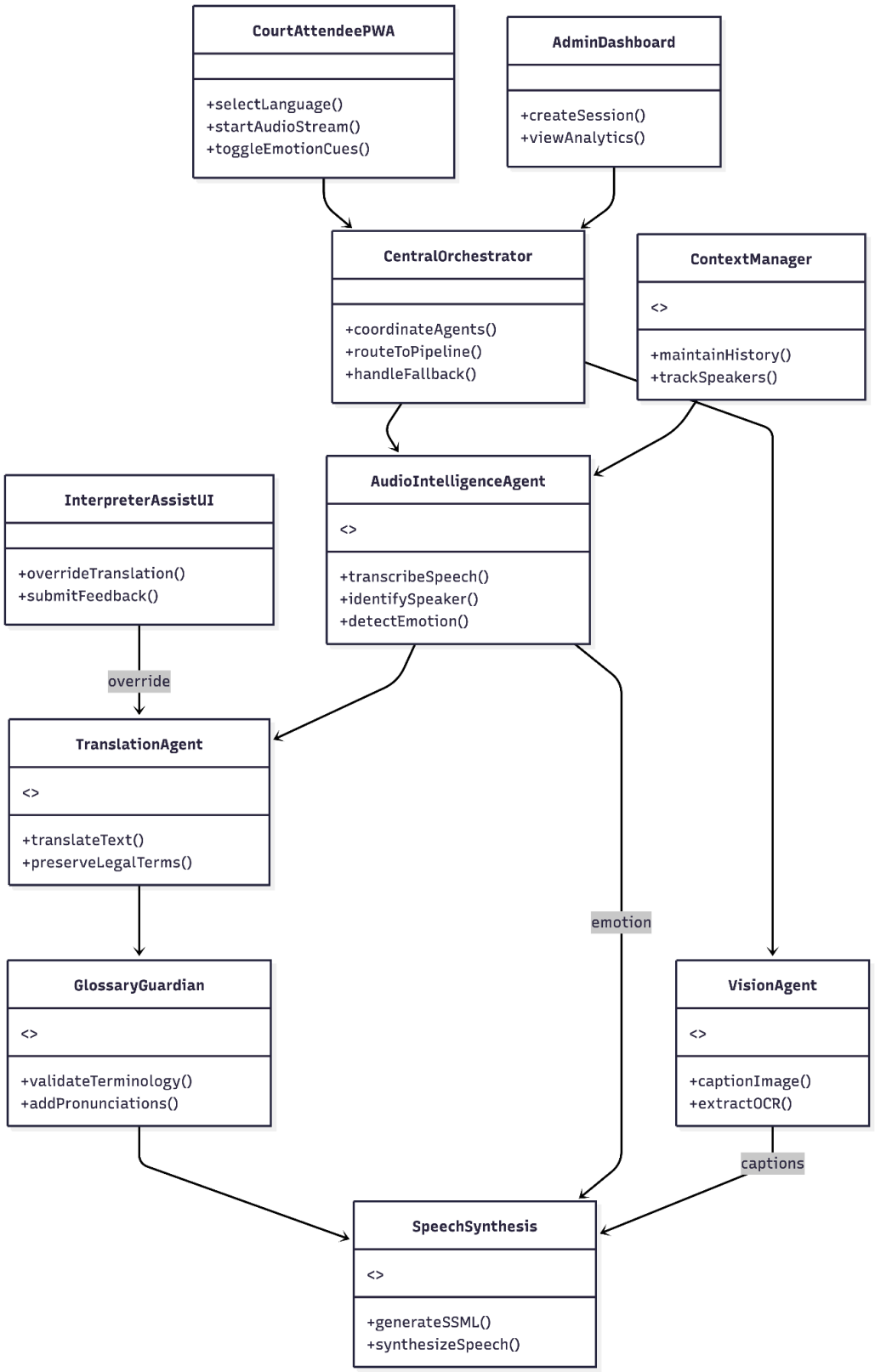
Google Cloud Services (Mode B): Cloud-assisted fallback providing Chirp 3, Cloud Translation, TTS WaveNet, and Gemini Vision API when edge resources are insufficient.

MLOps and Monitoring: Vertex AI for evaluation pipelines and model registry. Cloud Monitoring and Alerting for latency, WER, and confidence tracking.

Data Layer: AlloyDB (sessions, audit), Cloud Storage (artifacts, glossaries), and Cloud Logging (audit trail).

Data Flow: Audio flows from microphones through input capture to the orchestrator, then sequentially through Agent 1 → Agent 2 → Agent 3 → Agent 5, with Agent 6 providing context and Agent 1 passing emotion data to Agent 5. Visual evidence routes in parallel through Agent 4 → Agent 5. Synthesized audio reaches attendees via tablets and headsets. In Mode B, agents offload to Google Cloud services. All metrics feed into Vertex AI and Cloud Monitoring.

6.1.2 UML Class/Component Diagram



Overview
This diagram defines the modular software components of the platform, their key methods, and interconnections. It shows how three user interfaces, the Gemini-powered orchestrator, six Google AI-backed agents, and supporting layers work together.

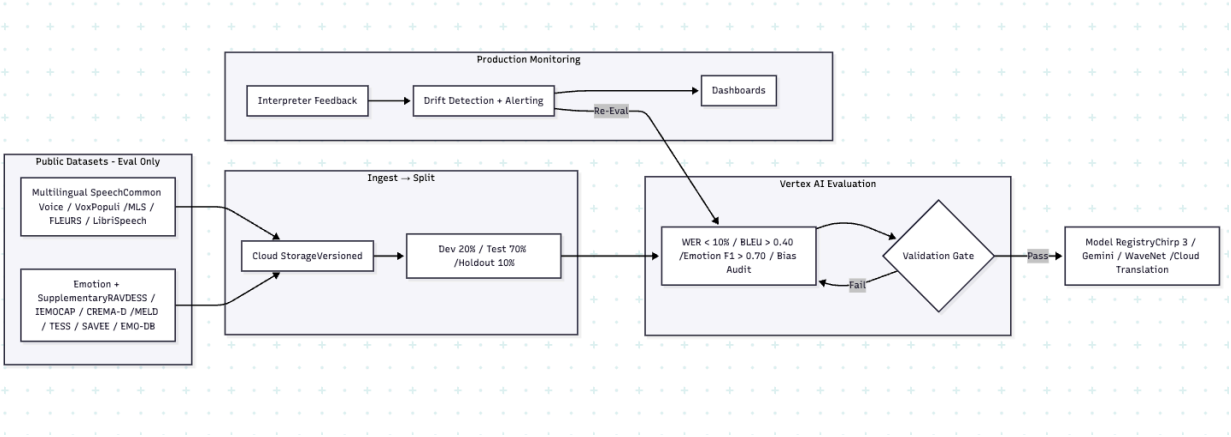
Key Components
Presentation Layer: CourtAttendeePWA (language selection, audio streaming, emotion toggles), AdminDashboard (session creation, analytics), and InterpreterAssistUI (translation overrides, feedback).

Central Orchestrator: Gemini 2.0 Flash coordinates all agents, routes inputs to audio or visual pipelines, and manages fallback between deployment modes.

Agent Layer: Agent 1 (Chirp 3 — transcription, speaker ID, emotion), Agent 2 (Cloud Translation — multilingual translation, legal term preservation), Agent 3 (glossary validation, pronunciations), Agent 4 (Gemini Vision — image captioning, OCR), Agent 5 (TTS WaveNet — SSML generation, speech synthesis), Agent 6 (conversation history, speaker tracking).

Data Flow: The orchestrator routes audio sequentially through Agent 1 → Agent 2 → Agent 3 → Agent 5, with context from Agent 6 and emotion data flowing from Agent 1 to Agent 5. Visual events route through Agent 4 → Agent 5. Interpreter overrides go directly to Agent 2.

6.1.3 MLOps and Data Pipeline - Evaluation, Monitoring, and Drift Detection



Overview
This diagram outlines the MLOps infrastructure for evaluating Google's pretrained models, detecting production drift, and maintaining quality through monitoring. No model training occurs, all datasets serve evaluation and benchmarking only.

Key Components

Public Datasets: Multilingual speech (Common Voice, VoxPopuli, MLS, FLEURS, LibriSpeech) and emotion recognition (RAVDESS, IEMOCAP, CREMA-D, MELD, TESS, SAVEE, EMO-DB) datasets for offline benchmarking.

Ingest and Split: Datasets are stored in versioned Cloud Storage, validated, and stratified into Dev 20% (calibration), Test 70% (WER, F1, BLEU, bias audit), and Holdout 10% (deployment simulation).

Evaluation Pipeline: Vertex AI runs evaluations against targets — ASR (WER < 10%), Translation (BLEU > 0.40), Emotion (F1 > 0.70), and Bias Audit. A validation gate promotes passing models to the registry or loops failures back.

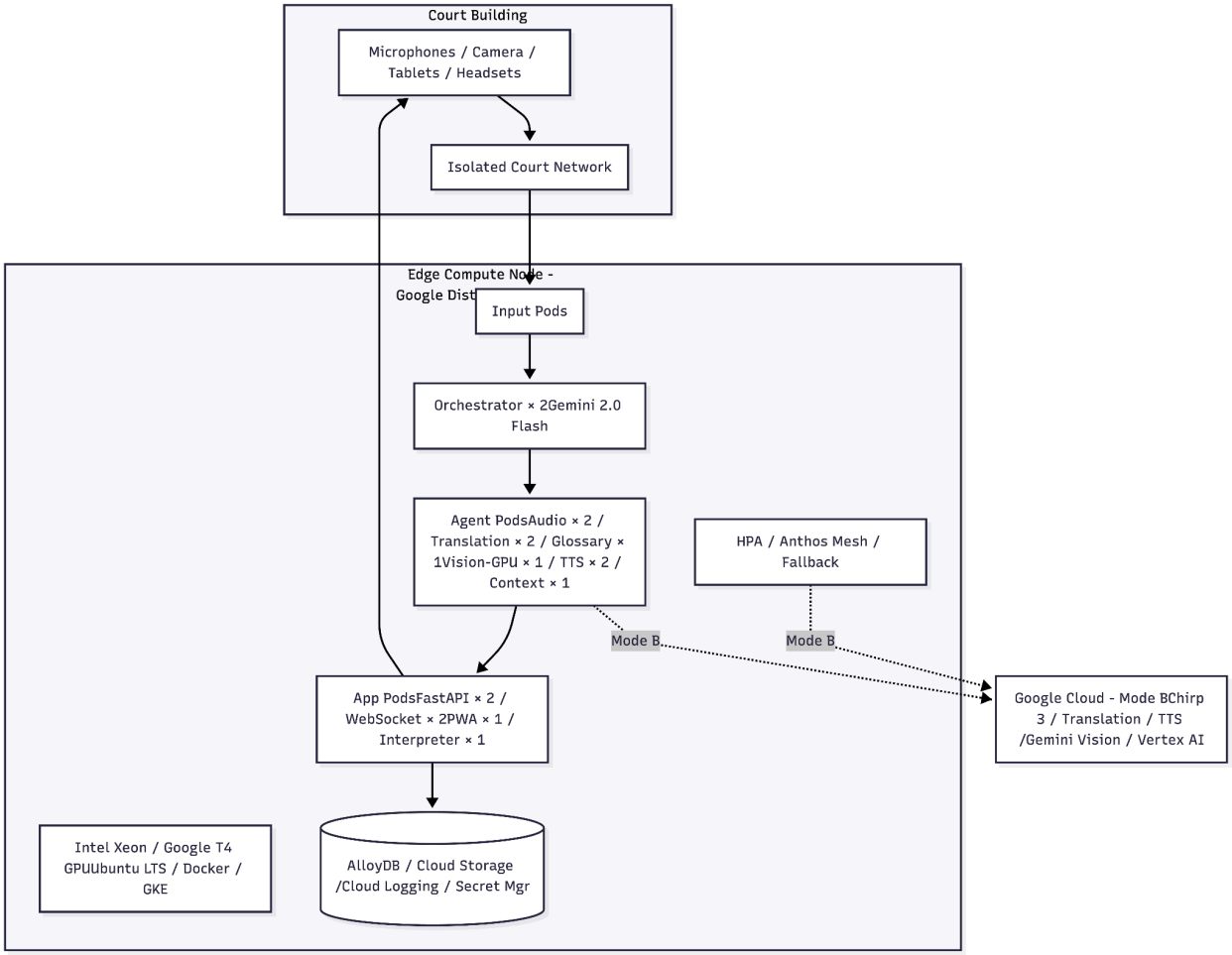
Model Registry: Vertex AI Registry versions Google STT Chirp 3, Gemini 2.0 Flash, Cloud Translation, and Cloud TTS WaveNet.

Production Monitoring: Drift detection tracks changes in audio characteristics, confidence trends, and error patterns. Alerting triggers re-evaluation on threshold breaches. Interpreter feedback feeds back into drift detection and evaluation. Dashboards cover operations, ML performance, and accessibility.

Data Flow

Datasets are ingested, validated, and split into partitions that feed the Vertex AI evaluation pipeline. Results pass through a validation gate into the model registry. In production, drift detectors trigger re-evaluation when thresholds are breached, creating a closed-loop cycle. Interpreter feedback feeds into both monitoring and evaluation.

6.1.4 Deployment Diagram - Edge Mode + Cloud-Assisted Mode



Overview

This diagram maps the physical and containerized deployment of the platform, showing courtroom hardware connecting through an isolated network to the edge compute node on Google Distributed Cloud, with optional Google Cloud fallback in Mode B.

Key Components

Court Building: Microphones, camera, tablets, and headsets connected via an isolated court network.

Edge Compute Node (Google Distributed Cloud): Intel Xeon CPU and Google T4 GPU running Ubuntu LTS with Docker and GKE. Containerized into Input Pods, Orchestrator (Gemini 2.0 Flash × 2), Agent Pods (Audio × 2, Translation × 2, Glossary × 1, Vision-GPU × 1, TTS × 2, Context × 1), App Pods (FastAPI × 2, WebSocket × 2, PWA × 1, Interpreter × 1), with AlloyDB, Cloud Storage, Cloud Logging, and Secret Manager for data services. Platform controls include Horizontal Pod Autoscaler, Anthos Service Mesh, and a Fallback Controller.

Google Cloud (Mode B): Chirp 3, Cloud Translation, TTS WaveNet, Gemini Vision, and Vertex AI activated by the fallback controller when needed. All communication is encrypted and transient.

Data Flow
Courtroom hardware connects through the isolated network to input pods, then the orchestrator routes to agent pods (sequentially for audio, parallel for vision), through app pods, and back to tablets and headsets. In Mode B, the fallback controller enables encrypted connections to Google Cloud services.

6.2 Bottleneck Detection

1. Legal & Compliance Bottlenecks
    a. Ambiguity around "no recording": Real-time processing vs. transient buffering is legally sensitive.

2. Audio & Environmental Complexity
    a. Overlapping speech: Judges, attorneys, witnesses, and interruptions
    b. Courtroom acoustics: Echo, distance from speaker, side conversations
    c. Emotionally charged speech: Raised voices, sarcasm, urgency, hard for ASR + NLP to interpret accurately

3. Context & Situational Awareness
    a. Non-verbal events: Objections, gestures, reactions, jury behavior
    b. Document handling: Exhibits bare being referenced verbally but not clearly identified visually
    c. Temporal ambiguity: "As stated earlier, "this document", "that image", without context anchoring

Improvements in the current process

1. Device & UX Improvements
    a. Single-purpose hardened device
    b. No general OS access: A custom UI can be designed for minimal and efficient handling by the users for the least friction in the
    c. Physical buttons for:
        i. Repeat
        ii. Pause
        iii. Detail level
        iv. Emergency mute

2. Evaluation & Trust
    a. Explainability layer
    b. Short clarifications:
        i. “This is a judge’s ruling.”
        ii. “Attorney is referencing a document.”
    c. Builds user trust and confidence
    d. Human-in-the-loop fallback
    e. The court accessibility officer can intervene if the system degrades

7. Metrics, Objectives & Business Goals

7.1 Metrics

7.1.1 Key Performance Indicators (KPIs)

System reliability

- System uptime ≥ 99.5% (measured by on-prem monitoring)
- Support for ≥ 50 concurrent users per courtroom under load

User satisfaction

- User satisfaction score > 4.5 / 5 (post-session surveys)

Translation quality

- Interpreter override rate < 20% (measured from human interpreter feedback)

Cost efficiency

- Processing cost < $0.10 per minute (resource usage estimation)

Latency

- End-to-end audio-to-audio latency: p95 ≤ 2–3 seconds
- TTS latency: p95 ≤ 1.5 seconds

Legal and compliance

- Legal glossary enforcement rate > 95%
- Zero data exfiltration (no outbound transmission of raw audio or transcripts)

7.1.2 Performance Metrics

Speech recognition

- WER: sampled streaming transcriptions vs. reference transcripts
- Speaker attribution: compared to ground truth when available
- Confidence distribution: per-segment confidence scores for flagging

Translation

- Glossary enforcement rate: % of glossary terms correctly preserved
- Confidence scores: per-segment confidence for review triggers

Emotion detection

- F1 score: on held-out validation set (target > 0.70)
- Bias audit: stratified performance by demographics (e.g., CREMA-D)
- Calibration: confidence calibration for emotion predictions

Accessibility

- Audio completion rate: % of finalized content delivered as audio (target 100%)
- TTS latency: p50, p95, p99 from text finalization to audio playback
- WCAG compliance: automated and manual audits

System health

- Latency percentiles (p50, p95, p99) for all API endpoints
- Error rates by service and error type
- Throughput and queue depth
- GPU utilization and memory use

7.2 Objectives

7.2.1 Strategic Objectives:

1. Real-time language coverage

Provide live translation for at least the top 10 languages used in Massachusetts courts (e.g., Spanish, Haitian Creole, Mandarin, Vietnamese, Portuguese, Arabic, Russian, Khmer, Cantonese, Turkish) within the first phase.

2. Speaker and emotional context

Identify speakers by role (Judge, Attorney, Witness, Defendant) and provide optional emotional context cues when confidence is sufficient, clearly marked as assistive.

3. Audio-first user experience

Deliver 100% of finalized spoken content as audio. Support real-time streaming with latency targets (e.g., 2–3 seconds end-to-end).

7.2.2 Technical Objectives:

Speech recognition and transcription

- Word Error Rate (WER) < 10% on courtroom-style audio
- Speaker diarization / attribution accuracy > 90% when ground truth is available
- End-to-end audio-to-audio latency ≤ 2 seconds (target), ≤ 3 seconds (maximum)
- Automatic flagging of low-confidence segments for interpreter review

Translation

- Legal glossary enforcement rate > 95% for covered terms
- At least 80% of translated segments understandable without correction
- Translation latency ≤ 1 second after transcript finalization
- Automatic language detection for input; user-selectable output language

Image captioning (visual evidence pipeline)

- Progressive caption delivery: short summary (2–3 s), detailed description (5–10 s), legal relevance (3–5 s)
- Captioning latency ≤ 10 seconds end-to-end for visual evidence

Accessibility and output

- Audio output for 100% of finalized spoken content
- TTS latency p95 ≤ 1.5 seconds after text finalization
- Automated WCAG audits for any visual components; no regression from baseline

7.3 Business Goals

1. Expand Language Access

Enable real-time spoken-language interpretation for blind users who can hear, supporting the top languages used in Massachusetts courts, so that non-English-speaking participants can understand and participate in proceedings through audio output.

2. Achieve ADA-Compliant Audio Communication

Provide effective communication through audio-only outputs for blind and low-vision users who can hear, consistent with ADA Title II requirements, without reliance on visual interfaces, or text-based accessibility features.

3. Maintain Legal Accuracy

Ensure that all spoken-language translations preserve legal meaning by enforcing custom legal glossaries and requiring human-in-the-loop oversight, preventing AI outputs from being treated as authoritative or certified interpretations.

4. Protect Data Sovereignty

Ensure that all audio processing and translation occur on-premises, with no external network transmission, preserving court confidentiality and compliance with regulated data-handling expectations.

5. Improve Courtroom Efficiency

Minimize delays caused by interpreter availability or setup, enabling hearings to proceed on schedule while maintaining accessibility standards. This matters because delays increase costs, backlog, and frustration for judges, staff, and litigants.

6. Enable Scalable Language Access

Create a platform that can be deployed across multiple courtrooms and locations with predictable costs and performance, without linear increases in staffing. This would help courts solve the need solutions that scale across jurisdictions and caseloads.

8. Failure Analysis

8.1 Speech Recognition Failures

- High background noise in courtroom exceeds model tolerance
- Overlapping speech from multiple speakers
- Strong accents or dialects not represented in training data
- Technical legal terminology not in vocabulary

Mitigation: Acoustic adaptation to courtroom environments, explicit flagging of low-confidence segments, continuous vocabulary expansion from interpreter feedback.

8.2 Translation Failures

- Legal phrases mistranslated due to lack of context
- Idioms translated literally instead of semantically
- Rare language pairs with insufficient training data
- Glossary misses for jurisdiction-specific terminology

Mitigation: Domain-specific legal glossaries, confidence thresholds requiring human review, interpreter override capabilities, quarterly glossary updates.

8.3 Emotion Detection Failures

- Cultural differences in emotional expression
- Acted vs. natural speech distribution mismatch
- Neutral speech misclassified as emotional
- Demographic bias in model predictions

Mitigation: Calibrate confidence thresholds, optional emotion display (user preference), regular bias audits.

8.4 Accessibility Output Failures

- Audio narration mispronouncing legal terms or names
- Screen reader incompatibility with dynamic content

Mitigation: Validation against SSML pronunciation guides for legal terms, automated WCAG compliance testing.

8.5 System Failures:

- GPU resource exhaustion under peak load
- Database connection pool depletion
- WebSocket connection drops during proceedings

Mitigation: Horizontal pod autoscaling, connection pooling, graceful degradation with cached responses, automated failover.

9. Deployment Infrastructure

9.1 Overview

The proposed platform supports two complementary deployment configurations to accommodate heterogeneous courtroom network conditions and operational constraints:

1. Fully Edge-Deployed Mode – ensures continuous accessibility services under limited or unstable network connectivity.
2. Cloud-Assisted Mode – enables full multilingual coverage and higher translation fidelity when reliable connectivity is available.

Both configurations preserve functional parity at the user interface level and rely exclusively on court-owned hardware and court-approved software environments.

9.1.1 Deployment Mode A: Fully Edge-Deployed (Offline-Capable)

The edge-only configuration executes the complete processing pipeline locally on a courthouse edge compute node.

Key Characteristics

- All inference (speech recognition, translation, emotion detection, and text-to-speech) executes on-device.
- Designed for continuity of service during network outages.
- Supports a reduced but jurisdiction-relevant language set.
- Manual model updates performed by court IT administrators.

Use Case

- Rural or bandwidth-constrained courthouses.
- Proceedings requiring uninterrupted ADA accommodations regardless of network availability.

9.1.2 Deployment Mode B: Cloud-Assisted (Hybrid)

The cloud-assisted configuration partitions processing between the local edge node and approved cloud services.

Key Characteristics

- Audio capture, preprocessing, orchestration, and fallback logic remain on the edge.
- Speech-to-text, translation, and speech synthesis may be delegated to cloud services.
- Encrypted, transient data exchange; no persistent storage of courtroom data.
- Enables broad language coverage and improved translation quality.

Use Case

- Urban courts with stable connectivity.
- Proceedings involving rare languages or high translation complexity.

9.2 Common Physical Infrastructure

1. **Courtroom Inputs**
    a. Existing courtroom microphone systems
    b. Optional fixed-angle assistive camera
2. **Edge Compute Node**
    a. Court-owned server (CPU + GPU)
    b. Executes preprocessing, orchestration, and fallback control
3. **User Access Device**
    a. Court-issued tablet or terminal
    b. Headset or assistive audio device
    c. Browser-based progressive web application (PWA)

9.3 Software Infrastructure Stack

1. **Operating System**
    a. Ubuntu Server LTS or Red Hat Enterprise Linux
2. **Containerization**
    a. Docker
3. **Orchestration**
    a. Kubernetes (optional, multi-court deployments)
4. **Backend Services**
    a. FastAPI or Express.js for agent coordination
5. **Frontend**
    a. React-based PWA
    b. WCAG 2.1 AA/AAA compliant
6. **Communication**
    a. WebRTC (audio capture)
    b. WebSockets (real-time audio streaming)

9.4 Reliability and Failover

1. Continuous monitoring of network health.
2. Automatic fallback from cloud-assisted to edge-only mode upon connectivity degradation.
3. No session interruption during mode switching.
4. Users are explicitly notified when operating in reduced-capability mode.

9.5 Security and Compliance

1. Encryption in transit for all cloud communications.
2. No persistent storage of audio, transcripts, or video.
3. Court-controlled access to hardware and system interfaces.
4. All outputs explicitly labeled as assistive and non-authoritative.
5. Alignment with ADA Title II effective communication requirements and court confidentiality norms.

9.6 Supported Platforms

1. **Compute Environments**
    a. On-premises edge servers
    b. Private cloud infrastructure
    c. Approved public cloud platforms
2. **Operating Systems**
    a. Ubuntu Server LTS
    b. Red Hat Enterprise Linux
3. **Browsers**
    a. Chrome
    b. Edge (Chromium-based)
4. **Assistive Hardware**
    a. Headsets
    b. Optional Braille displays (via WebBluetooth)

10. Monitoring Plan

10.1 Real-Time Monitoring

10.1.1 Service Health

- Track latency percentiles (p50, p95, p99) for all API endpoints
- Alert when p99 latency exceeds 3 seconds for 5 minutes
- Track error rates by service and error type; alert if error rate exceeds 2% over 5 minutes
- Track throughput and queue depth for congestion
- Monitor GPU utilization and memory and alert if utilization stays above 90% for 10 minutes

10.1.2 Model Performance

- Evaluate Word Error Rate on sampled streaming transcriptions; alert if WER increases >15% vs baseline

- Track translation confidence per segment; alert if mean confidence drops below 0.75
- Run periodic emotion classification F1 on a validation set (target >0.70)
- Check speaker diarization accuracy when ground truth is available

10.1.3 Accessibility Metrics

- Track audio narration completion (target 100% of finalized content)
- Track TTS latency (target p95 <1.5 seconds after text finalization)
- Run automated WCAG audits for screen-reader compatibility; alert on regression

10.2 Drift Detection

10.2.1 Feature Drift

- Audio characteristic distributions (SNR, duration, frequency spectrum)
- Language distribution changes
- Speaker demographic shifts

10.2.2 Prediction Drift

- Confidence score trends over time
- Error pattern changes
- Human override frequency increases

10.3 Dashboards

- Operations dashboard: System health, active sessions, error trends, resource use
- ML performance dashboard: WER, translation and emotion confidence distributions, drift indicators
- Accessibility dashboard: Audio usage, screen-reader compatibility, override rates

11. Success & Acceptance Criteria

This section defines the conditions under which the Iikshana system will be considered successful, acceptable, and deployable within a courtroom environment. Criteria are aligned with ADA Title II effective communication, court operational constraints, and system reliability expectations.

11.1 Accessibility & ADA Compliance Acceptance

Objective: Ensure blind and low-vision users who can hear receive effective, timely, and usable communication during court proceedings.

Acceptance Criteria

1. Audio output is available for 100% of finalized spoken courtroom content
2. Audio delivery is real-time, with no reliance on visual interfaces or text displays

3. All AI-generated outputs are clearly labeled as assistive, not official or certified
4. System can be operated using court-provided devices only, without personal smartphones
5. Blind users can:
    a. Identify who is speaking
    b. Understand what is being said
    c. Receive emotional context cues where confidence is sufficient
6. System behavior satisfies ADA Title II's requirement for effective communication, not merely technical access

Failure Conditions

1. Audio output is delayed enough to disrupt participation
2. Key speech content is omitted or inaccessible
3. Users are misled into treating AI output as authoritative
4. Accessibility depends on prohibited personal devices

11.2 Language Access & Translation Acceptance

Objective: Provide lawful spoken-language access for non-English-speaking blind users.

Acceptance Criteria

1. System supports the most commonly used languages in Massachusetts courts
2. Input language is automatically detected; no prior user configuration required
3. Spoken translations preserve legal meaning, supported by:
    a. Domain-specific legal glossaries
    b. Human-in-the-loop override capability
4. More than 80% of translated segments are understandable without correction, while allowing interpreter review

Failure Conditions

1. Legal terminology is altered or mistranslated silently
2. System presents translations as "official" or "certified"
3. Human interpreters cannot override AI output

11.3 Emotional & Contextual Awareness Acceptance

Objective: Supplement spoken content with assistive emotional context when appropriate.

Acceptance Criteria

1. Emotion cues are provided only when confidence exceeds defined thresholds
2. Emotion outputs are explicitly labeled as assistive context
3. Vision-based emotion detection does not perform identity recognition and does not store video

Failure Conditions

1. Emotion outputs are presented without uncertainty indicators
2. Emotion cues are treated as factual assessments
3. Vision processing exceeds assistive scope

11.4 Performance & Responsiveness Acceptance

Objective: Ensure the system functions in real courtroom conditions without disrupting proceedings.

Acceptance Criteria

1. End-to-end audio-to-audio latency ≤ 2–3 seconds
2. System remains usable under:
    a. Overlapping speech
    b. Elevated noise levels
    c. Rapid speaker changes
3. System degrades gracefully by:
    a. Flagging uncertainty
    b. Prioritizing intelligibility over completeness

Failure Conditions

1. Latency prevents real-time participation
2. System freezes or fails silently
3. Errors are hidden from users or interpreters

11.5 Privacy, Security & Court Policy Acceptance

Objective: Respect courtroom security rules and confidentiality.

Acceptance Criteria

1. All processing occurs on-premises
2. No raw audio, video, or derived outputs are transmitted externally
3. No operational data is stored beyond transient inference
4. Access is restricted to authorized court personnel
5. System operates on court-issued devices only

Failure Conditions

1. Any data leaves the local network
2. Persistent recording or storage occurs
3. System requires personal devices

11.6 Operational Acceptance by Courts

Objective: Ensure the system is practical for court adoption.

Acceptance Criteria

1. Court staff can deploy and operate the system with minimal training
2. Interpreter override and fallback procedures are clearly defined
3. System availability ≥ 99.5% uptime

Failure Conditions

1. System increases staff workload
2. Requires specialized hardware not typically available
3. Introduces legal or operational risk

12. Timeline Planning

Phase	Weeks	Focus Area	Key Activities	Deliverables
Phase 1	**1–2**	**Problem Definition & Compliance**	**Define scope, user personas, ADA Title II mapping, courtroom constraints analysis**	**Finalized problem statement, compliance boundaries**
Phase 2	**3–4**	**Architecture & Data Planning**	**Dual-pipeline design, edge + hybrid deployment planning, dataset selection, data governance**	**System architecture diagrams, data cards, privacy plan**
Phase 3	**5–6**	**Core Audio Pipeline Prototype**	**Audio capture, ASR (Chirp 3), speaker diarization, translation, TTS**	**Working audio-to-audio prototype, latency & WER baseline**
Phase 4	**7–8**	**Emotion & Context Awareness**	**Speech-based emotion detection, confidence gating, optional vision cues, bias analysis**	**Emotion-aware audio output, confidence calibration report**
Phase 5	**9–10**	**Accessibility UX & Device Constraints**	**Blind-accessible PWA, keyboard/voice controls, multi-voice output, device hardening**	**WCAG checklist, court-device UX demo**
Phase 6	**11–12**	**Evaluation & MLOps**	**Offline evaluation, bias audits, drift scenarios, monitoring setup**	**Evaluation report, failure analysis, monitoring plan**
Phase 7	**13–14**	**Deployment Simulation & Final Review**	**Edge + cloud-assisted simulation, interpreter override demo, documentation**	**Deployment diagrams, final report, recorded demo**

NOTE: Since the proposal was re-designed to suite Google / Government of Massachusetts requirements, the submission is late by a few days. We are currently in phase 3 of the project.

13. Additional Information

13.1 Assumptions

1. **Court-Provided Hardware:** The courthouse will provide dedicated tablets, headsets, and assistive audio devices to blind and low-vision attendees. No personal smartphones or recording devices are permitted in the courtroom, consistent with existing court security policies.

2. **Microphone Infrastructure:** The courtroom is assumed to have a functional microphone setup (e.g., individual microphones at the judge's bench, attorney podiums, and witness stand). The system does not provide or install its own microphone hardware.

3. **Network Availability:** For Mode A (edge-only), a stable local area network within the courthouse is assumed. For Mode B (cloud-assisted), reliable internet connectivity with sufficient bandwidth for encrypted API calls to Google Cloud is required. The fallback controller handles degradation between modes automatically.

4. **Google Service Access:** The project assumes continued access to Google's pretrained models (STT Chirp 3, Cloud Translation, Cloud TTS WaveNet, Gemini 2.0 Flash) through their respective APIs. No model training or fine-tuning is performed — the system relies entirely on Google's pretrained capabilities.

5. **Language Selection:** Users are assumed to know their preferred language in advance and can select it at session start. The system supports automatic input language detection but requires the user to specify their desired output language.

6. **Human Interpreter Availability:** A qualified human interpreter or court accessibility officer is assumed to be available during proceedings to monitor system output, perform translation overrides when necessary, and serve as the human-in-the-loop fallback.

7. **Single Courtroom Scope:** Each edge compute node is assumed to serve a single courtroom at a time. Multi-courtroom deployments would require separate edge nodes or Kubernetes-based orchestration, which is supported but not the primary deployment target.

8. **Audio as Primary Modality:** The system assumes that blind and low-vision users who use this platform can hear. Audio output (speech) is the primary delivery channel. Users

who are both blind and deaf are not within the current scope (see Out of Scope below).

9. **Emotion as Assistive Only:** Emotion detection outputs are treated as assistive contextual cues, not factual assessments. Users, interpreters, and court staff are assumed to understand that emotion labels carry uncertainty and are never used for legal decision-making.

10. **No Courtroom Modifications:** The system is designed to operate within existing courtroom infrastructure without requiring physical modifications, rewiring, or specialized installations beyond the edge compute node and standard network equipment.

13.2 Out of Scope

1. **Model Training and Fine-Tuning:** The system does not train, fine-tune, or adapt any AI models. All models are Google's pretrained offerings used as-is. Public datasets are used exclusively for offline evaluation, benchmarking, and bias auditing.

2. **Persistent Storage of Court Proceedings:** No audio, video, transcripts, or translated outputs are stored beyond transient inference. The system does not function as a court recording or transcription archival tool.

3. **Official Legal Translation or Certification:** All AI-generated translations are labeled as assistive and are not legally certified. The system is not a replacement for certified court interpreters in proceedings where official certified translation is legally mandated.

4. **Identity Recognition:** The assistive camera and vision pipeline are used solely for evidence captioning and facial expression-based emotion cues. The system does not perform facial recognition, identity verification, or biometric identification of any kind.

5. **Jury Monitoring or Behavioral Analysis:** The system does not observe, analyze, or report on jury behavior, reactions, or body language. Vision capabilities are limited to evidence captioning and speaker-area emotion cues.

6. **Personal Device Support:** The system runs exclusively on court-issued devices. There is no support for personal smartphones, tablets, laptops, or wearable devices, in compliance with courtroom security policies.

7. **Non-Google Services:** The project is constrained to Google products and services only. Third-party AI models (e.g., OpenAI, AWS, Azure services), non-Google speech or

translation engines, and external NLP tools are not used.

8. **Offline Transcript Editing or Post-Session Review:** The system provides real-time assistance only. There is no post-session transcript review, editing, or download functionality, as no session data is retained.

9. **Multi-Courtroom Orchestration:** While the architecture supports Kubernetes-based scaling, the current project scope targets single-courtroom deployment. Multi-courtroom orchestration across a courthouse complex is an architectural consideration but not a deliverable.

10. **Automated Legal Decision Support:** The system provides accessibility and communication aids only. It does not analyze legal arguments, predict case outcomes, assess witness credibility, or provide any form of legal reasoning or recommendation.

13.3 Why We Focus on Blind and Low-Vision Individuals

The decision to scope Iikshana specifically for blind and low-vision court attendees — rather than addressing a broader range of disabilities — is grounded in the following rationale:

1. An Underserved and Compounded Accessibility Gap While courtroom accessibility solutions exist for deaf and hard-of-hearing individuals (CART, sign language interpreters, real-time captioning), blind and low-vision individuals face a fundamentally different and largely unaddressed challenge. They can hear spoken testimony but miss all visual context — who is speaking, facial expressions, body language, visual evidence displays, and courtroom dynamics. Existing tools like screen readers and human describers provide fragmented assistance but fail to deliver real-time, comprehensive situational awareness. This gap is further compounded for non-English-speaking blind individuals who face both language barriers and missing visual context simultaneously.

2. ADA Title II Compliance Focus ADA Title II requires state and local governments — including courts - to provide "effective communication" to individuals with disabilities. For blind individuals who can hear, effective communication means delivering not just words but contextual, emotional, and situational information through audio. This is a distinct technical challenge from captioning or sign language interpretation, requiring a purpose-built solution that translates visual courtroom dynamics into rich audio narratives.

3. Audio-First Architecture Enables Depth By focusing on a single primary output modality (audio), the system can invest deeply in the quality of that experience — emotion-aware prosody, distinctive speaker voices, legal terminology preservation, real-time translation, and progressive image descriptions. Attempting to simultaneously serve visual, auditory, cognitive, and motor

disabilities would dilute the solution across multiple output modalities (captions, sign language avatars, simplified language, motor-accessible interfaces) and risk delivering a mediocre experience for all rather than an excellent one for the target population.

4. Scale of Impact According to the National Federation of the Blind (2023), approximately 12.2 million individuals in the United States are blind or visually impaired. The American Foundation for the Blind reports that 70% of these individuals encounter barriers when accessing legal proceedings, and the National Center for State Courts (2022) found that 40% of courts report inadequate ADA accommodations for blind litigants and witnesses. This represents a significant population with a clear, documented need.

5. Existing Solutions Address Other Disabilities More Effectively Courtroom accessibility for other disability categories, while still imperfect has more established tooling:

- **Deaf / Hard of Hearing:** CART services, real-time captioning, sign language interpreters, and assistive listening systems are widely deployed and legally mandated.
- **Motor / Mobility:** Physical courtroom accommodations (wheelchair access, accessible seating) are addressed through building codes and ADA Title III.
- **Cognitive / Intellectual:** Simplified language and plain-language summaries are emerging areas but require fundamentally different AI approaches (text simplification, reading-level adaptation) that diverge from this project's audio-centric architecture.

Blind and low-vision individuals remain the group with the least specialized real-time technology available in courtroom settings, making this the area where an AI-powered solution can deliver the greatest incremental impact.

6. Future Extensibility While the current scope targets blind and low-vision users, the platform's modular agentic architecture is designed for extensibility. The same pipeline that produces audio output could be extended in future iterations to support real-time captioning (for deaf users), simplified language output (for cognitive accessibility), or haptic feedback (for deafblind users) by adding new output agents without redesigning the core processing pipeline.