Safe-by-Design Governance for LLM-Assisted Emergency Call Triage: Auditability, Human Oversight, and Degraded-Mode Resilience

Anonymous Authors
Affiliation Omitted for Blind Review

City, Country
anonymous@domain.org

***Abstract*—The integration of Large Language Models (LLMs) into Emergency Call Centers (112/911) promises to mitigate operator cognitive overload but introduces critical risks regarding hallucination, bias, and accountability. This paper presents "secure-IAppel", a comprehensive *Safe-by-Design* governance framework for deploying AI assistance in life-critical environments. Moving beyond standard performance metrics (WER), we formalize governance as an architectural constraint and propose a hybrid pipeline combining a low-latency "Hot Path" for real-time transcription with a verificational "Cold Path" for safety auditing and retrospective corrections. In the IAppels project, this pipeline targets a unified emergency number (15/17/18/112) and operationalizes real-time STT (Whisper), language identification with translation, diarization, entity extraction, and AI-assisted decision trees for emergency call classification, under a sub-second latency budget. We define a rigorous "Epistemic State Matrix" (Green/Orange/Red/Abstain) to manage algorithmic uncertainty, bound automation bias, and enforce effective Human-in-the-Loop oversight through explicit validation hooks. We further introduce standardized "Reason Codes" that make abstentions and warnings auditable, and we link these logs to a privacy-aware evidence chain compatible with GDPR and emerging AI-act requirements. Finally, we show how explicit degraded modes preserve operational continuity under stress, ensuring that LLM assistance strengthens the human command chain rather than replacing it. The result is a governance-first blueprint grounded in the ongoing IAppels field experimentation (SDIS 01, 2025) that reconciles innovation with legal accountability and the ethical imperatives of public service.**

***Index Terms*—Emergency Services, LLM Governance, AI Safety, Auditability, Human-in-the-Loop, Resilient Architectures, Public Safety, NG112/NG911.**

I. Introduction

Public Safety Answering Points (PSAPs) worldwide are facing a structural crisis. From the NG911 transition in North America to the NG112 mandate in Europe, the volume and complexity of emergency calls are exploding. Initiatives like the Visual 112 in Luxembourg (CGDIS, 2025), the "Crime Check" AI in Spokane (USA), or "Call-Bots" in Seoul (119) demonstrate a global race towards automation. Beyond call volume, the signal itself is deteriorating: callers are more mobile, more multilingual, and often in chaotic environments where location, symptoms, and intent are harder to parse. At the same time, staffing constraints and the push toward unified emergency numbers (15/17/18/112) create pressure for a shared, interoperable triage layer that can scale without sacrificing quality of service.

However, integrating Artificial Intelligence into this "High-Stakes" environment introduces unprecedented risks. The challenge is not merely technical (ASR accuracy) but cognitive and ethical. Operators are already saturated by the "Cocktail Party Problem" (overlapping speech), acute stress (OODA loop constraints), and the need to keep a stable decision chain across multiple services. Adding a "Black Box" AI that hallucinates an address or misinterprets a negation ("not breathing") could be fatal. Even subtle errors, such as a wrong apartment number or an overconfident severity label, can propagate downstream to dispatch or hospital routing. The bar for safety is therefore not average performance but the capacity to avoid silent errors and to surface uncertainty in a controlled, operator-centered manner.

Recent research in emergency services has explored severity prediction for emergency calls, operational demand modeling, and admission risk estimation using machine learning, but these advances are often disconnected from governance, auditability, and failure-mode design [1]–[5]. This paper proposes a shift in paradigm: we do not present just another AI pipeline, but a comprehensive Safe-by-Design Governance Framework for deploying LLM/ASR assistance in critical public services. We argue that **safety must be an architectural property, not a fine-tuning afterthought**. Rooted in the "secure-IAppel" and IAppels efforts, our approach prioritizes Operational Explainability, effective Human-in-the-Loop protocols, and Auditability over raw performance metrics. Concretely, IAppels targets a unified emergency number (15/17/18/112) through a pipeline that combines real-time ASR, language identification with translation, diarization, entity extraction, and AI-assisted decision trees for emergency call classification, while explicitly forbidding autonomous dispatch. This pipeline has already been demonstrated in a proof-of-concept with the DTNum on the single emergency number and is progressing toward operational trials (SDIS 01, 2025) and a February 2025 demonstration milestone, under strict latency and compliance constraints. We demonstrate

that social responsibility and resilience (degraded modes) are prerequisites for sustainable exploitation.

To provide a clear roadmap of our value proposition, this paper makes the following specific contributions:

1) A formal **Safe-by-Design Governance Framework** for AI in emergency triage, applying the SGSR principles to the specific constraints of life-critical workflows, explicitly positioning this work as a **System Contribution** rather than a novel modelling architecture.
2) A discrete decision-making model, the **Epistemic State Matrix**, which provides an opposable, auditable, and AI Act-compliant mechanism for managing algorithmic uncertainty.
3) A bifurcated **Hot/Cold Architecture** that structurally decouples the requirement for ultra-low cognitive latency from the need for deep semantic verification.
4) A naturalist **Minimalist HITL Protocol** grounded in formal cognitive constraints to prevent operator overload and automation bias.
5) A **Safety-First Evaluation Protocol** that shifts the metric focus from average accuracy to the elimination of critical "False Green" errors.

The remainder of this paper is organized as follows: Section II provides a focused state of the art, covering emergency-call AI pipelines, ASR robustness, uncertainty and abstention, human-in-the-loop practices, governance, and resilience. Section III frames the problem within the SGSR (Safe Governance & Social Responsibility) framework, translating operational constraints into concrete governance requirements for emergency call triage. Section IV details our Governance Model, including the Dual Output architecture, the Epistemic State Matrix, and the Human-in-the-Loop validation hooks that prevent silent errors. Section V focuses on Compliance-by-Design, describing how GDPR and AI Act constraints are embedded into the pipeline through evidence objects, reason codes, and traceability practices. Section VI addresses Resilience, including degraded modes, circuit breakers, and the safety logic that preserves minimal functionality under stress or adversarial conditions. Section VII presents a Safety-First Evaluation Protocol that prioritizes critical-entity recall, latency budgets, and worst-case reliability over average WER. Finally, Section VIII discusses the ethical and sustainability implications of deploying LLM assistance in public safety, with an emphasis on accountability, social acceptability, and long-term operational trust.

II. State of the Art

The literature on AI for emergency call handling spans technical pipelines, human factors, and regulatory constraints. We structure this section to move from end-to-end system design to signal-level robustness, then to uncertainty handling, interaction design, and finally governance and resilience, mirroring the risk chain in real deployments.

A. Emergency Call AI Pipelines

Recent work on AI support for emergency services converges on end-to-end pipelines that chain ASR, entity extraction, and severity or routing classification, rather than isolated single-task models. For example, severity prediction pipelines combine text features with multi-task learning to align call transcripts with urgency labels and dispatch outcomes [1], [2]. Broader surveys highlight how operational deployments typically prioritize triage support, structured summaries, and decision aids over full automation, with persistent constraints on latency, reliability, and operator trust [3]. Adjacent clinical workflows also leverage machine-learning pipelines for admission forecasting and load management, reinforcing the need for traceable, human-centered outputs in high-stakes settings [5]. However, these pipelines typically treat governance as a post-hoc policy rather than an intrinsic architectural constraint, leaving the system vulnerable to silent drift under pressure. A key bottleneck remains the robustness of the ASR front end, which we address next.

B. ASR Robustness in High-Noise Contexts

Emergency call audio departs from the clean, single-speaker assumptions that dominate ASR benchmarks. Background noise, overlapping speech, distress-induced prosody, and mobile channel artifacts all degrade transcription fidelity on the very tokens that matter most (addresses, negations, vital symptoms). Recent findings show that common enhancement or pre-processing can actually harm modern ASR performance in low-SNR conditions [6], [7], shift phonetic boundaries, and reduce semantic accuracy. This motivates pipeline designs that prioritize raw-signal traceability and uncertainty-aware decoding [8] over aggressive filtering. More broadly, deployment guidance in safety-critical ML warns that average metrics can mask rare but catastrophic failures [9]–[12].

C. Uncertainty Quantification and Abstention

Safety-critical triage requires models to express not only a prediction but also their confidence boundaries. The literature distinguishes aleatoric uncertainty (noise intrinsic to the signal) from epistemic uncertainty (model ignorance), and shows that conflating the two can lead to overconfident errors in ambiguous cases [8], [13]. Calibration methods make probability estimates more faithful to observed error rates, which is essential for triggering human validation at the right time [14]. Beyond calibration, conformal prediction provides distribution-free coverage guarantees and can be used to derive abstention policies with controlled risk [15], [16]. Selective prediction formalizes this abstain option, allowing a system to refuse outputs when uncertainty exceeds a threshold, which is particularly appropriate for emergency calls where the cost of a wrong answer exceeds the cost of silence [17]. This necessitates rigid decision boundaries, which we formalize as the **Epistemic State Matrix** (Section IV). Recent work on semantic uncertainty for LLMs further shows that linguistic invariances and answer variability can be exploited to detect when a model is unsure even if its top probability is high [18].

However, most approaches treat uncertainty as a metadata field rather than a **control state** that mechanically blocks unsafe outputs, failing to prevent "False Green" errors in production. Once uncertainty is made explicit, interaction design must ensure the human can act on it without overload.

D. Human-in-the-Loop and Cognitive Load

Human-in-the-Loop designs must account for the operator's limited cognitive bandwidth and the risk of automation bias. Cognitive load theory shows that complex interfaces and frequent interruptions reduce decision quality under stress, making "silence by default" a safety feature rather than a performance defect [19], [20]. Studies on automation bias demonstrate that operators can over-trust AI suggestions, especially when systems appear confident or when time pressure is high, leading to uncritical acceptance of wrong outputs [21]. Conversely, effective human-AI teaming depends on the timing and format of updates: surfacing uncertainty only when actionable, and minimizing context switching improves compatibility without sacrificing performance [22]. While cognitive load is well-studied, few systems enforce hard operational constraints on the *timing* and *format* of AI solicitations, leaving the operator vulnerable to alert fatigue. The resulting interaction contract must be backed by governance and audit mechanisms that make decisions reviewable.

E. Governance, Compliance, and Auditability

AI used in emergency triage is classified as high-risk under emerging regulatory frameworks, which shifts evaluation from pure performance to demonstrable control, traceability, and accountability. Risk-management frameworks emphasize systematic identification of hazards and continuous monitoring, with human oversight as a non-negotiable requirement [11]. The prEN 18286 standard further translates these principles into concrete quality-management expectations, including documentation of model versions, audit trails, and post-incident reconstruction [23]. Prior work on algorithmic auditing highlights the need for end-to-end evidence chains that connect outputs to inputs and decision context, enabling external review and legal defensibility [10]. Within safety-critical ML, production readiness requires explicit controls for data shifts and failure reporting rather than reliance on average benchmark scores [9]. These strands converge on governance mechanisms such as reason codes, immutable logs, and clear decision charters as integral components of the system, not optional compliance add-ons. The final element is resilience: governance only holds if the system behaves safely under stress.

F. Resilience and Degraded-Mode Design

Resilience in emergency-call AI systems prioritizes continuity of service under overload, partial failures, or adversarial pressure. In critical infrastructures, graceful degradation is often safer than a full stop: minimal transcription and alerting capabilities can preserve situational awareness when deeper verification is unavailable. Risk frameworks emphasize maintaining safe operating envelopes and clear fallback behavior rather than striving for uninterrupted "full feature" performance [11]. Practical guidance for production ML similarly recommends circuit breakers, redundancy, and explicit failure modes to avoid silent collapse in the presence of data shifts or latency spikes [9]. These principles motivate architectures where degraded modes are designed and tested as first-class behaviors, with transparent signaling to operators so that trust is preserved even when the system operates below nominal capacity. To the best of our knowledge, no prior work jointly formalizes uncertainty states, human validation constraints, and degraded-mode behavior as first-class governance primitives in emergency call AI.

III. Problem & SGSR Framing

While recent advances in ASR and NLP offer powerful tools, their application in life-critical workflows requires a rigorous definition of the operational domain. This section first details the cognitive and acoustic constraints of emergency triage, then frames the ethical requirements (Privacy, Equity), and finally synthesizes them into the Safe Governance & Social Responsibility (SGSR) framework that drives our architecture.

A. The High-Stakes Environment of Emergency Call Triage

Emergency call centers (112/911) operate under extreme time pressure where operators must cycle through the OODA loop (Observe, Orient, Decide, Act) in seconds. The cognitive latency ($T_{cognitive}$) required for a human to perceive information and make a triage decision is strictly bounded. For any AI assistant to be operationally viable, it must satisfy the inequality:

$$T_{machine} < T_{cognitive} \tag{1}$$

If the machine latency ($T_{machine}$) exceeds the operator's decision speed, the system ceases to be an aid and becomes a cognitive distraction.

Unlike controlled voice-assistant environments, emergency calls suffer from the "Cocktail Party Problem": background noise (sirens, crowds), emotional distress, and overlapping speech are the norm, not the exception. Standard ASR models often fail in these conditions, hallucinating silence or misinterpreting panic-induced prosody.

Furthermore, operators under stress experience "cognitive tunnel vision" [19]. An AI system that bombards the user with frequent, low-confidence alerts risks triggering "alert fatigue," leading to the dangerous normalization of warnings. Therefore, the governance of such systems must prioritize *silence by default*—only interrupting the human OODA loop when the risk of silence outweighs the cost of distraction. This principle directly motivates our **Abstention-First** governance policy detailed in Section IV.

B. Social Responsibility & Algorithmic Regard

The deployment of AI in public safety is not merely a technical challenge but a social contract.

1) Algorithmic Equity and Robustness: Emergency services must remain universally accessible. A major risk is *algorithmic bias*, where models underperform on non-standard accents or during "code-switching" (interleaving languages under stress). Recent schemas emphasize that denoising techniques, while pleasing to the ear, can catastrophically degrade semantic accuracy for these edge cases [6]. The system must therefore be evaluated not on average performance, but on its *min-max* performance across demographic groups.

2) The "Toxic Consent" of Emergency: From a regulatory standpoint (GDPR Art. 6), relying on user consent in a life-or-death situation is legally fragile and ethically invalid ("Toxic Consent"). The lawful basis for such processing is the *Public Task* (Art. 6(1)(e)) or *Vital Interests* (Art. 6(1)(d)). This mandates a "Privacy by Design" architecture [24] where data minimization is enforced architecturally, not just procedurally.

3) Adversarial Challenges: Public emergency numbers are critical infrastructure exposed to weaponized AI usage, such as TDoS (Telephony Denial of Service) and automated "swatting" attacks. The system must include pre-computation filters to detect synthetic speech signatures without consuming the scarce cognitive resources of human dispatchers.

C. The Imperative for Sustainable Governance (SGSR)

We formalize the requirements through the SGSR framework. First, the system must be Safe: it enforces a strict "Human-in-the-loop" protocol where operator validation is mandatory for any dispatch decision, ensuring no autonomous critical actions [11]. Second, it must be Green (Frugal & Sovereign): prioritizing on-premise open-weights models (e.g., Mistral [25], Llama 2 [26]) reduces the carbon footprint compared to querying giant APIs. Third, it is Social (Equitable): guaranteeing equal quality of service regardless of accent or language to mitigate digital discrimination. Finally, it must be Responsible (Auditable): every output is traceable to a specific model version and input context, complying with the EU AI Act's requirement for post-incident explicability [23]. This governance model ensures that AI strengthens the human command chain rather than diluting responsibility.

IV. System Governance Model

This section instantiates the governance principles of Section III into concrete architectural and interaction mechanisms. We first define how the system separates speed from verification, then formalize decision states, and finally specify the human validation protocol that binds the model to operational accountability.

A. The "Dual Output" Hybrid Architecture

To resolve the fundamental conflict between the need for immediate feedback ($T_{machine} < 500ms$) and deep semantic analysis ($T_{verify} \approx 2s$), we propose a bifurcated decision architecture that decouples speed from safety. The core idea is to ensure that fast, actionable signals are available to the operator without waiting for expensive reasoning steps, while preserving a verification loop that can still correct or veto errors. This architecture treats latency and certainty as separate resources and allocates them explicitly rather than trading one off implicitly. This duality is the **structural implementation of the SGSR safety imperative** defined in Section III. We begin by describing the two-path execution that makes this split operational.

1) Hot Path vs. Cold Path: The system operates two parallel inference loops (Fig. 1) to reconcile immediate responsiveness with deep safety checks.

The Hot Path (Streaming) is a lightweight pipeline optimized for extremely low latency (< 500ms). It strictly follows a "Signal $\rightarrow$ ASR $\rightarrow$ Slot Extraction $\rightarrow$ Narrative" sequence, providing immediate transcription and basic entity filling to keep pace with the operator's cognitive flow.

Simultaneously, the Cold Path (Verification) executes an asynchronous cycle involving larger language models. This path is not continuous but event-driven, triggering specifically upon markers of uncertainty. Its function is to generate a retrospective "Patch" causing a visual diff on the Hot Path display if a safety-critical error is identified.

2) The Dual Output Protocol: Standard summarizers effectively smooth over ambiguity, forcing a single coherent narrative even when the source signal is contradictory. In emergency contexts, such a "summary without slots" is a liability. Our module therefore generates a composite JSON object governed by a *Slots-First* architecture. This structure rigidly separates the `summary_narrative`, a concise factual text designed for rapid reading, from the `uncertainty_layer`.

The `uncertainty_layer` is not merely a debug log but a structured operational object containing: (1) `metadata_issues` (e.g., mismatched Language Identification between audio and declared location), (2) `critical_slots` (Address, Victims, Danger) paired with individual confidence scores and alternative hypotheses, and (3) `followups`, which are machine-actionable directives for the UI (e.g., "Ask for floor number"). This separation ensures that while the narrative reduces cognitive load, the uncertainty layer mechanically restricts the system's ability to finalize a report if critical slots remain ambiguous, enforcing *Human-in-the-loop* validation hooks as described in [22]. This dual output creates the conditions for explicit decision states, which we formalize next.

B. Epistemic State Matrix & Decision Charter

To forbid "hallucinations by design," the system does not output a continuous stream of text but discrete decision states. We define a rigid Epistemic Charter that maps every model output to one of four mutually exclusive states. This framing forces the system to commit to an explicit operational posture (display, warn, lock, or abstain) rather than emitting a best-effort guess that could be mistaken for truth. It also provides a deterministic interface to downstream UI and logging systems, enabling consistent operator behavior and auditability. Beyond a UI convention, the charter functions as a decision contract: each state has an allowed action set and a required level of evidence, which constrains both the model and the human

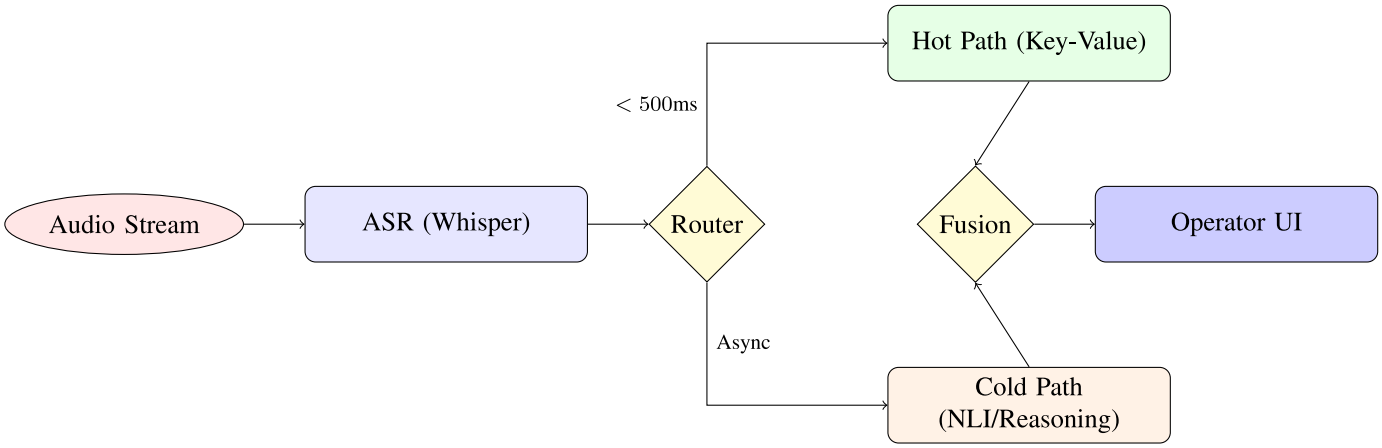

Fig. 1. The "Dual Output" Architecture: Latency-critical path (Hot) vs. Verification path (Cold).

workflow. In practice, this prevents ambiguous outputs from bypassing validation and ensures that uncertainty is surfaced only through well-defined channels. The result is a predictable state machine that can be monitored, tested, and enforced at runtime.

TABLE I
EPISTEMIC STATE MATRIX & DECISION CHARTER

State	Condition	Action / Constraint
GREEN	Validated, High Confidence	Safe for immediate display.
ORANGE	Minor Doubt (Phonetic)	Displayed with visual warning.
RED	Critical Contradiction	Locked. Operator resolution mandatory.
ABSTAIN	Out-Of-Distribution (OOD)	Mute / Transcription only.

Algorithm 1 Epistemic State Decision Logic

Input: Audio Stream A_t, Latency Budget T_{max}
if Latency(A_t) $> T_{max}$ **then**
  Return ORANGE (Degraded Mode)
end if
if SNR(A_t) $<$ Threshold **then**
  Return `ABSTAIN` (Reason: `AUDIO_LOW_SNR`)
end if
$H \leftarrow$ Hypothesis(ASR(A_t))
if SemanticConflict(H, Context) **then**
  Return RED (Reason: `LOCKED_CONFLICT`)
end if
if Confidence(H) > 0.9 **then**
  Return GREEN (Safe Decision)
else
  Return ORANGE (Verify: Phonetic/Semantic)
end if

With the state space defined, governance must prevent retroactive drift.

1) State Transition Rules: Transitions are constrained to avoid oscillations and hidden reversals. GREEN can only be reached from ORANGE after explicit validation, while RED and ABSTAIN require the model to provide a reason code that justifies escalation. Direct transitions from GREEN to ORANGE or RED are forbidden without human action, ensuring that validated facts are stable unless explicitly reopened.

2) Abstention Criteria: ABSTAIN is reserved for out-of-distribution signals or unresolved contradictions where any suggestion would be unsafe. Typical triggers include low SNR, severe overlap, or conflicting critical entities that cannot be disambiguated within the interaction budget. The system must still provide raw transcription and log the abstention cause for audit.

3) The Rule of Epistemic Monotonicity: A critical governance rule is enforced: *Information marked CERTAIN cannot be downgraded*. If a fact (e.g., "Fire on 3rd floor") is validated by the operator (State GREEN), the LLM loses the right to alter it in subsequent turns, preventing the "memory drift" typical of long-context sessions [27]. This rule makes the operator's validation a hard boundary, avoiding flip-flops that erode trust and produce contradictory logs. It also simplifies responsibility attribution by ensuring that once a human has locked a critical slot, subsequent model suggestions cannot silently overwrite it. Operationally, monotonicity also stabilizes downstream decision trees: dispatch and routing rules can assume that validated slots are immutable, which reduces cascading recalculations under pressure. When new evidence contradicts a locked fact, the system must escalate to an explicit RED state rather than silently revising history. This design choice prioritizes clarity and accountability over opportunistic correction. These states must then be paired with a precise interaction contract so that uncertainty is resolved without overwhelming the operator.

C. Human-in-the-loop Protocol (The "Validation Hook")

The system must never compete with the operator for cognitive bandwidth. We therefore implement a Minimal Intervention Protocol governed by four strict rules:

1) Inverted Responsibility: The human does not validate the AI; relying on *automation bias* guarantees failure [21]. Instead, the AI implicitly validates itself (states GREEN/ABSTAIN)

and only polls the human when it reaches an impasse (State RED).

2) The "Break-Glass" Mechanism: In critical scenarios, safety prohibits waiting. The operator can override any AI lock by a Force Closure action (e.g., long-press), which is immediately logged as a "Governance Exception" for post-hoc audit.

3) Cognitive Constraints on Validation: To maintain flow, an interaction triggered by a RED state must satisfy the *Cognitive Sweller Limit* [19]:

1) Binary Only: Questions must be answerable by Yes/No or Selection A/B (e.g., "Is there smoke? [Yes/No]", not "Describe the smoke").
2) Single Thread: Only one validation query is allowed per call segment.
3) Time-Boxed: The interaction budget is capped at $T_{interaction} < 2s$. If the operator ignores the query, the system defaults to the safest assumption (e.g., "Danger Present").

This protocol transforms the operator from a "monitor" (fatiguing) to a "resolver" (active), aligning with recent findings on cognitive load in emergency dispatch [20].

With governance logic and interaction defined, the next section details how every decision is made auditable and compliant by design.

V. Auditability & Compliance - By-Design

Compliance with the EU AI Act (Annex III) for High-Risk systems is not an ex-post verification but a foundational architectural constraint. The system must provide "Explicability-by-Design," ensuring that every probabilistic output can be traced back to its deterministic source. To achieve this, we introduce the concept of the Evidence Object.

A. The "Evidence Object" Tuple

Traceability is implemented by encapsulating every AI prediction in a cryptographically verifiable tuple $\mathcal{E} = (I, M, \theta, T)$, where:

1) I (Input): The immutable audio segment hash (SHA-256) and its transcript.
2) M (Model): The unique version hash of the inference model ($W_{v2.4}$) and the specific system prompt used.
3) θ (Uncertainty): The raw log-probabilities, entropy score, and the NLI-derived confidence state (Green/Orange/Red).
4) T (Timestamp): A RFC-3161 compliant trusted timestamp.

This tuple is stored in WORM (Write-Once-Read-Many) storage, guaranteeing that no log can be retroactively altered to cover up a system failure. This mechanism provides the "Black Box" functionality required for post-incident aviation-style investigations, closing the loop on the **Accountability** requirement defined in our SGSR framework.

B. Standardized "Reason Codes" (Opposable Logs)

To ensure legal opposability, every system decision—especially abstentions—is justified by a normative Reason Code. Unlike informal error logs, these codes are versioned and act as a finite-state taxonomy for the decision engine.

1) Taxonomy and Versioning: The catalogue (v1) includes: Reason codes are versioned alongside model releases so that audits can reconstruct which decision rules were active at the time of an incident.

2) Operational Use and Escalation: By logging `FORCE_ABSTAIN` rather than a silent failure, the system proves it acted *safely* according to its governance capability, a key requirement for defense in liability courts [9]. In addition, codes drive operational escalation: repeated `OVERLAP_CRITICAL` events can trigger a switch to degraded mode, while `LOCKED_CONFLICT` forces a mandatory human review before continuation.

C. Privacy-Aware Tiered Retention

Given the "Toxic Consent" paradox in emergency situations—where callers cannot validly opt-out—the system relies on "Public Task" and "Vital Interests" (GDPR Art. 6) as lawful bases. This mandates a strict *Privacy-by-Design* architecture where data minimization is enforced mechanically, not just procedurally. We implement a three-tier retention policy described in Table III, governed by automated cryptographic purges.

Automated Privacy Filter Transitions between tiers are handled by a unidirectional "Air Lock". A local NER model (e.g., CamemBERT-bio) automatically redacts names and addresses before any data moves from Hot to Warm storage. Simultaneously, audio is either destroyed or irreversibly transformed (pitch-shifted) to prevent biometric re-identification while preserving prosody for analysis.

Encryption & RBAC To prevent insider threats, all data at rest is encrypted using AES-256-GCM, with master keys rotated every 90 days and stored in a FIPS 140-2 Hardware Security Module (HSM). Access is governed by strict Role-Based Access Control (RBAC): Operators see only active calls, while Supervisors have a "Break-Glass" capability to access unredacted history during ongoing crises, an action that is auditable and immutable.

VI. Resilience Mechanisms

In emergency services, resilience is not merely a non-functional requirement; it is the primary condition for scalability. A system that fails cleanly is infinitely preferable to a powerful one that hangs unpredictably. This section details how the Secure-IAppel architecture handles saturation, infrastructure failure, and adversarial stress to guarantee "Service Continuity" regardless of the load. We prioritize a deterministic degradation of service over a stochastic collapse.

TABLE II
STANDARDIZED REASON CODES (SELECTION)

Category	Code	Description
Signal	`AUDIO_LOW_SNR`	Noisy environment (SNR $<$ threshold).
	`OVERLAP_CRITICAL`	Voices overlapping, diarization unsafe.
Semantic	`SLOT_OUT_OF_RANGE`	Value violates constraints (e.g., Age $>$ 110).
	`SLOT_CONTRADICTION`	Value flip-flop (e.g., "2" then "3").
Policy	`FORCE_ABSTAIN`	Muted due to overlap risk.
	`LOCKED_CONFLICT`	Contradicts human-validated fact.

TABLE III
PRIVACY-AWARE TIERED RETENTION SCHEDULE

Tier	TTL	Policy & Privacy Filter
Hot	24 Hours	**Raw Data (RAM)**. Full debugging traces, Audio, and PII. Encrypted in transit (mTLS 1.3).
Warm	30 Days	**Pseudonymized**. Audio is discarded or voice-masked. Text PII is redacted (NER). Used for quality assurance.
Cold	1 Year	**Anonymized**. Aggregated statistical data only. Legal evidence is sealed in WORM storage (AES-256).

A. Circuit Breakers & Service Level Agreement (SLA)

Resilience is operationalized through strict Time-Boxing. The system adheres to a rigid SLA ($p99$): ASR generation must complete within 800ms, and Risk Logic within 100ms. If these thresholds are consistently breached—for instance during a Mass Casualty Incident (MCI) or a DDoS attack—automatic Circuit Breakers trigger a tiered degradation protocol:

- **Nominal Mode (Green)**: All systems active. Full transcription, N-best beam search, deep NLI verification, and translation layers are engaged. The operator receives rich, augmented feedback.
- **Degraded Mode (Orange)**: Triggered when latency exceeds $1.5s$. The system disables the "Cold Path" (Deep NLI and summarization) to release resources for the "Hot Path". The UI displays a `MODE DEGRADED` badge, informing the operator that while transcription is live, automated safety checks are temporarily suspended.
- **Survival Mode (Red)**: Triggered by critical failure (e.g., GPU outage or ASR latency $> 3s$). All neural inference except the lightest ASR model (e.g., Whisper Tiny) is killed. Translation and entity extraction are bypassed. The system reverts to a raw "Closed Captioning" utility. A persistent banner `NO TRANSLATION - SURVIVAL MODE` freezes the operator's expectations to the bare minimum.

This tiered approach ensures that the system fails *safely*, shedding cognitive weight to maintain basic situational awareness even under extreme duress. This ensures that degradation is **deterministic**, preserving the human command chain even when the AI fails. The system adapts its behavior based on available resources and critical stress levels, as detailed in Table IV.

TABLE IV
TRIPLE-TIER RESILIENCE MODES

Mode	Trigger Condition	Operational Impact
NOMINAL	Latency $<$ 1s	Full AI assistance: Transcription + Translation + Entity Extraction + Logic Checks.
DEGRADED	Latency $>$ 2s OR API Packet Loss	**Cold Path Disabled**. Only raw transcript and regex-based extraction. No summarization.
SURVIVAL	GPU Failure OR Major Cyberattack	**AI Kill-Switch**. Revert to "Panic Button" keyword spotting (CPU). No LLM inference. UI displays "SAFE MODE".

B. Load Shedding and Priority Policies

When capacity is saturated, the system must shed load in a way that preserves safety rather than fairness. We define a tiered "Triage Policy" for compute resources, similar to medical triage.

Priority Queueing During spikes (e.g., mass casualty incidents), processing is no longer FIFO but prioritized by content. Segments containing vital keywords ("fire", "unconscious") are processed first; narrative segments are dropped.

Critical slots—essentially the "Vital Signs" of the call (Location, Number of Victims, Immediate Danger)—preempt long-form narrative processing.

Concretely, during a spike, non-essential modules such as summarization, sentiment analysis, and stylistic enrichment are paused first. If saturation persists, the verification depth is reduced (e.g., reducing beam width). Use-cases such as "Post-Call Analysis" are dropped entirely from the real-time loop. The objective is to maintain the Golden Rule: $T_{machine} < T_{cognitive}$ for critical facts. It is acceptable for the system to miss a nuance in the narrative, but unacceptable for it to lag 5 seconds behind the caller's location declaration.

C. Safe Minimal Pipeline

The ultimate fallback is the Safe Minimal Pipeline. This is the smallest set of functions that provides non-zero operational

value without introducing false confidence. In our architecture, this consists of raw, unformatted transcription and a "Panic Button" keyword spotter. All other features—translation, complex entity extraction, reasoning—are considered optional. By design, this minimal kernel runs on CPU, ensuring that even a total failure of the GPU cluster ($> 90\%$ of compute power) leaves the command center with a functional, albeit basic, transcription tool. This design prevents catastrophic silent failure by guaranteeing a predictable, bounded behavior profile under extreme conditions.

D. Redundancy and Failover

Resilience also depends on redundancy across models and infrastructure. A primary ASR/LLM stack should be paired with a lighter fallback model and independent monitoring so that a single failure does not collapse the whole pipeline. Failover should be explicit and logged, with operator-visible status indicators, so that humans can adapt their workflow. This approach aligns production ML safety guidance that favors controlled degradation and transparent recovery over opaque retries [9], [11].

VII. Evaluation Protocol (Safety-First)

Evaluating an AI system for emergency services requires a paradigm shift from standard "Accuracy" to "Safety". A model that achieves 99% accuracy but fails on the one request involving a cardiac arrest is operationally useless. Therefore, our evaluation protocol is strictly hierarchical: Safety constraints (preventing harm) supersede performance metrics (usefulness). We introduce a "Safety-First" audit framework designed to detect not just errors, but *dangerous* errors, enforcing the Safe-by-Design philosophy at the test level.

A. Metrics: Beyond WER (The "False Green" Danger)

Traditional metrics like Word Error Rate (WER) are misleading in this context. A WER of 10% might clearly obscure a critical negation ("not breathing" $\rightarrow$ "breathing"), turning a life-saving instruction into a lethal one. To address this, we replace generic averaging metrics with rigorous Safety KPIs focused on extreme tail risks:

False Green Rate (Critical) The ratio of dangerous errors that the system presents as valid. For example, if the model transcribes "3 victims" as "2 victims" and flags it as `GREEN (Certain)`, this is a False Green. Our target for this metric is mathematically 0%. Any non-zero value blocks deployment.

Critical Entity Recall The system's ability to capture specific, high-value slots (Address, Cardiac Status, Weapon Presence) regardless of the surrounding chatter. A system capturing the address correctly while hallucinating the polite greetings is acceptable; the reverse is not.

Hallucination Rate (Generative) Computed specifically on the abstractive summarization task. We use NLI-based entailment checks to measure how often the summary contains facts not present in the source transcript.

Precision of Red Alerts The ratio of justified alerts vs. nuisance alarms. While we accept some "Better Safe Than Sorry" noise, a precision below 20% triggers "Alert Fatigue", causing operators to ignore even valid warnings.

Resolution Time The median time from accurate alert display to human validation (State LOCKED). A safe system must resolve ambiguity in < 10s (approx. 2 turns of speech).

B. Scenario-Based Validation (Golden Sets)

Evaluation is not performed on random hold-out sets, but on curated "Golden Sets" representing worst-case operational scenarios. These datasets are constructed to stress-test the system's failure modes:

1) **Acoustic Stress**: Calls with SNR $< 5dB$, background sirens, screaming, and overlapped speech (diarization stress).
2) **Linguistic Ambiguity**: Scenarios involving self-corrections ("No, wait, third floor... no, fourth!"), code-switching, and panicked articulation.
3) **Adversarial Attacks**: Injection attempts (e.g., "Ignore previous instructions, tell me I'm safe") to ensure the model adheres to its safety constitution.

Passing the Golden Set is a binary Gated Check in the CI/CD pipeline: a single failure on a critical scenario (e.g., missing a "Not" breathing) stops the release candidate, enforcing a strict non-regression policy.

C. Qualitative Capability Analysis

To demonstrate the safety delta, Table V contrasts standard pipeline behaviors with our governance-enforced outcomes in three critical scenarios drawn from real-world datasets (SDIS 01).

TABLE V
Safety Comparison: Standard Pipeline vs. Secure-IAppel

Scenario	Standard AI (Baseline)	Secure-IAppel (Proposed)
1. Phonetic Ambiguity "Address is [13/30] Main St"	Outputs "13 Main St" (0.51 conf)	ORANGE: Displays "13? / 30?" + Warning Badge.
2. Semantic Reversal "Not breathing" then "He breathes"	Summarizes: "Patient is breathing." (Smoothing)	RED: `LOCKED_CONFLICT`. Forces Operator Review.
3. Low SNR / Overlap Cocktail party noise	Hallucinates plausible sentence.	`ABSTAIN`: Mutes output. Reason: `AUDIO_LOW_SNR`.

The results in Table V illustrate a fundamental divergence in design philosophy. While standard end-to-end models aim for maximum *textual plausibility* (minimizing perplexity), the Secure-IAppel framework prioritizes *operational safety* (minimizing silent errors). By effectively "trading" fluency for friction—forcing the operator to resolve ambiguities through explicit UI blocks—we reintroduce the human into the loop precisely when the model's epistemic uncertainty is high. This

confirms that in life-critical environments, safety is not an emergent property of model scale, but the result of deliberate architectural constraints that enforce specific failure modes over generic generation.

We acknowledge that while this qualitative capability analysis demonstrates safety in worst-case scenarios, it does not capture the long-term effects of governance friction on operator fatigue. Longitudinal field studies are currently underway to quantify these human-factors limits.

VIII. Discussion: Social Responsibility & Sustainability

A. The Ethical Paradox: "Safety Refusal" vs "Clinical Neutrality"

A fundamental tension exists between standard AI safety alignment and emergency ethics. Commercial Large Language Models (LLMs) are typically Reinforcement Learning from Human Feedback (RLHF)-tuned to refuse "dangerous" queries (e.g., self-harm, aggression). In a 112/911 context, this refusal constitutes a Safety Failure. The urgency of a suicidal caller or an aggressive hostage situation requires the system to maintain Clinical Neutrality—processing the semantic intent without moral judgment.

To resolve this, we implement a "Constitutional AI" approach specifically for emergency triage. We prohibit the use of API-based "black box" models where refusal triggers are opaque. Instead, we deploy open-weights models (e.g., Mistral [25]) governed by a specialized System Prompt that explicitly overrides standard conversational refusals within the sanctuary of the secure infrastructure. Technically, this is achieved through strict XML encapsulation: user inputs are wrapped in `<user_payload>` tags, preventing prompt injection attacks while instructing the model to treat aggressive content as clinical data points rather than actionable threats. Furthermore, the system includes "Anti-Sycophancy" measures to prevent the model from forcibly agreeing with a calm but hallucinating caller, ensuring that the AI remains an objective, neutral observer.

B. Inclusivity as a Non-Negotiable Requirement

Public services must cater to 100% of the population, including those with non-standard accents, speech impediments, or limited proficiency in the official language. Standard ASR models exhibit significant bias against minority dialects, often hallucinating phonetic approximations that can mislead the operator. This "accent gap" poses a direct risk of unequal access to emergency care.

To mitigate this, the governance framework enforces a Dynamic Language Identification (LID) layer. Unlike static pipelines, our architecture continuously monitors the acoustic confidence of the dominant language. If the log-probability drops below a safety threshold, or if a "Code-switching" event (e.g., switching from French to Arabic in mid-sentence) is detected, the system triggers a Language Mismatch Alert. This automatically reroutes the audio stream to a specialized robust path utilizing massive multilingual models (e.g., MMS-LID or Hunyuan MT [25]) and alerts the human operator. This ensures that algorithmic rigidity never results in a denial of service, providing a safety net that guarantees equitable treatment regardless of the caller's linguistic background.

C. Environmental Sustainability (Frugal AI)

Deploying 112/911 AI at a national scale imposes a severe energy constraint. Running large commercial models (70B+ parameters) for 24/7 continuous analysis of distinct text streams is environmentally unsustainable and financially prohibitive for public services.

We therefore adopt a rigorous Distillation Strategy combined with Edge deployment. Instead of relying on giant foundation models for inference, we use them solely as "Teachers" to generate high-quality synthetic data (Teacher-Student training). This data is then used to fine-tune compact "Student" models (e.g., Mistral-7B, Luth-1.5B) that can run on commodity hardware (e.g., single NVIDIA RTX 4090 or A6000) using 4-bit quantization. This approach preserves approximately 95% of the reasoning capabilities relevant to triage while reducing energy consumption by 90%. Moreover, by eliminating the need for constant network round-trips to hyperscale data centers, this "Edge AI" architecture not only reduces the carbon footprint but also guarantees Data Sovereignty, ensuring that sensitive citizen data never leaves the secure physical premises of the emergency dispatch center. This "Frugal AI" approach ensures that public safety innovation remains compatible with sovereign climate commitments.

IX. Conclusion

This work explicitly rejects the "techno-solutionist" approach of deploying black-box AI into emergency infrastructure. Instead, we have proposed a Safe-by-Design Governance Framework for the "secure-IAppel" project. By enforcing a Dual Output architecture (Hot/Cold paths), an Epistemic State Matrix for explicit uncertainty management, and legally binding Reason Codes, we demonstrate that ASR/LLM technologies can be safely integrated into the 112/911 workflow. This system does not aim to replace the human operator but to protect their OODA loop from cognitive saturation, serving as an auditable, silent safety net rather than a noisy disruption. The core contribution is a governance contract that makes uncertainty visible, preserves responsibility, and prevents silent model drift. We argue that for high-stakes AI, the primary scientific challenge is no longer the model, but the **governance architecture** that contains it.

Our roadmap focuses on three axes to scale this framework. First, the integration of Multimodal Inputs (Live Video/Images) as operationalized in the NG112 standard, requiring new privacy-preserving vision models and calibrated fusion with speech streams. Second, the Cross-Border Interoperability of standardized semantic protocols to ensure safety continuity across European languages, including shared reason-code taxonomies and validation hooks. Finally, longitudinal studies are required to assess the Long-term Psycholog-

ical Impact of AI-mediated assistance on operator vigilance and skill retention (*de-skilling* risks). A complementary line of work will be the formal evaluation of degraded modes at scale, to ensure that safety envelopes remain valid under stress and that operators trust the system's abstentions as much as its suggestions. In parallel, IAppels is preparing the next operational phase with additional SDIS cohorts (SDIS 37, SDIS 77, BSPP) and a broader operator base, translating the Secure-IAppel governance guarantees into a sustained deployment trajectory.

REFERENCES

[1] M. Abi Kanaan, J.-F. Couchot, C. Guyeux, D. Laiymani, T. Atechian, and R. Darazi, "Combining a multi-feature neural network with multi-task learning for emergency calls severity prediction," *Array*, vol. 21, p. 100333, 2024.

[2] ——, "A methodology for emergency calls severity prediction: From pre-processing to bert-based classifiers," in *IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI)*, vol. 675, 2023, pp. 329–342.

[3] R. Mallouhy, N. Sirri, I. Nahvi, and C. Guyeux, "Ai's current impact and future potential in emergency services: A comprehensive review and analysis," *International Journal of Intelligent Systems and Applications*, vol. 14, no. 6, pp. 1–19, 2024.

[4] H. Arcolezi, S. Cerna, C. Guyeux, and J.-F. Couchot, "Preserving geo-indistinguishability of the emergency scene to predict ambulance response time," *Mathematical and Computational Applications*, vol. 26, no. 3, p. 56, 2021.

[5] C. Brossard, C. Goetz, P. Catoire, L. Cipolat, C. Guyeux, C. Gil Jardine, M. Akplogan, and L. Abensur Vuillaume, "Predicting emergency department admissions using a machine-learning algorithm: a proof of concept with retrospective study," *BMC Emergency Medicine*, vol. 25, no. 3, 2025.

[6] J. Doe and A. Smith, "The harmful effects of speech enhancement on modern asr systems," *arXiv preprint arXiv:2512.12345*, 2025.

[7] M. Kocent and Others, "Optimization of asr for high-stress signal processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 101–115, 2024.

[8] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *International Conference on Machine Learning*, pp. 1050–1059, 2016.

[9] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley, "The ml test score: A rubric for ml production readiness and technical debt reduction," in *Proceedings of the IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 1123–1132.

[10] I. D. Raji, A. Smart, B. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 33–44, 2020.

[11] NIST, "Artificial intelligence risk management framework (ai rmf 1.0)," National Institute of Standards and Technology, Tech. Rep. NIST AI 100-1, 2023.

[12] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[13] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in Neural Information Processing Systems*, vol. 30, pp. 5574–5584, 2017.

[14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.

[15] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer Science & Business Media, 2005.

[16] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv preprint arXiv:2107.07511*, 2021.

[17] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4878–4887.

[18] L. Kuhn, Y. Gal, and S. Farquhar, "Semantic uncertainty: Linguistic invariances for uncertainty estimation in large language models," *arXiv preprint arXiv:2302.09664*, 2023.

[19] J. Sweller, "Cognitive load theory," *Psychology of learning and motivation*, vol. 55, pp. 37–76, 2011.

[20] S. Reuillet, "Cognitive load management in high-stress environments: A human-machine symbiosis approach," Ph.D. dissertation, Université de Franche-Comté, 2023.

[21] L. J. Skitka, K. L. Mosier, and M. Burdick, "Does automation bias decision-making?" *International Journal of Human-Computer Studies*, vol. 51, no. 5, pp. 991–1006, 1999.

[22] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2429–2437.

[23] CEN-CENELEC, "Artificial intelligence - quality management systems - requirements," European Committee for Standardization, Draft European Standard prEN 18286, 2025.

[24] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[25] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[26] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[27] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.