Auditable Clinical Translation Orchestration

A verifier-backed approach to real-time medical language bridging.

AUTHOR Freddy Rojas · Orquor

VERSION v0.1 draft · May 2026 · subject to revision

CONTACT research@orquor.com · orquor.com

ABSTRACT

Why ACTO is necessary, and why now.

*|1**|1**|1*

research@orquor.com

1. Introduction

Clinical communication across language boundaries is one of the highest-stakes natural-language tasks in modern healthcare. A mistranslated drug dose, a misunderstood symptom, or a miscommunicated consent can produce direct patient harm and significant legal liability. The Joint Commission has documented language barriers as a contributing factor in adverse events for over two decades, and the steady internationalization of healthcare — medical tourism, cross-border telehealth, immigration-driven demographic shifts in the United States, refugee health programs in Latin America — has continued to grow demand for accurate, instantaneous, and defensible medical interpretation.

The existing market response splits into two categories.

Video Remote Interpretation (VRI) and Over-the-Phone Interpretation (OPI) services like LanguageLine Solutions, Cloudbreak Health, Stratus Video, and Voyce Global provide human interpreters on demand. They offer high accuracy at high cost — typically USD 1.50 to 2.50 per minute — with latency measured in tens of seconds for connection and onboarding. The audit trail consists of session metadata; the actual translation is ephemeral.

AI-only translation systems like Wordly and direct deployments of generic large language models (GPT, Claude, DeepL) offer near-zero latency at low marginal cost but cannot meet the burden of proof required in regulated healthcare. They produce text but no verifiable evidence that the text faithfully represents what the clinician said. When errors occur — and they do, at rates that vary from a few percent on common terms to double-digit percentages on specialty vocabulary — there is no remediation path and no forensic record.

Neither category solves the actual problem facing healthcare administrators in 2026: producing medical interpretation that is fast enough for real clinical use, cheap enough for emerging-market budgets, and *defensible enough* for adverse-event investigation, regulatory audit, and litigation discovery.

We propose **Auditable Clinical Translation Orchestration (ACTO)** as the third category. An ACTO system orchestrates human interpreters and AI components into a verified pipeline: every translation output passes through a deterministic test suite before reaching the clinician, every session is logged with cryptographic timestamping, every consequential decision (specialty register selection, fallback to human interpreter, escalation) is auditable. ACTO is not a replacement for VRI or for AI translation — it is the orchestration layer that makes either component usable in a healthcare-grade context.

The remainder of this paper formalizes the ACTO concept, describes a reference architecture, articulates the verifier-design methodology adapted from current agent-evaluation research, and proposes a benchmark methodology that future systems can be measured against.

2. The Categorical Gap

We frame the gap quantitatively along seven axes that matter to a hospital procurement decision.

AXIS	VRI / OPI (HUMAN-LED)	AI-ONLY TRANSLATION	ACTO (PROPOSED)
Latency to first token	15–60 seconds	0.5–2 seconds	0.5–2 seconds
Cost per minute (USD)	1.50–2.50	0.02–0.10	0.20–0.80
Word Error Rate, medical	1–3%	8–15%	Target ≤ 5%
Translation faithfulness, clinical (COMET-22)	High variance, dependent on interpreter	0.78–0.84	Target ≥ 0.86
Audit log granularity	Session-level metadata	None or token log only	Per-utterance with cryptographic timestamping
Regulatory defensibility (HIPAA, Law 29733)	Strong for human, weak for AI	Weak	Mathematically demonstrable
Specialty register adaptation	Manual, interpreter-dependent	Generic	Department-selected, register-tagged

The ACTO category is defined by simultaneous performance on all seven axes, not by superiority on any single axis. A system that achieves low latency and high audit defensibility but mediocre translation quality is not ACTO; it is an audit log for mediocre translation. A system that achieves high translation quality and low latency but no audit defensibility is not ACTO; it is fast translation with legal exposure.

3. Reference Architecture

We describe the reference ACTO architecture in five components. Implementations may vary in specifics, but a system that omits any component fails to meet the ACTO definition.

3.1 The orchestration substrate

The orchestrator is the central process that coordinates the other components. It accepts an audio stream and metadata (department, patient identifier, clinician identifier, session identifier, consent flags, register selection) and produces a translated text stream, an audit record, and a verification result.

The orchestrator must support deterministic replay: given the same audio input and the same metadata, the system must produce a bit-identical audit record. This requirement rules out

implementations that depend on non-deterministic LLM inference without temperature pinning and seed control.

3.2 The ASR component (Automated Speech Recognition)

The ASR component transcribes the source-language audio. ACTO requires that the ASR be fine-tuned on medical vocabulary in the source language. A baseline Whisper-Large-v3 model achieves approximately 11–14% Word Error Rate on medical terminology in English; an ACTO-compliant ASR achieves ≤ 5%, validated against a fixed evaluation set described in Section 5.

The ASR component must emit confidence scores per token. The orchestrator uses these scores as a primary input to the verifier and to the human-escalation decision.

3.3 The MT component (Machine Translation)

The MT component translates source-language transcripts into target-language clinical text. ACTO requires that the MT be specialized for the medical domain and capable of register adaptation: a Formal/Clinical register for physician-physician communication, a Simplified register for patient-facing communication, a Legal register for documentation that may enter the medical record or be cited in adverse-event review.

We do not prescribe a specific MT architecture. The reference implementation uses a fine-tuned sequence-to-sequence transformer initialized from NLLB-200 and adapted with paired corpora from clinical settings. Other implementations could use prompted LLMs with verifier feedback, ensembles, or interactive translation with the human interpreter as an editor.

3.4 The verifier suite

This is the component that distinguishes ACTO from prior categories. Before any translated output reaches the screen, it passes through a deterministic test suite — a verifier — that evaluates the output against outcome-based criteria.

A verifier is a Python function that returns `pass` or `fail` (with structured reasons) for a given input-output pair. Verifiers are designed using the methodology described in Section 4. Examples of verifier checks include:

- The source utterance mentions a numeric dosage; the translated utterance contains a numeric dosage with the same numeric value
- The source utterance mentions a known controlled-substance name; the translated utterance contains the corresponding controlled-substance term in the target language
- The source utterance does not mention a person by name from a predefined exclusion list (privacy redaction); the translated utterance does not contain those names
- The translated utterance's clinical-register classification matches the requested register
- The translated utterance's measured COMET score against a sentence-level reference, if one is available, exceeds 0.85

A translation that fails verification is not silently delivered. Depending on the failure type, the orchestrator either re-runs the MT with the failure signal as input, escalates the utterance to a human interpreter, or marks the session as requiring post-session review with the failure recorded in the audit log.

3.5 The audit primitive

Every utterance in an ACTO session is logged with the following minimum fields:

- Session ID (universally unique, immutable for session duration)
- Utterance ID (sequential within session)
- Source audio reference (URI of the encrypted blob; the blob itself is not in the log)
- ASR transcript with per-token confidence
- MT output with selected register
- Verifier result and reasons (structured)
- Timestamp from an external cryptographic timestamping authority (RFC 3161, OpenTimestamps via Bitcoin or Ethereum)
- Digital signature of the log entry by the orchestrator's hardware key

The audit log is append-only. Modification is detected by signature verification. The log is portable: a hospital can demand and receive its complete audit log on demand, in a format that an external auditor can independently verify against the cryptographic timestamping authority without trusting the operator.

4. Verifier Design Methodology

The verifier suite is the technical heart of ACTO. Designing verifiers is a discipline adapted from recent work in agent evaluation, in particular the outcome-based rubrics methodology that has emerged in projects such as OpenClaw Atlas and the HiL-Bench Blocker Injection framework. We summarize the relevant principles.

Principle 1: Verifiers evaluate outcomes, not traces. A verifier checks that the final translated output has a property, not that the system's intermediate steps followed a particular pattern. This makes verifiers robust to implementation changes: as MT models improve, the verifiers continue to apply unchanged.

Principle 2: Verifiers are atomic. Each verifier evaluates exactly one property. A verifier that combines multiple checks via conjunction or disjunction obscures failure attribution. The standard implementation form is a `pytest`-style test function that returns boolean and a structured reason on failure.

Principle 3: Verifiers are weighted explicitly. Different failure modes have different severity. A misidentified medication name is catastrophic; a stylistic deviation in register is mild. Each verifier carries a weight (e.g. ±5 for catastrophic, ±3 for moderate, ±1 for stylistic) and the orchestrator's escalation logic is parameterized by the weighted sum.

Principle 4: Negative verifiers are required. A complete verifier suite includes both positive verifiers ("the translated output contains the dosage") and negative verifiers ("the translated output does not contain names from the exclusion list"). Without negative verifiers, the suite is incomplete because it cannot detect leakage of disallowed content.

Principle 5: Verifiers are versioned. When a verifier changes, the version is recorded in the audit log. A translation verified under verifier-suite v1.2 is auditable as such even after the suite is updated to v1.3.

The current reference implementation of Orquor Clinical maintains a verifier suite of approximately 60 verifiers organized by clinical specialty (Emergency, Operating Room, ICU, Pharmacy, Pediatrics, Psychiatry, Admissions, Legal, 911 Dispatch, General) and by register (Formal/Clinical, Simplified, Legal). The full suite is available on request under a research collaboration license and will be published as an open-source standard within twelve months of this paper.

5. Benchmark Methodology

We describe a benchmark methodology for ACTO systems. The intent is to make ACTO systems directly comparable and to enable hospital procurement teams to evaluate vendors on a uniform basis.

5.1 Evaluation set

The evaluation set consists of 1,000 utterances per specialty, sampled from de-identified clinical conversations or synthesized from medical reference materials. Each utterance is paired with a gold-standard transcription, a gold-standard translation in the target language(s), and a clinical-register annotation.

We commit to publishing the first version of the evaluation set as a public benchmark within six months of this paper, with appropriate de-identification, consent, and IRB review.

5.2 Metrics

For each specialty and register, we compute:

- **WER** (Word Error Rate) of the ASR component against the gold transcript
- **COMET-22** translation quality of the MT component against the gold translation, measured at the sentence level
- **Verifier pass rate** of the full pipeline output
- **Latency** to first translated token, p50/p95/p99
- **Latency** to verified output (including any verifier-driven retries), p50/p95/p99
- **Human escalation rate** as percent of utterances

5.3 Reporting

We propose that ACTO systems publish their benchmark results in a standardized format that includes hardware, software versions, model versions, verifier suite version, and date of measurement. Reproducibility is required: the system vendor must be able to re-run the benchmark and produce results within ± 0.5 percentage points of the reported values.

5.4 Current results

For the Orquor Clinical reference implementation, we report preliminary results on a 500-utterance evaluation set spanning Emergency, Operating Room, and Pharmacy specialties. These are preliminary; the full peer-reviewed benchmark will follow in a subsequent paper.

SPECIALTY	WER	COMET-22	VERIFIER PASS RATE	LATENCY P50	HUMAN ESCALATION
Emergency	4.1%	0.87	94.2%	0.9s	5.8%
Operating Room	3.6%	0.88	95.6%	0.8s	4.4%
Pharmacy	3.2%	0.89	96.1%	0.7s	3.9%

These numbers are preliminary, drawn from a single-site pilot, and should not be interpreted as production guarantees. They are presented here as evidence that the ACTO architecture is achievable in practice, not as a competitive claim. The intent of this paper is categorical, not commercial.

6. Regulatory Posture

ACTO is designed to meet the documentation requirements of four major regulatory frameworks that govern healthcare data and clinical decision support in 2026.

HIPAA (United States). The audit log primitive satisfies the access logging and audit control requirements of 45 CFR § 164.312(b). The cryptographic timestamping satisfies the integrity controls of § 164.312(c). Business Associate Agreements (BAA) with hospital customers cover the ACTO operator's role as a downstream processor.

Peruvian Law 29733 (Protección de Datos Personales). ACTO supports the data subject rights under Articles 18–26, including access, rectification, and the right to oppose processing. The audit log facilitates compliance with the reporting requirements of the National Authority for Personal Data Protection.

Brazilian LGPD. The legal basis for processing patient data is the legitimate interest of healthcare delivery under Art. 7, IX, combined with the explicit consent recorded at session initiation. The audit log supports the data subject's right to portability under Art. 18, V.

EU GDPR. The architecture supports data minimization (Art. 5(1)(c)), storage limitation (Art. 5(1)(e)), and integrity (Art. 5(1)(f)). The cryptographic audit log materially exceeds the GDPR

baseline for healthcare processing.

We note that no software system is automatically compliant with any of these frameworks. ACTO provides the *technical primitives* on which a compliance program can be built; the compliance program itself remains the responsibility of the operating organization.

7. Related Work

The closest prior art falls into three groups.

Medical machine translation research. A substantial body of work studies neural MT performance on clinical text, including Costa-jussà et al. on the Health Mention task, the MEDLINE bilingual corpora, and recent benchmarks on prescription label translation. This work focuses on translation quality in isolation, not on orchestration or auditability.

Agent evaluation frameworks. OpenClaw Atlas, HiL-Bench, and the broader literature on outcome-based agent rubrics provide the methodological foundation for the verifier design described in Section 4. ACTO applies these methods to a specific high-stakes domain.

Healthcare interpretation services. Commercial offerings from LanguageLine Solutions, Cloudbreak Health (Martti, InDemand), Stratus Video, Voyce, and Boostlingo are operational systems but do not articulate a public technical category. They do not publish benchmarks. They do not provide a cryptographic audit primitive. ACTO is proposed as the category they implicitly should have been operating in but have not chosen to formalize.

To our knowledge, no prior published work proposes the synthesis of verifier-backed translation with cryptographic audit logging as a named technical category applicable to healthcare interpretation.

8. Conclusion and Call

We have proposed **Auditable Clinical Translation Orchestration** as the technical category that resolves the quality-cost-defensibility trade-off currently bifurcating the medical interpretation market. ACTO is defined by simultaneous performance on translation quality, latency, cost, and regulatory-grade auditability. It is realized through an orchestration substrate, a fine-tuned ASR, a register-adaptive MT, a verifier suite of weighted outcome-based tests, and a cryptographically timestamped audit log.

We invite three communities to engage with this proposal.

To hospital procurement and CIO offices: include ACTO-aligned criteria in your next translation services RFP. Specifically: require per-utterance audit logs with cryptographic timestamping, require verifier pass rates by specialty, require benchmark publication.

To clinical informatics researchers: contribute to the public evaluation set we will release in 2026. Independent benchmarks reduce vendor lock-in across the industry.

To other system builders: adopt the ACTO category. We are not pursuing this as a trademark; the *category* is open. The implementations will compete. The category benefits everyone.

The next decade of healthcare AI will be defined by the systems that bridged languages safely under regulatory scrutiny. ACTO is the framework under which those systems should be designed.

Appendix A — Verifier Suite Sample

We include three representative verifiers from the production suite as illustration. The full suite of approximately 60 verifiers is documented separately.

```
def verify_dosage_preserved(source: str, translated: str) → VerifierResult:
    """Numeric dosage in source must appear identically in translated.
    Weight: +5 (catastrophic if violated)."""
    source_dosages = extract_numeric_dosages(source)
    translated_dosages = extract_numeric_dosages(translated)
    missing = source_dosages - translated_dosages
    if missing:
        return VerifierResult(
            passed=False,
            weight=5,
            reason=f"Dosage value(s) {missing} present in source but not in transla
        )
    return VerifierResult(passed=True, weight=5)


def verify_no_excluded_names(translated: str, exclusion_list: list[str]) → Verifie
    """Translated output must not contain names from exclusion list.
    Weight: -5 (privacy violation if triggered)."""
    found = [name for name in exclusion_list if name.lower() in translated.lower()]
    if found:
        return VerifierResult(
            passed=False,
            weight=-5,
            reason=f"Excluded name(s) {found} appeared in translation",
        )
    return VerifierResult(passed=True, weight=-5)


def verify_register_match(translated: str, requested_register: str) → VerifierResu
    """Clinical register of translated must match requested register.
    Weight: +2 (stylistic)."""
    detected = classify_register(translated)
    if detected ≠ requested_register:
        return VerifierResult(
            passed=False,
            weight=2,
            reason=f"Requested register={requested_register}, detected={detected}",
        )
    return VerifierResult(passed=True, weight=2)
```

Appendix B — Glossary

- **ACTO** — Auditable Clinical Translation Orchestration. The category proposed in this paper.
- **ASR** — Automated Speech Recognition.
- **MT** — Machine Translation.

- **VRI** — Video Remote Interpretation.
- **OPI** — Over-the-Phone Interpretation.
- **WER** — Word Error Rate.
- **COMET-22** — Crosslingual Optimized Metric for Evaluation of Translation, 2022 reference implementation.
- **Verifier** — Deterministic test function that evaluates an outcome property of system output.
- **Cryptographic timestamping** — Method (RFC 3161 or OpenTimestamps) for producing tamper-evident proof that a document existed at a specific moment.
- **HIPAA BAA** — Business Associate Agreement under the US Health Insurance Portability and Accountability Act.
- **Law 29733** — Peruvian law on Protection of Personal Data.
- **LGPD** — Brazilian Lei Geral de Proteção de Dados.

Acknowledgments

The author thanks the broader agent-evaluation research community, in particular contributors to OpenClaw Atlas, the HiL-Bench framework, and the labelers and researchers at platforms including Outlier, Scale AI, and Labelbox who have advanced outcome-based rubric design as a practical discipline. The verifier-design methodology described in Section 4 is directly indebted to this body of work.

This research was conducted at Orquor. The author acknowledges the clinical partners who participated in the single-site pilot that generated the preliminary benchmark results reported in Section 5.4; their institutions requested anonymity pending institutional review.

The author used the following open-source tools in the preparation of this work: Whisper (OpenAI) for ASR baselines, NLLB-200 (Meta AI) for MT initialization, COMET (Unbabel) for translation quality evaluation, and OpenTimestamps for cryptographic timestamping reference implementation.

The remaining mistakes are the author's.

Contact

For research collaboration, benchmark access, or hospital pilot inquiries: research@orquor.com