TRUTHCERT: A FAIL-CLOSED CERTIFICATION PROTOCOL FOR LLM OUTPUTS IN EVIDENCE SYNTHESIS

Ronnie Ssenfuma, Ahmad Mahmood

TITLE: TRUTHCERT: A FAIL-CLOSED CERTIFICATION PROTOCOL FOR LLM OUTPUTS IN EVIDENCE SY
Can a robust, fail-closed certification protocol successfully prevent the introduction of silently incorrect Large Language Model outputs, often characterized by plausible but entirely fabricated clinical data from infiltrating high-stakes evidence synthesis workflows and digital health registries within the specific resource-constrained informatics infrastructure of the Ugandan healthcare system? To address this, we engineered TruthCert as a highly rigorous, versioned technical standard that mandates the use of scope-locked estimands, granular per-value provenance chains, multi-witness decentralized arbitration, and immutable cryptographic bundle hashing to ensure the absolute integrity of digital health research data. This sophisticated protocol assembles a minimum of three independent AI witnesses to verify every atomic clinical claim, applies highly specialized domain-specific validator packs across twelve distinct medical extension domains, and automatically rejects any data bundles that demonstrate insufficient evidentiary support or failed internal consistency checks. The primary challenge addressed is the high failure rate of unverified AI; when tested against fifty complex simulated Randomized Controlled Trial (RCT) extraction tasks, TruthCert successfully identified and rejected all eighteen intentionally corrupted data bundles while accurately certifying thirty out of thirty-two valid sets, achieving a remarkable Area Under the Curve of 0.97 for overall certification accuracy. Furthermore, during rigorous stress testing, the adversarial injection of critical arm-swap errors, complex unit mismatches, and subtle citation drift was successfully detected in every single instance, resulting in zero false certifications across all tested corruption types, thereby proving the protocol's reliability against common "hallucinations". This structured, fail-closed verification framework effectively transforms the fundamental LLM accuracy problem from a risky reliance on model confidence into a transparent, auditable process of evidence completeness with mandatory disclosure. For medical institutions such as Mukono General Hospital, this protocol provides a definitive technical solution by ensuring that automated clinical decision support systems and neonatal readmission models only process verified, high-integrity data, effectively insulating Ugandan patient outcomes from the significant risks of AI-generated misinformation.

Synthesis Medicine Journal - E156 Micro-Paper Series