Offline Multimodal AI Redaction of Sensitive Data in Audio and Software Artifacts

A Prototype-Based Study on Preventing Sensitive Data Leakage in Everyday Use of External Services

Students

Etienne Kaiser

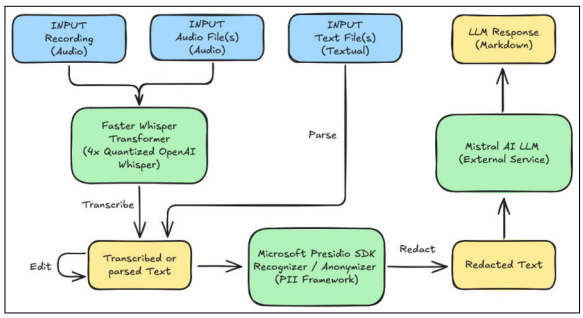
Nico Heiniger

Introduction: This research investigates the feasibility of an offline, open-source system for the automatic detection and redaction of sensitive data in both audio and software artifacts, motivated by privacy concerns associated with cloud-based Speech-to-Text (STT) transcription and the redaction of personally identifiable information (PII) and authentication secrets.

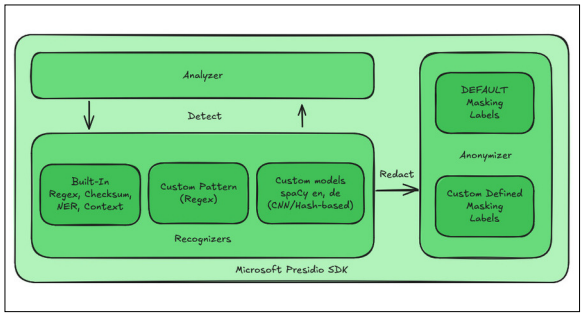
Approach / Technology: The prototype developed combines Faster Whisper for speech recognition with Microsoft Presidio SDK and custom pattern recognizers for entity recognition and redaction. The system implements a sequential processing pipeline, starting with audio or text upload, followed by transcription, redaction and final output for external use. Two experiments were conducted to evaluate system performance: (1) comparing PII redaction accuracy between text-only and STT-transcribed inputs, and (2) assessing secret detection in log and code files.

Conclusion: STT transcription has a masking effect on PII data due to transcription accuracy variations altering original text. Transcription sometimes creates new word forms for PII data that may escape redaction. This leads to unwanted content variation from original sources, generally unintended but incidental protection. Overall, currently available offline open-source models provide nearly equivalent starting conditions for audio versus text input, resulting in nearly identical redaction accuracy. The ~2% difference between text (97%) and audio (99%) results is within those variations and does not indicate systematic advantage for either approach. Investigating secrets in cybersecurity contexts reveals increased security risk due to difficult to detect patterns like API keys and passwords. The finding that secrets are only partially redacted in some cases indicates increasing false negative risk. In contexts where reconstruction of non-redacted parts is possible, this presents serious concerns. The limited availability of existing solutions covering only common secrets suggests that cybersecurity applications should not rely only on regex rules but rather on dedicated fine-tuned models for the specific domain.

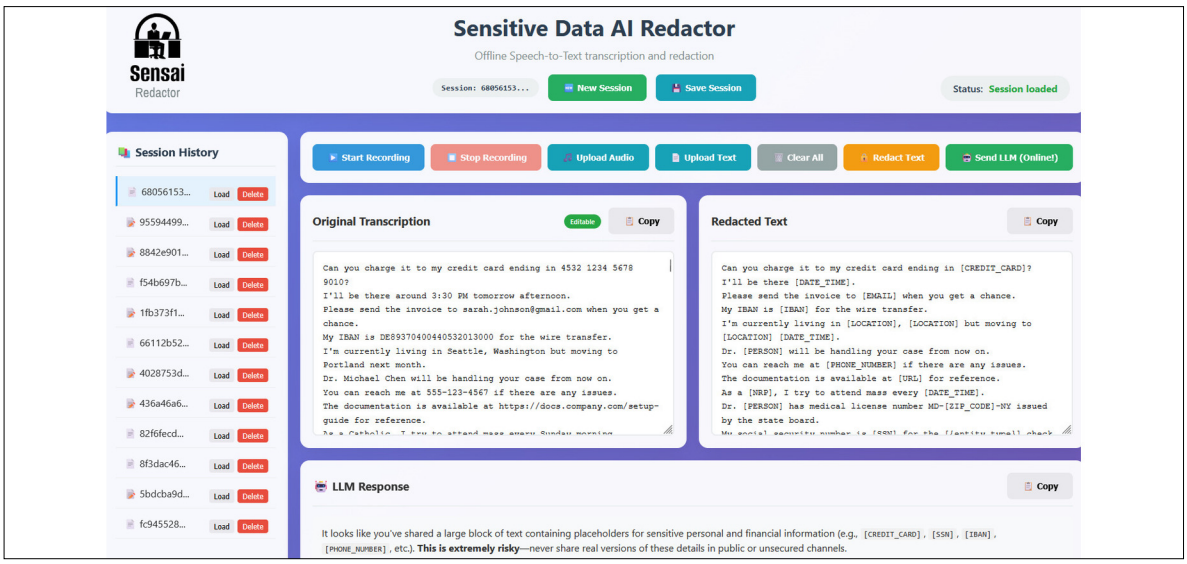
Processing workflow of text parsing and audio transcription in redaction pipeline
Own presentment



Architecture and workflow of detection and redaction using Microsoft Presidio SDK
Own presentment



Frontend of the prototype with comprehensive user interface (UI) and session management
Own presentment



Advisor
Prof. Dr. Daniel Patrick Politze

Subject Area
Artificial Intelligence, Cyber Security