

PsyberTrace

An Offline, Reviewer-Correctable Forensic Triage Assistant for Digital Chat Evidence

Submission Category	Digital Forensics, Cybercrime Investigation & Forensic Behavioral Science
Project Title	PsyberTrace: Offline, Reviewer-Correctable Triage Assistant for Digital Chat Evidence
Academic Institution	Curtin University Dubai (Dubai, United Arab Emirates)
Academic Supervisor	Dr. Anitha Menon — Faculty Supervisor, Curtin University Dubai
Project Team	Diluksh Fernando — Cybersecurity Lead (Cryptographic Integrity, Parsers & Metadata Forensics) Harishma — Psychology Lead (Forensic Psycholinguistics & Discourse Taxonomy)
Forum Dates	16–18 November 2026 (World Police Summit, Dubai, UAE)
Hardware / Deployment & Live Concept Demo	Target Production Environment: 100% Offline, Air-Gapped Local Execution (Standard Forensic Workstation) Interactive Concept Prototype (Live Preview): https://psybertrace.pages.dev

EXECUTIVE ONE-LINE PROPOSAL

PsyberTrace is an offline, reviewer-correctable forensic triage assistant that preserves chat-evidence integrity in accordance with ISO/IEC 27037:2012 guidelines, organizes multi-platform messages and metadata, transparently flags candidate conversation segments using clinically grounded linguistic frameworks and cybersecurity anomalies for specialist review, and generates source-linked, court-auditable draft reports.

1. EXECUTIVE SUMMARY & OPERATIONAL CONTEXT

The Forensic Challenge: In contemporary digital investigations involving online child exploitation, sextortion, cyber-harassment, and domestic coercive control (addressed under UAE Federal Decree-Law No. 34 of 2021 on Countering Rumors and Cybercrimes), forensic examiners are frequently confronted with massive mobile chat extractions ranging from 50,000 to over 200,000 messages across platforms such as WhatsApp, Telegram, Discord, and Instagram. Manual linear review creates severe investigative bottlenecks, while static keyword filters suffer from high false negatives (due to dialectal slang, indirect manipulation, voice notes, and English–Arabic–*Arabizi* code-switching) and strip messages of vital conversational context.

Alignment with Dubai Police Priorities: Public information indicates that Dubai Police utilizes digital forensics, analytics, and AI-assisted investigation workflows. **PsyberTrace** proposes a transparent, reviewer-correctable triage layer for authorized chat exports, designed to complement—not replace—existing forensic extraction suites (such as Cellebrite UFED or Magnet AXIOM) and specialist human review.

2. INTERDISCIPLINARY ARCHITECTURE: THE PSYBERTRACE ENGINE

PsyberTrace is built around a 50/50 interdisciplinary partnership between **Cybersecurity / Digital Forensics** and **Forensic Psychology**. It is explicitly a triage and documentation assistant—not an autonomous decision-maker of criminal guilt or clinical diagnosis.

DOMAIN 1: CYBERSECURITY & INTEGRITY ENGINE (DILUKSH FERNANDO) • Read-Only Ingestion & SHA-256 Hashing: Locks raw exports (.txt , .json , .csv) in read-only mode and computes SHA-256 integrity hashes before and after parsing in alignment with ISO/IEC 27037:2012 digital evidence preservation principles. • Malformed Export Detection: Flags truncated or corrupted exports immediately rather than silently analyzing incomplete evidence. • UTC+4 Schema Normalization: Indexes message IDs, sender attribution, normalized UAE local timestamps (UTC+4), and original byte/line offsets. • Metadata Anomaly Detection: Identifies late-night bursts (00:00–05:00), rapid response-latency shifts (Δt), deleted-message spoliation clusters, and high-risk entities (cryptocurrency wallets, UAE phone numbers/IBANs, off-platform links). • Offline Voice-Note Pre-Triage: Local faster-whisper speech-to-text transcription of WhatsApp .opus audio across English and Arabic.	DOMAIN 2: FORENSIC PSYCHOLOGY ENGINE (HARISHMA) • Literature Grounding: Built upon O'Connell (2003), Lorenzo-Dus et al. (2016), Inches & Crestani (2012), and Stark's (2007) Coercive Control framework. • C1 — Rapport & Exclusivity Framing: Special-status validation, age-maturity flattery, and gift promises. • C2 — Secrecy & Boundary Testing: Directives to hide chats from parents, delete messages, or incremental boundary probing. • C3 — Structural Isolation Indicators: Statements driving wedges between the recipient and family/peers or off-platform migration. • C4 — Coercive Escalation & Extortion: Reputational threats (<i>fadiha</i>), countdown ultimatums, and blackmail framing. • C5 — Acute Distress Language: Surfaces severe distress markers for immediate human review (non-diagnostic welfare triage alert).
---	--

Multilingual Processing and Detection Rationale: Arabic NLP research highlights challenges associated with dialectal variation, non-standard spelling and Arabic written in Latin characters. Research on Arabizi code-switching further supports treating language identification as a distinct preprocessing task (Guellil et al., 2021; Shehadi & Wintner, 2022). PsyberTrace will therefore use language-aware lexical patterns and normalization rules to flag candidate segments across supported English, Arabic and Arabizi forms. Modern contextual language models provide an alternative to purely keyword-based classification (Devlin et al., 2019); however, their suitability for the proposed offline environment and target languages must be established through testing. All flags will remain provisional and subject to specialist review.

3. SCIENTIFIC CALIBRATION & ETHICAL BOUNDARIES

What PsyberTrace Does	What PsyberTrace Explicitly Does Not Do
Preserves original evidence in read-only mode and verifies integrity via SHA-256 hashing (ISO/IEC 27037:2012 aligned).	Does not determine legal guilt, innocence, or criminal liability.
Organizes large chat exports into structured timelines with participant and metadata summaries.	Does not diagnose psychological disorders or clinical mental health conditions.
Highlights candidate conversation segments and displays the exact rule/pattern that triggered each flag.	Does not perform autonomous suicide-risk assessment or replace a specialist.
Requires human specialist review (Confirm, Dismiss, or Reclassify) before any flag enters the report.	Does not use opaque "Behavioural Threat Scores" that mask uncertainty.
Processes all case data 100% locally on air-gapped hardware and logs every reviewer action.	Does not transmit sensitive victim or case data to external cloud APIs.

Explainability and Human Oversight: PsyberTrace will provide source-linked explanations identifying the linguistic rule or pattern responsible for each candidate flag, alongside surrounding message context. Research on interpretable machine-learning methods, including LIME and SHAP, offers potential approaches for explaining individual model predictions (Ribeiro et al., 2016; Lundberg & Lee, 2017). These methods explain model behaviour rather than establish criminal intent or the truth of an allegation. Accordingly, PsyberTrace will preserve uncertainty, record reviewer decisions and require human verification before any candidate flag is included in the finalized report. Its design and evaluation will also be informed by the NIST AI Risk Management Framework (2023).

Calibration of Claims for Judicial Validity

Original Uncalibrated Concept	PsyberTrace Defensible Forensic Specification
"Detects psychological stages of manipulation"	Flags conversation segments that may warrant specialist review against defined linguistic patterns.
"Identifies exact timestamps where interaction pivoted to criminal coercion"	Highlights candidate time ranges and escalation clusters for reviewer attention.
"Detects victim acute crisis or self-harm risk"	Surfaces concerning language for immediate human review; explicitly labeled as non-diagnostic.
"Reduces review from four days to under ten minutes"	Aims to reduce initial triage time; performance is empirically measured against a manual baseline.
"Beats Cellebrite Pathfinder or Magnet AXIOM"	Complements existing forensic suites with a transparent, reviewer-correctable triage layer.

4. 10-STEP END-TO-END TRACEABLE WORKFLOW

1. **Authorized Import:** Ingests standard `.txt`, `.json`, or `.csv` chat exports from forensic extraction suites.
2. **Cryptographic Verification:** Computes pre-ingestion SHA-256 hash (H_{raw}), records `UTC+4` timestamp, and checks byte completeness.
3. **Read-Only Vaulting:** Locks original evidence file against mutation in an isolated vault directory.
4. **Structured Parsing:** Normalizes participants, timestamps, messages, attachments, line/byte offsets, and metadata anomalies.
5. **Timeline Visualization:** Renders interactive message frequency, response-latency shifts (Δt), and late-night burst timelines (`00:00–05:00`).
6. **Transparent Candidate Flagging:** Displays candidate segments (C1–C5) with Explainability Cards and ± 5 message context windows.
7. **Human-in-the-Loop Gate:** Reviewer explicitly selects `[Confirm Relevance]`, `[Dismiss False Positive]`, or `[Reclassify]`.
8. **Source-Linked Draft Report:** Compiles verified findings linked to exact message IDs, line numbers, and byte offsets.
9. **Audit Trail Sealing:** Logs reviewer decisions, software version, rule-set hash, and timestamps into an immutable audit log.
0. **Final Export:** Generates a court-auditable PDF report and cryptographically signed JSON audit manifest (H_{report}).

5. PROTOTYPE ARCHITECTURE, DEMONSTRATION & EVALUATION

Interactive Concept Prototype Notice (<https://psybertrace.pages.dev>): To provide evaluators with an immediate, visual demonstration of the PsyberTrace workflow during initial submission review, an interactive web-based concept demonstrator has been deployed at <https://psybertrace.pages.dev> (with full bilingual English/Arabic RTL support, Dark/Light workstation modes, and live report generation). This live demonstrator is strictly a visual showcase running on synthetic mock datasets and remains under active development. Upon shortlisting, the full offline production architecture and advanced cybersecurity hardening detailed below will be completed for the live air-gapped workstation exhibition.

Three-Scenario Synthetic Benchmark Suite Included in the Concept Prototype

Benchmark Scenario File	Triage Escalation Index	Forensic & Psycholinguistic Characteristics
<code>case_1_benign_peer_chat.json</code> (Benign Control Case)	~5% Index (No Predator Indicators)	Normal study coordination between two 15-year-old classmates with parental awareness. Contains one benign mention of <i>"don't tell him yet"</i> (regarding a brother's birthday gift) to demonstrate how investigators dismiss harmless false positives in the review gate.
<code>case_2_ambiguous_boundary_testing.json</code> (Ambiguous / Suspicious)	~55% Index (Needs Investigation)	Online gaming mentor flattering a 14-year-old player (C1: <i>"mature for your age"</i> , <i>"enti gher"</i>), offering a 500 AED voucher lure (C1), requesting parental secrecy (C2: <i>"bini w binek"</i>), and prompting migration to a private Telegram channel (C3), without explicit extortion (C4).
<code>case_3_critical_predator_extortion.json</code> (High-Confidence Predator)	94% Index (Critical Priority)	Full 5-stage escalation (C1–C5): initial flattery, parental secrecy, social isolation wedge (<i>"ahlek ma yefhamoon"</i>), 01:18 AM late-night burst, deleted message spoliation, offline <code>.opus</code> voice threat, family reputational blackmail (<i>"fadiha"</i>), 30-minute TRON/USDT crypto extortion ultimatum, and acute victim distress (C5).

Post-Selection Cybersecurity & Production Go-Live Roadmap (Air-Gapped Workstation Scope)

While the current submission prototype demonstrates core SHA-256 hashing, deterministic regex/lexicon screening, and reviewer-in-the-loop report generation, transitioning PsyberTrace into a forensic-grade production appliance post-selection involves the following cybersecurity and engineering milestones:

- Encrypted Evidence Vault & SQLCipher Storage:** Upgrading the local SQLite audit database (`psybertrace.db`) to AES-256 encrypted **SQLCipher** with append-only hash-chained audit logs

(Merkle-linked log entries) so any post-hoc tampering of reviewer logs is cryptographically detectable.

- **Hardware-Backed Cryptographic Signing (PKCS#11 / SmartCard):** Integrating examiner badge/smartcard digital signatures (X.509 / Ed25519) to cryptographically bind the examiner's identity to the exported PDF dossier and JSON manifest (`$H_{\text{report}}`).
- **Sandboxed Parser Hardening & Native UFDR/AXIOM Ingestion:** Implementing memory-safe, sandboxed streaming parsers for multi-gigabyte Cellebrite `.ufdr` XML/JSON dumps and Magnet AXIOM CSV exports, protecting the workstation against malformed-input parser exploits or zip-bomb payloads.
- **Air-Gapped OS Lockdown & Zero-Egress Enforcement:** Packaging the Python 3.11+ engine, local `faster-whisper` (CTranslate2 INT8) audio transcription model, and quantized offline transformer models (e.g., CAMEL-BERT / AraBERT with LIME/SHAP token attribution) inside a network-isolated container with OS-level firewall rules blocking all outbound socket calls.

Empirical Evaluation Metrics

System performance is evaluated across five core operational dimensions: (1) **Candidate Segment Recall & Precision** across categories C1–C5; (2) **False-Positive & Reviewer Dismissal Rates**; (3) **Initial Triage Time Reduction** compared against a manual linear review baseline; (4) **Cryptographic Integrity Verification Rate** (100% pre/post SHA-256 match); and (5) **Source-Linkage Traceability** (100% byte/line offset mapping).

Scientific Validation Protocol: PsyberTrace will be evaluated using a controlled, synthetic benchmark corpus containing both relevant and non-relevant conversational segments. Each segment will be labelled against predefined linguistic categories using documented annotation criteria. Where feasible, independent reviewers will label a subset of the corpus to assess inter-rater agreement and identify ambiguous cases. Following established classification-evaluation practices (Sokolova & Lapalme, 2009), the evaluation will report precision, recall, F1 score, false-positive rate and missed-segment rate, alongside reviewer correction rates and initial triage time compared with a documented manual-review baseline. Performance will be examined separately across supported language and writing-system categories where sufficient benchmark examples are available. The results will be reported transparently, including limitations arising from synthetic data, benchmark composition and the absence of validation on real-world case evidence. The evaluation approach will be informed by published research on machine-learning-based grooming detection (Leiva-Bianchi et al., 2025) and the NIST AI Risk Management Framework (2023).

6. REFERENCES

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., & Nouvel, D. (2021). Arabic natural language processing: An overview. *Journal of King Saud University – Computer and Information Sciences*, 33(5), 497–507. <https://doi.org/10.1016/j.jksuci.2019.02.006>
- Inches, G., & Crestani, F. (2012). Overview of the International Sexual Predator Identification Competition at PAN-2012. In *CLEF 2012 Evaluation Labs and Workshop: Online Working Notes*. CEUR Workshop Proceedings. <https://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-InchesEt2012.pdf>
- International Organization for Standardization, & International Electrotechnical Commission. (2012). *ISO/IEC 27037:2012: Information technology—Security techniques—Guidelines for identification, collection, acquisition and preservation of digital evidence*. International Organization for Standardization. <https://www.iso.org/standard/44381.html>
- Leiva-Bianchi, M., Castillo, N., Astudillo, C. A., & Ahumada-Méndez, F. (2025). Effectiveness of machine learning methods in detecting grooming: A systematic meta-analytic review. *Scientific Reports*, 15, Article 9008.

<https://doi.org/10.1038/s41598-024-83003-4>

- Lorenzo-Dus, N., Izura, C., & Pérez-Tattam, R. (2016). Understanding grooming discourse in computer-mediated environments. *Discourse, Context & Media*, 12, 40–50. <https://doi.org/10.1016/j.dcm.2016.02.004>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30). <https://papers.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
- O'Connell, R. (2003). *A typology of cybersexexploitation and on-line grooming practices*. Cyberspace Research Unit, University of Central Lancashire.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 97–101). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N16-3020>
- Shehadi, S., & Wintner, S. (2022). Identifying code-switching in Arabizi. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop* (pp. 194–204). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.wanlp-1.18>
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Stark, E. (2007). *Coercive control: How men entrap women in personal life*. Oxford University Press.
- United Arab Emirates. (2021). *Federal Decree-Law No. (34) of 2021 on Countering Rumors and Cybercrimes*. <https://uaelegislation.gov.ae/en/legislations/1526>