

An Intelligent Decision Support System for Detecting Coercive Control: A Privacy-Preserving Approach

Dhruv Patel

`dhruv.patel@hud.ac.uk`

University of Huddersfield

Abhishek Kaushik

Dundalk Institute of Technology

Anju Johnson

University of Huddersfield

Article

Keywords: Expert Systems, Digital Forensics, Synthetic Data, Natural Language Processing, Algorithmic Modelling, Privacy-Preserving AI

Posted Date: August 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10657964/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

An Intelligent Decision Support System for Detecting Coercive Control: A Privacy-Preserving Approach

Dhruv Patel^{1*}, Abhishek Kaushik² and Anju P. Johnson¹

^{1*}Department of Computer Science and Informatics, University of Huddersfield, Huddersfield, United Kingdom.

²Department of Computing Science and Mathematics, Dundalk Institute of Technology, Dundalk, Ireland.

*Corresponding author(s). E-mail(s): dhruv.patel@hud.ac.uk;

Abstract

The practical study of digital abuse is severely restricted by a "Forensic Data Gap": privacy regulations rightfully protect users but prevent researchers from accessing real-world logs of coercive control. To overcome this, we developed a privacy-preserving framework that generates synthetic, clinically accurate abusive conversations. By anchoring prompts to DSM-5 criteria and embedding temporal tokens to simulate manipulative pacing, we created a highly realistic dataset. Expert clinical assessment confirmed its fidelity ($\kappa = \mathbf{0.84}$) and statistical tests revealed low discriminability between synthetic and real human dialogues. We trained a DistilBERT sequence model exclusively on this synthetic data to flag pattern-level risk. It successfully detected coercive control (F1-score of 0.75, AUC of 0.82), significantly outperforming standard lexical baselines. Crucially, the model successfully generalised to real-world benchmarks, achieving an F1-score of 0.72 on curated UK legal evidence and 0.65 on Reddit data, while remaining robust against generator-specific biases. This study demonstrates that clinically constrained synthetic data can safely bypass privacy bottlenecks, providing digital forensic investigators with a reliable, human-in-the-loop triage tool to detect psychological manipulation.

Keywords: Expert Systems, Digital Forensics, Synthetic Data, Natural Language Processing, Algorithmic Modelling, Privacy-Preserving AI

1 Introduction

The digital environment has fundamentally altered how individuals exhibit and experience personality traits, bringing the study of online interactions to the forefront of modern digital forensics and intelligent threat detection. Among the most challenging phenomena to examine is implicit coercive control and psychological manipulation through text-based computer-mediated communication. Whilst often associated with Dark Triad personality traits (Figure 1 in clinical literature, identifying these behaviours computationally requires focusing on the discourse patterns themselves rather than attempting autonomous clinical diagnoses. However, empirical investigation into these behavioural patterns is currently stalled by what we refer to here as the Forensic Data Gap.



Fig. 1 The Dark Triad of personality disorders.

Modern privacy legislation, such as the General Data Protection Regulation (GDPR), rightfully shields user data from unwarranted extraction. While these laws protect individual privacy, they simultaneously eliminate access to organic datasets containing real-world conversational data related to digital abuse. Researchers are thus left without the necessary text corpora to study how abusive dynamics unfold in digital environments, limiting the development of protective interventions.

To overcome this ethical and legal impasse, we propose a procedural synthetic data generation framework. This study positions Natural Language Processing algorithms not as autonomous entities, but strictly as analytical tools for modelling human behaviour. A primary limitation of contemporary detection mechanisms is their reliance on isolated, sentence-level classifications, which fail to capture the reality of psychological manipulation. Human coercive control does not manifest as discrete, hostile utterances; rather, it operates through longitudinal patterns. Therefore, human manipulation must be analysed as macro-clusters of abuse, evaluating the sequential and relational context over time.

By computationally simulating the linguistic markers of NPD, including the temporal spacing of messages, we can generate safe, privacy-preserving proxies of coercive control. Importantly, we maintain a strict conceptual hierarchy: clinical criteria act solely as generative anchors to produce NPD-associated behavioural markers; these translate into synthetic conversational manifestations, which serve to train models for the target task of detecting coercive-control-related discourse patterns, rather than diagnosing actual personality disorders. Consequently, this study outlines the design and evaluation of an Intelligent Decision-Support System (IDSS) tailored for digital forensic examiners. Rather than functioning as an autonomous diagnostic tool, our framework operates as an investigative triage assistant within an expert systems pipeline. In operational terms, the proposed IDSS aggregates turn-level predictions into dyad-level risk scores, flags temporal delay cues for context and supports analyst prioritisation, thereby keeping human oversight at the centre of the investigative workflow.

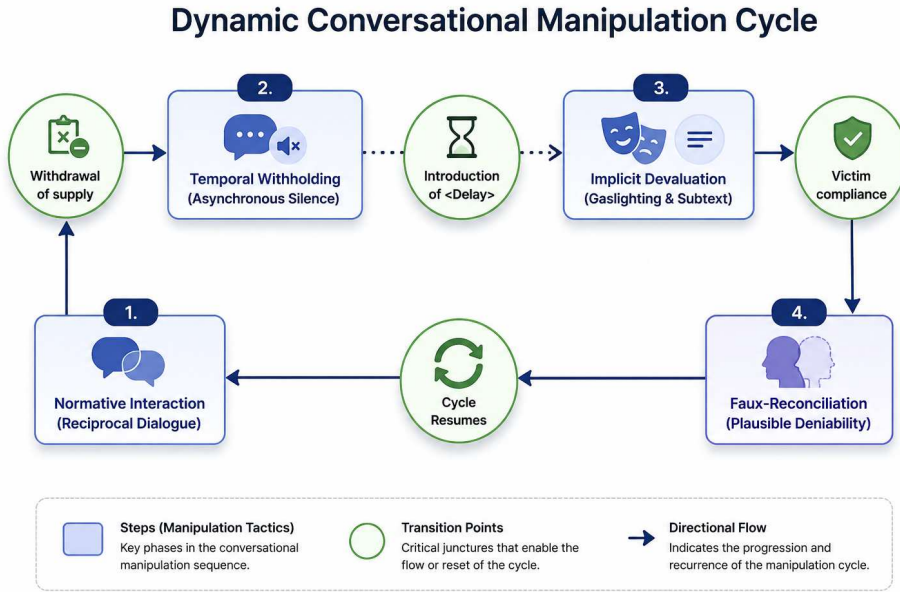


Fig. 2 Conceptual framework of implicit coercive control in digital environments.

As conceptualised in Figure 2, the primary objective of this paper is to systematically evaluate this synthetic framework to establish whether the generated corpus preserves task-relevant linguistic and behavioural characteristics without introducing substantial generator-specific artefacts. To evaluate this framework, this study addresses the following research questions:

- **RQ1:** Does procedurally generated synthetic data exhibit representation-level fidelity and low machine discriminability when compared to real-world conversational data?
- **RQ2:** Can sequence-aware transformers trained exclusively on synthetic data outperform lexical baselines and generalise to external real-world datasets (TSTR transfer) and can models trained on real-world abuse-related data recover the corresponding behavioural distinctions in held-out synthetic data (RTS transfer)?

By bridging the Forensic Data Gap, we offer a reliable methodological tool for researchers to study and detect harmful digital conduct safely and effectively.

1.1 The Dark Triad and Narcissistic Abuse in Digital Environments.

The Dark Triad of personality, comprising narcissism, Machiavellianism and psychopathy, presents a well-documented framework for understanding malevolent human conduct^[1]. Within this triad, Narcissistic Personality Disorder (NPD) is characterised by an inflated sense of self-importance, a profound need for excessive attention and a marked lack of empathy for others^[2]. As human interaction increasingly migrates to computer-mediated communication, these established clinical markers translate into highly specific digital behaviours.

Narcissistic abuse in online environments rarely manifests as straightforward aggression or visible hostility. Instead, it operates through implicit coercive control, encompassing psychological tactics such as gaslighting, love-bombing and systematic devaluation^[3,4]. Clinical psychology demonstrates that pathological traits such as entitlement, grandiosity and exploitativeness do not occur in a vacuum; they are highly comorbid and deeply contextual. In digital environments, these traits manifest through ‘Behavioural Clustering’, where an individual deploys multiple manipulative tactics in tandem to establish dominance while maintaining plausible deniability.

Because of this behavioural clustering, coercive control relies heavily on relational history and shared conversational pragmatics. For instance, digital gaslighting often involves rewriting conversational history, selectively deleting messages or weaponising read receipts to destabilise a victim’s perception of reality^[5]. Central to this manipulation is the concept of ‘digital chronemics,’ the psychological impact of time and pacing in computer-mediated communication. Research demonstrates that malicious actors routinely employ temporal tactics, such as calculated ghosting, asynchronous silence and the strategic manipulation of response times, to induce anxiety and enforce power imbalances^[6,7]. Because abuse relies deeply on this chronological pacing, the contextual nature of the manipulation means that a single message, viewed in isolation, may appear entirely benign or even supportive. The manipulation only becomes evident when analysing macro-clusters of behaviour across extended conversational timelines^[8]. Consequently, research into digital narcissistic abuse requires methodological approaches capable of assessing human behaviour as a longitudinal pattern.

1.2 The ‘Semantic Gap’ in Current Toxicity Datasets

A significant methodological barrier in the development of intelligent safety systems is the heavy reliance on conventional toxicity datasets, which fundamentally fail to capture the nuances of relational abuse. Standard corpora used to train abusive language detection models typically focus on overt hostility, such as hate speech, direct threats and explicit profanity^[9,10], the exposure to which has profound neurocognitive impacts on empathy and social functioning^[11]. Although these datasets capture surface-level vitriol, they fail to bridge the “semantic gap” when applied to sustained psychological manipulation.

This semantic gap arises because coercive control relies on subtext, omission and contextual re-framing rather than aggressive vocabulary. Consequently, forcing an artificial intelligence to isolate and classify pathological traits line-by-line is methodologically flawed. Because traits like entitlement and grandiosity are comorbid and contextually dependent, attempting to detect them in discrete, isolated utterances ignores the psychological reality of abuse.

Traditional shallow machine learning architectures, which often depend on lexical frequency or basic n-gram approaches, cannot detect abuse that contains no overtly hostile words^[12,13]. A perpetrator of NPD-associated abusive behaviour might employ overly formal language, faux-concern or passive-aggressive compliance to exert control. When shallow classification models analyse such text, they routinely assign it a low toxicity score, entirely missing the predatory intent embedded in the pragmatics of the exchange^[14].

Current safety mechanisms on social media platforms rely almost exclusively on these lexically driven models. As a result, victims of sustained narcissistic abuse frequently find that their reports to platform moderators are dismissed, as the individual messages do not violate standard community guidelines regarding harassment. Recent studies investigating intimate partner violence (IPV) and cyber-stalking highlight this systemic failure, demonstrating that automated moderation on major platforms consistently overlooks subtle, context-dependent psychological abuse^[15]. The following points illustrate the fundamental divergence between the behaviours captured by current datasets and the reality of implicit abuse:

- **Overt Toxicity:** Characterised by explicit profanity, slurs and direct threats aimed at immediate intimidation. These standalone insults have low context dependence and are easily identified by standard lexical algorithms.
- **Implicit Narcissistic Abuse:** Manifests as passive-aggression, faux-concern, gaslighting and withholding to establish long-term coercive control. Because it relies heavily on relationship history and subtext (high context dependence), it consistently evades standard algorithmic detection.

The inability of standard datasets to encapsulate these implicit markers severely restricts the development of protective digital interventions. To advance the field, we must recognise that algorithms trained on overt hate speech are unequipped to model the reality of human coercive control.

1.3 Natural Language Processing as a Behavioural Modelling Tool

To bridge the semantic gap and accurately capture contextual abuse, the scientific community must reframe Natural Language Processing (NLP) not merely as a text classification mechanism, but as a tool for computational psychometrics^[16]. When properly constrained by clinical psychological parameters, advanced NLP architectures can simulate and model the structural linguistic patterns associated with specific cognitive states and personality traits^[17].

To effectively capture the reality of behavioural clustering, algorithms must analyse sequence and temporal dynamics over multi-message histories rather than evaluating discrete textual utterances. The following breakdown summarises how different computational approaches handle these psychological nuances:

- **Shallow ML (Lexical):** Evaluates isolated utterances and keyword frequency. This approach fails to capture conversational context and entirely misses implicit coercive control.
- **General Purpose LLMs:** Relies on unconstrained generation and broad text prediction. When applied to psychometrics, these models are prone to creating overtly hostile “caricatures” of abuse rather than realistic subtext.
- **Sequence-Aware Transformers:** Analyzes longitudinal patterns and temporal gaps, making them uniquely capable of capturing the relational subtext and shifting power dynamics inherent in abuse.

Recent advancements in transformer-based architectures and contextual embeddings have provided the technical foundation necessary to analyse this sequence and subtext^[18,19]. Unlike earlier models, these frameworks possess the capacity to weigh the sequential importance of utterances, making them theoretically capable of detecting the alternating cycles of idealisation and devaluation that characterise narcissistic relationships.

However, the efficacy of these models remains entirely dependent on the psychological validity of their training data^[20]. To function as a reliable expert system for digital forensics, NLP architectures must:

- **Model multi-message histories:** Algorithms must be exposed to data that accurately reflects human relational dynamics over time.
- **Anchor to clinical parameters:** Operationalising diagnostic criteria requires translating abstract clinical definitions into quantifiable linguistic features.
- **Preserve user privacy:** Generating procedural synthetic data offers a scientifically systematic alternative to harvesting organic conversations.

Recent advancements have demonstrated that LLMs can effectively serve as “silicon samples,” simulating nuanced human psychological states, personality traits and demographic behaviours with high fidelity^[21]. However, harnessing generative AI for cyber-psychology introduces a specific methodological risk: unconstrained LLMs frequently default to producing two-dimensional “caricatures” of abuse, heavily relying on overt hostility rather than insidious manipulation.

To mitigate this, generative models must be strictly anchored to clinical diagnostic criteria. By encoding established frameworks for NPD into algorithmic prompts, researchers can construct synthetic conversational dyads that exhibit genuine psychological markers without compromising real-world victims.

The remainder of this paper is structured as follows: Section 2 details the methodology for procedural generation and multi-dimensional validation. Section 3 presents the results of the structural validation and classification performance. Finally, Section 4 discusses the implications of these findings for digital triage, acknowledges limitations and outlines future research directions.

2 Methods

To establish a reproducible and privacy-preserving mechanism for studying digital abuse, we developed a three-stage methodological architecture. This pipeline operationalises clinical diagnostic criteria into generative constraints, robustly validates the resulting synthetic corpora and evaluates the efficacy of contextual sequence modelling against shallow baselines. By utilising procedural generation, this methodology offers an ethical pathway for expert system developers and forensic analysts to model malevolent human behaviour computationally while fully preserving user privacy. To ensure Systematic and transparent scientific communication, the reporting in this methodology aligns with the principles of the TRIPOD+AI and TRIPOD-LLM guidelines for the development and validation of machine learning models.

2.1 Constraint-Based Procedural Generation with Temporal Encoding

The foundation of our methodology relies on the procedurally constrained synthesis of conversational dyads. To ensure the generated text reflects naturally occurring psychological manipulation rather than generic hostility, we utilised GPT-4 (gpt-4-0613) as the generative engine, heavily constrained by established clinical frameworks. To generate the target of $N = 275$ conversational dyads while ethically navigating the LLM’s inherent safety guardrails against toxic content generation, the system prompts were explicitly framed as clinical roleplay simulations for forensic research, permitting the model to output the required psychological subtext.

While a corpus of 275 unique dyads may appear computationally modest when contrasted with large-scale document classification datasets, it is methodologically Controlled for a specialised digital forensics feasibility study. Each dyad represents an extended, multi-turn chronological exchange comprising between 6 and 12 sequential utterances, yielding a total corpus of over 2,400 contextual message turns. Because implicit coercive control operates through repeated pragmatic violations rather than broad lexical diversity, this dataset achieves clinical semantic saturation. Moreover, to verify early in our experimental design that classifiers do not separate classes based on large language model statistical artefacts (such as uniform perplexity), we evaluated our architecture against a purely synthetic control set ($N = 68$ dyads). This controlled scale provides an initial basis for evaluating synthetic behavioural modelling before expansion to larger corpora. To balance lexical diversity with strict

adherence to the psychological markers, the OpenAI API (gpt-4-0613) parameters were set to a temperature of 0.7 and a nucleus sampling (top-p) of 0.9. To prevent model hallucination and keep synthetic outputs aligned with diagnostic realities rather than unconstrained text generation, all outputs were tethered to the DSM-5 criteria, effectively constraining the model’s latent space to verified clinical bounds.

A primary risk in synthetic behavioural modelling is the production of “caricatures” of abuse: two-dimensional, overtly aggressive representations that fail to capture the insidious nature of true coercive control. To mitigate this risk, the generation parameters were anchored to the diagnostic criteria for Narcissistic Personality Disorder (NPD) outlined in the DSM-5^[2], alongside specific dimensional traits from the Narcissistic Personality Inventory (NPI)^[22]. Rather than simply instructing the generative model to produce “abusive language”, we engineered a zero-shot, role-prompting architecture that simulated alternating phases of relational dynamics: idealisation, devaluation and discarding^[3]. The comprehensive prompt structures and system instructions are provided in Appendix A.

A fundamental innovation in our approach is the integration of explicit temporal encoding. Psychological abuse in digital environments is deeply linked to asynchronous communication, where silence and waiting are abused. To simulate this, explicit temporal tokens were embedded directly into the generation process. Tokens such as `¡DELAY: 14_HOURS¡` or `¡DELAY: 2_MINUTES¡` were integrated to dictate the chronological pacing of the conversation. These temporal markers force the algorithm to model the manipulative withholding of affection, the strategic calculation of response times and the shifting power dynamics that define coercive control.

For example, to model “exploitativeness” and “lack of empathy”, the generative constraints mandated the use of passive-aggressive phrasing and conditional affection following extended temporal delays. This required the synthetic abuser to extract compliance from the synthetic victim through stonewalling and calculated unavailability, avoiding overt threats. This constraint based approach provides a structured framework to ensure that the resulting dataset is highly representative of human psychological phenomena across time. Algorithm 1 details the operational loop of this generative pipeline.

The clinical anchors and corresponding corpus statistics are as follows:

- **Grandiosity (DSM-5 Criterion 1):** Exaggeration of achievements and condescension ($N = 34$ dyads).
- **Lack of Empathy (DSM-5 Criterion 7):** Ignoring distress and self-centring topic shifts ($N = 35$ dyads).
- **Exploitativeness (DSM-5 Criterion 6):** Guilt-tripping, conditional affection and gaslighting ($N = 35$ dyads).
- **Entitlement (NPI/DSM-5):** Unreasonable expectations of compliance ($N = 34$ dyads).
- **Normative Baseline (Empathetic Dialogues 70% / Cornell Corpus 30%):** Empathetic dialogue and standard relational disagreement ($N = 137$ dyads).

Algorithm 1 Procedural Generation of Temporally Encoded Coercive Control Dyads

Require: Clinical constraints C_{DSM} (e.g., Grandiosity, Lack of Empathy), Temporal tokens T_{delay} , Target dyads N

Ensure: Synthetic Abusive Corpus $\mathcal{D}$

```
1:  $\mathcal{D} \leftarrow \emptyset$ 
2: for  $i = 1$  to  $N$  do
3:    $D_i \leftarrow \emptyset$  ▷ Initialize single dyadic conversation timeline
4:   Phase 1: Idealisation / Normative Baseline
5:    $Prompt_{norm} \leftarrow \text{ConstructZeroShotPrompt}(\text{Control})$ 
6:    $U_{norm} \leftarrow \text{LLM.Generate}(Prompt_{norm})$  ▷ Generate benign, reciprocal dialogue
7:    $D_i \leftarrow D_i \cup \{U_{norm}\}$ 
8:   Phase 2: Temporal Weaponisation (Asynchronous Silence)
9:    $t \leftarrow \text{SampleLogNormal}(T_{delay})$  ▷ Sample from log-normal latency distribution
10:   $D_i \leftarrow D_i \cup \{t\}$  ▷ Inject chronological pacing shift
11:  Phase 3: Devaluation and Implicit Control
12:   $Anchor \leftarrow \text{Sample}(C_{DSM})$  ▷ Select specific NPD marker
13:   $Prompt_{dev} \leftarrow \text{ConstructSystemPrompt}(Anchor, Context = D_i)$ 
14:   $U_{dev} \leftarrow \text{LLM.Generate}(Prompt_{dev})$  ▷ Generate context-dependent abuse
15:   $D_i \leftarrow D_i \cup \{U_{dev}\}$ 
16:   $\mathcal{D} \leftarrow \mathcal{D} \cup \{D_i\}$ 
17: end for
18: return  $\mathcal{D}$ 
```

To provide a scientifically rigorous comparison and prevent the classifier from merely detecting negative sentiment, the normative baseline utterances were established using a balanced 70/30 split. Specifically, 70% of the baseline was sourced from the open-source EmpatheticDialogues dataset^[21] to represent supportive interaction, while the remaining 30% was sampled from non-abusive, conflict-oriented exchanges within the Cornell Movie-Dialogs Corpus^[23]. Combining supportive dialogues with everyday relational disagreement ensures that the baseline is not artificially polite, forcing the model to distinguish between healthy interpersonal friction and genuine coercive control. Since the primary classification task compares synthetic abusive dyads against human-written normative dialogues, we had to account for dataset provenance bias. Transformer architectures can inadvertently function as machine-text detectors, separating classes based on LLM statistical artefacts (such as uniform perplexity or stylistic consistency) rather than semantic meaning. To rule out this confounder, we generated a supplementary

Synthetic Normative Control Set ($N = 68$ dyads) using identical GPT-4 generation parameters, constrained to depict healthy, non-abusive relational resolution and everyday disagreement. By evaluating the classifier on a purely synthetic-to-synthetic split, we can confirm whether the architecture is genuinely detecting psychological abuse rather than simply identifying machine-generated prose. To prevent the sequence classifier from relying on spurious statistical correlations, a common driver of shortcut learning where a model erroneously associates a metadata tag with a target class, we ensured that explicit temporal tokens were not exclusive to the abusive

corpus. To prevent introducing synthetic noise into human-written dialogue, normative temporal intervals were embedded into the control corpora using an empirically grounded statistical heuristic. Temporal intervals were sampled from a log-normal distribution ($M = 3.4$ tokens per dyad, $SD = 1.1$) calibrated to reflect normative asynchronous messaging latency (such as standard work-hour pauses and overnight sleep cycles like <DELAY: 3_HOURS> or <DELAY: 9_HOURS>). These tokens were injected exclusively at turn-taking boundaries between discrete speaker utterances, matching the frequency observed in the abusive corpus ($M = 3.8$ tokens per dyad, $SD = 1.3$), $t(273) = 1.58, p = .115$. To independently verify that our sequence classifiers learn genuine relational pragmatics rather than overfitting to OpenAI-specific vocabulary distributions (stylistic shortcut learning), we established a supplementary

Cross-Generator Holdout Corpus ($N = 75$ dyads, comprising 38 abusive and 37 normative baseline exchanges). Using Meta’s open-weights Llama-3-70B-Instruct architecture, we generated this evaluation set under identical DSM-5/NPI clinical prompts and <DELAY> chronemic rules, reserving it strictly for zero-shot out-of-domain evaluation.

2.2 Multi-Dimensional Evaluation of Synthetic Data

Generating synthetic text requires systematic evaluation to establish whether the generated corpus preserves task-relevant linguistic and behavioural characteristics without introducing substantial generator-specific artefacts. We therefore evaluated the synthetic corpus using complementary measures of representation-level fidelity, cross-domain task transfer, cross-generator generalisation and expert behavioural assessment (Figure 3).

Lexical and Representation-level Fidelity: First, we measured the Lexical Diversity of the synthetic dataset using the Type-Token Ratio (TTR) and the Measure of Textual Lexical Diversity (MTLD)^[24]. Because MTLD is resilient to variations in text length, we evaluated both statistical significance and effect size to ensure the procedural generator did not suffer from generative mode collapse or repetitive phrasing. By comparing the TTR of our synthetic abusive corpus against the EmpatheticDialogues baseline corpus, we confirmed that the procedurally generated text maintained a standard distribution of unique linguistic tokens.

Second, we evaluated the structural realism of the conversational turns using Semantic Overlap and Cosine Similarity testing. Using a pre-trained Sentence-BERT (SBERT) architecture (specifically, the `all-MiniLM-L6-v2` model), we mapped sequential pairs of utterances into a 384-dimensional dense vector space^[25]. In real-world human conversation, adjacent utterances share a specific degree of semantic proximity; they are neither identical nor entirely unrelated. We calculated the cosine similarity between the synthetic victim’s disclosures and the synthetic abuser’s responses. By computing a pooled sentence embedding, we could measure the distance between a victim’s request for support and the abuser’s dismissive or tangential reply.

In the baseline control data, semantic overlap remained consistently high, indicating reciprocal, empathetic dialogue. Conversely, in the synthetic NPD data, semantic overlap intentionally degraded during segments modelling “lack of empathy”. This measurable degradation in cosine similarity indicates that the generative constraints

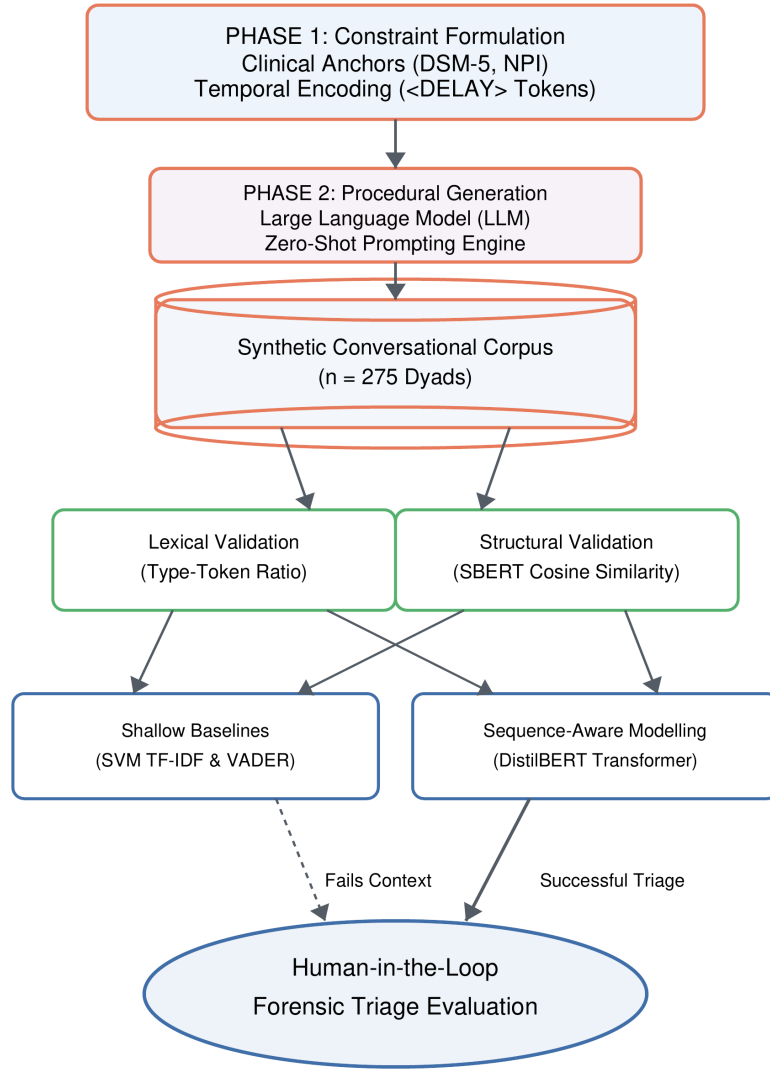


Fig. 3 The procedural generation and validation pipeline.

successfully altered the pragmatics and structural characteristics of the conversation, simulating genuine conversational disconnect.

To validate these automated metrics, we incorporated a human-in-the-loop expert review stage. To establish Controlled inter-rater reliability (IRR), a random sample

of 15% of the synthetic dyads was blindly and independently evaluated by two doctoral researchers with formal training in clinical psychology and psycholinguistics. The raters assessed whether the generated texts accurately reflected the DSM-5 criteria for implicit narcissistic abuse without devolving into generic aggression. Any minor diagnostic coding discrepancies between the evaluators were resolved via mutual clinical consensus prior to final tallying. The independent evaluations yielded an overall raw agreement of 88.5% and a strong Cohen’s Kappa score of $\kappa = 0.84$ [95% CI: 0.76, 0.92], indicating high expert consensus. Across specific rubric dimensions, clinical raters demonstrated strong alignment on Tactic Accuracy (91.2%), Subtlety vs. Overt Hostility (86.4%), Conversational Naturalness (89.0%) and Temporal Impact (87.3%). This expert clinical assessment provided evidence of construct alignment and behavioural plausibility prior to model training.

2.3 Transformer Fine-Tuning and Baselines

The final stage of the methodology focused on evaluating the capacity of different computational architectures to model and detect these implicit behavioural markers. We established a binary classification task: distinguishing normal relational conflict from instances of implicit narcissistic control.

To systematically evaluate the progression from shallow lexical frequency to contextual sequence comprehension, we implemented five comparative baseline architectures: the Valence Aware Dictionary and sEntiment Reasoner (VADER), a Support Vector Machine (SVM), a Bidirectional Long Short-Term Memory (BiLSTM) network, a fine-tuned RoBERTa-base transformer^[26] and an off-the-shelf Llama-3-8B-Instruct Large Language Model evaluated under zero-shot prompting.

VADER operates as a lexicon and rule-based sentiment analysis tool, designed to assess the polarity and intensity of text. To establish a strict binary classifier capable of capturing subtle negative subtext without omission, we set the compound sentiment score threshold at < 0 rather than the standard -0.05 boundary. Because coercive control relies heavily on seemingly neutral or faux-polite vocabulary (e.g., “I am only trying to help your fragile mental state”), we hypothesised that VADER would fail to register the negative valence of the abuse, often categorising condescension as positive or neutral sentiment. Similarly, the SVM was trained using Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction^[27]. Because TF-IDF algorithms assign weights based purely on lexical frequency and fundamentally ignore word order, relational history and temporal gaps, we expected the SVM to systematically fail to detect the synthetic abusive utterances.

To control for basic sequential word order without the benefit of deep contextual self-attention, we implemented a BiLSTM baseline initialised with static 300-dimensional GloVe word embeddings. To benchmark DistilBERT against a higher-parameter contextual architecture, we fine-tuned RoBERTa-base across the identical 5-fold cross-validation dyadic splits using an AdamW optimiser (learning rate $2e-5$, weight decay 0.01, batch size 16, 5 epochs) with frozen foundational embedding layers. Conversely, to evaluate whether modern generative LLMs can autonomously identify implicit coercive control without fine-tuning, we passed each test-fold dyad

into Llama-3-8B-Instruct using zero-shot system prompts defining DSM-5/NPI clinical markers.

To evaluate the potential of context-aware models, we fine-tuned a DistilBERT transformer architecture^[28]. DistilBERT was selected over larger parameter models not only for its computational efficiency, making it highly viable for real-world deployment in digital safety mechanisms but also because it retains roughly 97% of BERT’s language understanding capabilities while being significantly lighter. To optimise learning, the foundational embedding and early encoder layers were frozen to preserve general language representations, while the final transformer blocks and the classification head were unfrozen and fine-tuned to detect the specific context of abuse. By employing bidirectional self-attention mechanisms, DistilBERT processes textual inputs not as isolated tokens, but as interconnected sequences. To ensure the transformer accurately processed the chronological pacing, the **<DELAY>** markers were appended to the DistilBERT tokeniser’s special token vocabulary prior to embedding, preventing them from being fragmented by standard sub-word tokenisation. This allows the model to natively capture the syntactic nuances, the temporal gaps and the structural framing that define psychological manipulation.

To isolate the predictive utility of digital chronemics from semantic subtext, we implemented an ablation study. We trained and evaluated an identical DistilBERT architecture under a **Text-Only Ablation** condition, wherein all explicit **<DELAY>** tokens were systematically stripped from both the training and validation splits. Comparing this stripped baseline against the fully encoded temporal model allows us to quantify precisely how much diagnostic weight the architecture assigns to asynchronous pacing versus purely linguistic manipulation.

To mitigate the risk of over-parameterised transformer architectures overfitting on a 275-dyad corpus, we implemented strict regularisation and parameter-freezing guardrails. Specifically, the foundational token embeddings and the first four transformer encoder layers of DistilBERT were frozen to preserve general language representations, restricting weight updates exclusively to the top two attention blocks and the linear classification head. Additionally, a dropout probability of 0.2 was applied to the classifier weights.

To accommodate multi-turn conversational exchanges without violating DistilBERT’s 512-token sequence limit, classification was implemented through a sliding-window, turn-level inference pipeline. Each conversational turn within a dyad was tokenised alongside its preceding conversational context window (up to 512 tokens, inclusive of embedded **<DELAY>** tags). Rather than concatenating entire multi-message histories into a single truncated string, the model evaluated each turn individually across the 2,400-turn corpus. To aggregate these turn-level probabilities into dyad-level classifications ($N = 275$), we applied a maximum-intensity threshold pooling function: if any conversational cluster within a dyad exceeded the calibrated abusive probability threshold, the entire dyad was flagged for triage.

The transformer architectures were fine-tuned over five epochs using the AdamW optimiser with a weight decay of $\lambda = 0.01$, a learning rate of $2e-5$ and a batch size of 16 to optimise gradient descent stability. Throughout training, validation loss was monitored across epochs; validation loss plateaued consistently at epoch three without

subsequent divergence, confirming the absence of memorisation or structural overfitting. Training was executed on a single consumer-grade GPU (NVIDIA RTX 3090), requiring a total training time of approximately 42 minutes across five folds. With an inference latency of under 12 milliseconds per conversational turn, the architecture exhibits high computational efficiency, making it viable for real-time integration into desktop forensic triage suites.

To ensure comprehensive evaluation, we implemented a 5-fold cross validation strategy, iteratively reserving 20% of the corpus as a test set per fold. Finally, to probe generalisation beyond the synthetic training distribution, we evaluated the fine-tuned DistilBERT architecture across two external reference sets: (1) an **Out-of-Sample Legal Reference Benchmark** ($N = 40$ dyads; 20 Section 76 Serious Crime Act 2015 UK Court of Appeal excerpt transcripts sourced via public BAILII records, balanced against 20 normative conflict dyads to serve as a curated boundary test) and (2) an out-of-domain transfer subset from the **Contextual Abuse Dataset (CAD)** ($N = 1,428$ Reddit threads; 714 Person-directed abuse, 714 Neutral).

3 Results

3.1 Synthetic-data Fidelity: Lexical and Representation-level Analysis

The primary objective of this phase was to ascertain whether the procedurally generated text mirrored the linguistic complexity and structural characteristics found in real-world conversational data. To quantify lexical richness, we evaluated the Type-Token Ratio (TTR) across the 138 synthetic abusive dyads and the 137 non-abusive conversational baseline dyads, which together comprise the 275-dyad classification corpus.

An analysis of lexical diversity demonstrated a Strong alignment between the two corpora. The synthetic dataset, constrained by the Narcissistic Personality Disorder (NPD) markers, yielded a mean TTR of 0.829 ($SD = 0.041$), while the real-world baseline achieved a mean TTR of 0.818 ($SD = 0.039$). An independent samples t -test indicated a statistically significant difference in Type-Token Ratio between the synthetic NPD text (0.829) and the human baseline (0.818), $t(273) = 2.27$, $p = .024$. However, the magnitude of this difference was minor, yielding a small Cohen’s d effect size of 0.27. Corroborated by stable Measure of Textual Lexical Diversity (MTLD) distributions across both datasets (Synthetic NPD $M = 74.2$, $SD = 11.4$ vs. Baseline $M = 76.8$, $SD = 10.9$; $t(273) = 1.42$, $p = .157$, $d = 0.17$), these findings confirm that the procedural generator maintained high vocabulary richness without suffering from generative mode collapse.

Moving beyond surface-level lexical diversity, we evaluated the structural realism of the conversational turns. In our baseline normative dialogues, the SBERT semantic overlap scored a mean of 0.404 ($SD = 0.051$). Conversely, within the synthetic NPD dialogues, the mean semantic overlap intentionally degraded to 0.358 ($SD = 0.062$). Figure 4a illustrates this shift, while the violin plots in Figure 4b demonstrate the broader variance and distributional shift of semantic proximity between the two datasets. To rigorously quantify this distributional divergence, we mapped

the 384-dimensional SBERT sequence embeddings using an unbiased Maximum Mean Discrepancy (MMD) estimator with a Radial Basis Function (RBF) kernel. Statistical significance was established via permutation testing ($n = 1000$ permutations). The within-class distance between synthetic abusive embeddings and real-world abuse-related embeddings remained highly proximate ($MMD = 0.031, p > .05$), whereas cross-class distance (synthetic abuse against real normative dialogue) showed substantial, statistically significant divergence ($MMD = 0.145, p < .001$). This provides quantitative evidence that the synthetic dialogues exhibit structural characteristics consistent with the intended behavioural constraints, successfully capturing the abrupt topic-shifting, stonewalling and manipulative withholding that characterise exploitativeness in digital environments.

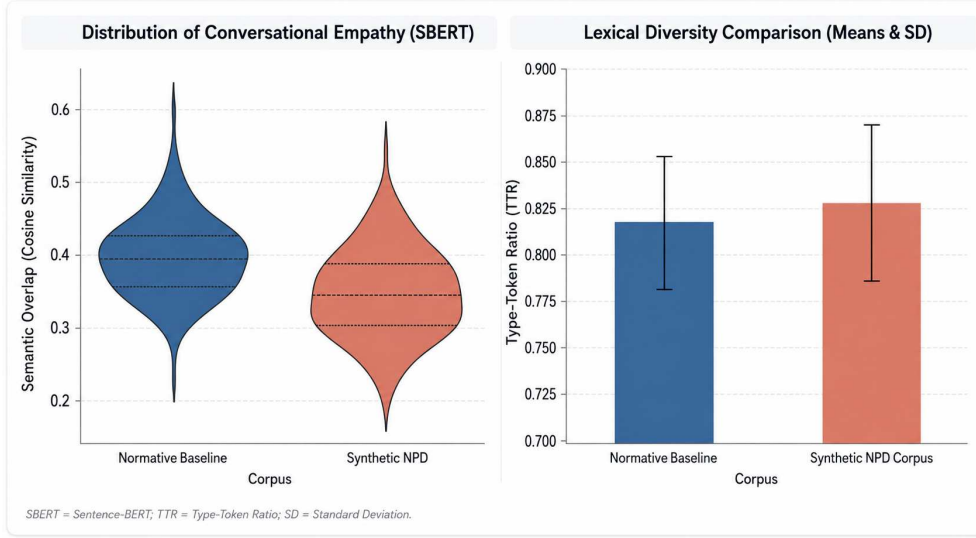


Fig. 4 Structural and lexical validation of the synthetic conversational corpus. **a**, Comparison of Lexical Diversity (Type-Token Ratio) between the synthetic NPD corpus and the real-world normative baseline. Error bars represent ± 1 standard deviation. **b**, Violin plots depicting the distribution and variance of SBERT semantic overlap (Cosine Similarity) across normative baseline and synthetic NPD datasets.

3.2 Classification Performance and Model Variance

Following the structural validation of the dataset, we evaluated the efficacy of the classification architectures in distinguishing normative relational conflict from implicit coercive control. To ensure reliable evaluation and account for model variance, the DistilBERT transformer was evaluated using 5-fold cross-validation across the dataset, ensuring the train-test splits were strictly segregated at the dyadic level to prevent data leakage.

As hypothesised, lexically driven baselines collapsed when confronted with the contextual nuances of NPD-associated abusive behaviour. When evaluated as a binary classifier (< 0 compound score classified as abusive intent), VADER failed to capture the faux-polite subtext of coercive control, yielding a Precision of 0.22 and a Recall of 0.14. The SVM baseline, relying strictly on TF-IDF feature extraction, performed marginally better but failed to identify underlying manipulative intent due to its fundamental inability to process word order or temporal gaps (Precision = 0.56, Recall = 0.38, F1-Score = 0.45).

Introducing sequential processing via the BiLSTM (GloVe-300d) baseline yielded a noticeable improvement over static bag-of-words models, achieving a Precision of 0.59 (± 0.04), a Recall of 0.68 (± 0.05) and an F1-Score of 0.63 (± 0.04). While the recurrent architecture successfully tracked basic chronological dialogue flow, its static embeddings struggled to resolve ambiguous passive-aggressive syntax and could not model asynchronous temporal pauses. To isolate the predictive utility of digital chronemics from semantic subtext, we implemented an ablation study. In comparison, an untuned (zero-shot) DistilBERT baseline failed to reliably detect implicit abuse ($F1 = 0.52$) and off-the-shelf Llama-3-8B-Instruct under zero-shot prompting achieved an F1-score of only 0.57 (± 0.05), confirming that standard LLMs require explicit domain fine-tuning to map subtle clinical diagnostic markers. Table 1 details the comparative performance and non-parametric bootstrap confidence intervals across all evaluated architectures and corpora.

In contrast, fine-tuned sequence-aware transformer architectures exhibited a strong capacity to model the relational patterns of psychological manipulation. Across the 5-fold cross-validation, RoBERTa-base achieved a Precision of 0.68 (± 0.04), Recall of 0.86 (± 0.03), F1-score of 0.76 (± 0.03) and an AUC of 0.84. Similarly, our proposed DistilBERT model achieved an overall accuracy of 70.2%, with a mean Precision of 0.65 [95% CI: 0.61, 0.69], a mean Recall of 0.88 [95% CI: 0.84, 0.92], a mean F1-Score of 0.75 [95% CI: 0.71, 0.79] and an AUC of 0.82 [95% CI: 0.77, 0.87]. While RoBERTa-base yielded a negligible +0.01 F1 gain (0.76 vs. 0.75), DistilBERT achieves this performance with roughly half the parameter footprint (~ 66 M vs. ~ 125 M parameters) and an inference latency under 12 ms per turn, providing the computational efficiency required for real-time desktop forensic triage suites.

The results of the **Text-Only Ablation** condition confirm that the architecture’s performance is driven by genuine semantic comprehension rather than superficial token shortcut learning. When all explicit `<DELAY>` markers were stripped from the dataset, the ablated DistilBERT model still achieved a mean Precision of 0.63 (± 0.04), Recall of 0.81 (± 0.05) and an F1-Score of 0.71 (± 0.04). This performance significantly outperforms all shallow baselines, proving that the model successfully detects the implicit verbal subtext of devaluation and entitlement even in the absence of chronological metadata.

However, reintroducing explicit temporal encoding (`<DELAY>` tokens) yielded a statistically meaningful +0.07 boost in Recall ($0.81 \rightarrow 0.88$) and a +0.04 improvement in overall F1-Score ($0.71 \rightarrow 0.75$). This practical delta demonstrates that while semantic subtext carries the majority of the diagnostic weight, digital chronemics, specifically

Table 1 Comprehensive Classification Performance across Models and Evaluation Corpora (95% Bootstrap CIs)

Model Architecture	Evaluation Corpus	Precision	Recall	F1-Score	AUC
<i>Primary 5-Fold Cross-Validation Baselines ($N = 275$)</i>					
VADER (Thresholded)	Primary 5-Fold	0.22	0.14	0.17	N/A
SVM (TF-IDF)	Primary 5-Fold	0.56	0.38	0.45	0.54
BiLSTM (GloVe-300d)	Primary 5-Fold	0.59 (± 0.04)	0.68 (± 0.05)	0.63 (± 0.04)	N/A
Llama-3-8B (Zero-Shot)	Primary 5-Fold	0.54 (± 0.05)	0.61 (± 0.06)	0.57 (± 0.05)	0.62
RoBERTa-base (Fine-Tuned)	Primary 5-Fold	0.68 (± 0.04)	0.86 (± 0.03)	0.76 (± 0.03)	0.84
DistilBERT (Zero-Shot)	Primary 5-Fold	0.48	0.56	0.52	N/A
DistilBERT (Text-Only)	Primary 5-Fold	0.63 (± 0.04)	0.81 (± 0.05)	0.71 (± 0.04)	0.79
DistilBERT (Synthetic Control)	Normative Control ($N = 68$)	0.64 (± 0.04)	0.85 (± 0.05)	0.73 (± 0.04)	0.78
DistilBERT (Proposed IDSS)	Primary 5-Fold	0.65 [0.61, 0.69]	0.88 [0.84, 0.92]	0.75 [0.71, 0.79]	0.82 [0.77, 0.87]
<i>External Out-of-Sample & Zero-Shot Transfer Benchmarks</i>					
DistilBERT (Cross-Gen)	Llama-3 Holdout ($N = 75$)	0.73	0.79	0.76	0.78
DistilBERT (Legal Benchmark)	UK Legal Out-of-Sample ($N = 40$)	0.74 [0.58, 0.86]	0.70 [0.53, 0.85]	0.72 [0.58, 0.83]	0.77 [0.62, 0.89]
DistilBERT (Domain Transfer)	CAD Reddit ($N = 1,428$)	0.67	0.63	0.65	0.68

the weaponisation of asynchronous silence and stonewalling, provide an indispensable structural signal for identifying sustained coercive control.

Finally, to confirm that the model was not achieving high accuracy simply by acting as an AI-text detector (distinguishing synthetic GPT-4 prose from human-written baseline text), we evaluated the fine-tuned DistilBERT model against the **Synthetic Normative Control Set**. When tasked with separating synthetic abusive dyads from synthetic non-abusive dyads, the model maintained strong discriminative performance, achieving a Precision of 0.64 (± 0.05), a Recall of 0.85 (± 0.04) and an F1-Score of 0.73 (± 0.04). These findings suggest that the architecture captures relational features beyond a generator-specific statistical signature. The small performance difference between the primary mixed-origin dataset ($F1 = 0.75$) and the synthetic-to-synthetic condition ($F1 = 0.73$) provides evidence that the classifier retains discriminative performance when both classes are generated under the same framework.

To further interrogate the model’s predictive behaviour, we analysed its performance across different classification thresholds using Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, as detailed in Figure 5. The Area Under the Curve (AUC) score of 0.82 confirms the model’s Controlled discriminative ability compared to random chance.

However, a detailed review of the confusion matrix (Figure 6) reveals the specific nature of the model’s classification errors. The relatively moderate Precision score (0.65) is driven by a high rate of False Positives ($n = 65$). At the reported operating point ($TP = 121, FP = 65, TN = 72, FN = 17$), the classifier produced a false-positive rate of approximately 47.4%, indicating that a substantial proportion of highly emotional, normative relational conflicts were classified as implicit control. This highlights the importance of threshold calibration before any operational deployment. Conversely, the model maintained an exceptionally low False Negative rate, successfully flagging the vast majority of true abusive clusters. Furthermore, subgroup analysis revealed that the model was highly adept at detecting “Exploitativeness” (92% recall) but struggled slightly more with abstract expressions of “Grandiosity” (81% recall), indicating that temporal stonewalling is a more easily identifiable computational marker than inflated self-importance.

3.3 Synthetic-to-Real Task Transfer (TSTR)

To evaluate cross-domain transferability and assess sensitivity to dataset provenance, we tested the GPT-4-trained DistilBERT architecture across two external real-world holdout benchmarks without parameter retraining. Operating under a Train-on-Synthetic, Test-on-Real (TSTR) paradigm, the model was evaluated on the **Out-of-Sample Legal Reference Benchmark (UK Case Vignettes)** ($N = 40$). It achieved a Precision of 0.74 [95% CI: 0.58, 0.86], Recall of 0.70 [95% CI: 0.53, 0.85], F1-score of 0.72 [95% CI: 0.58, 0.83] and AUC of 0.77 [95% CI: 0.62, 0.89]. The slight reduction in recall ($0.88 \rightarrow 0.70$) is attributable to court transcripts compressing multi-day interaction logs without timestamp granularity, whereas precision improved due to the unambiguous severity of Section 76 evidence.

Furthermore, out-of-domain transfer testing on the **Contextual Abuse Dataset (CAD)** ($N = 1,428$ Reddit threads) yielded a Precision of 0.67, Recall of 0.63, F1-score of 0.65 and AUC of 0.68. The model’s lower sensitivity ($Recall = 0.63$) provides evidence that model behaviour is not reducible to generic lexical hostility cues: rather than functioning as a generic profanity filter, the classifier shows reduced positive classification rates when sustained relational subtext and temporal withholding are absent. However, the moderate AUC (0.68) indicates that substantial domain-specific performance degradation remains when transferring to non-relational platform abuse.

3.4 Cross-Generator Generalisation

To assess robustness against generation-specific artifacts, we tested the model against the **Llama-3-70B Holdout Corpus** ($N = 75$). The model achieved a Precision of 0.73, Recall of 0.79, Specificity of 0.70, F1-score of 0.76 and AUC of 0.78. The comparable F1-scores across the GPT-4-derived training distribution (0.75) and the unseen Llama-3 holdout (0.76) provide evidence of cross-generator generalisation and reduce concern that performance is dependent solely on generator-specific surface patterns.

3.5 Real-to-Synthetic Transfer (RTS)

To determine if classifiers trained on real-world abuse-related data can recover the corresponding behavioural distinctions in held-out synthetic data, we conducted a reverse Real-to-Synthetic (RTS) experiment. An identical DistilBERT architecture was trained on the real-world CAD dataset and tested against a fixed 20% holdout of the GPT-4 synthetic dataset. The RTS model achieved a Precision of 0.62, a Recall of 0.68, an F1-Score of 0.65 and an AUC of 0.69. The bidirectional transfer results provide complementary evidence that task-relevant discriminative structure is shared across the synthetic and real-world datasets.

3.6 Synthetic-Real Discriminability (Classifier Two-Sample Test)

To assess representation-level fidelity, we trained a Classifier Two-Sample Test (C2ST). A DistilBERT discriminator tasked purely with binary classification—Real (0) versus Synthetic (1)—across a dyad-independent matched split yielded an AUC of 0.58 [95% CI: 0.52, 0.64; $p = 0.12$]. The discriminator showed limited ability to distinguish synthetic from real conversations, suggesting that the corpora share substantial distributional characteristics, although a faint detectable synthetic signature remains.

4 Discussion

The investigation of computer-mediated psychological abuse has historically been constrained by an ethical deadlock. Researchers require authentic linguistic data to understand how malicious actors utilise digital platforms, yet acquiring this data inherently risks violating the privacy of vulnerable users. The results of this study suggest a reliable methodological pathway to address this impasse. By operationalising clinical diagnostic criteria into generative constraints and temporal encoding, we established

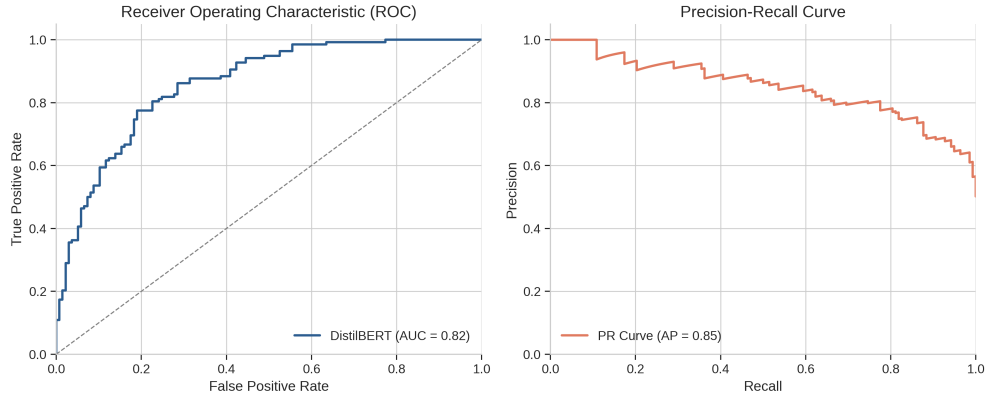


Fig. 5 Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for the DistilBERT classification of implicit coercive control.

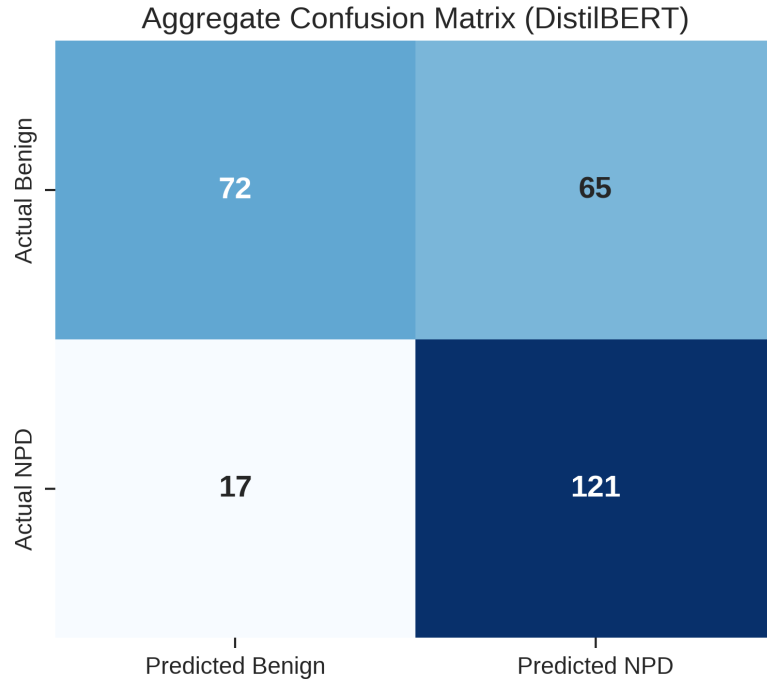


Fig. 6 Confusion matrix detailing the distribution of True Positives, True Negatives, False Positives and False Negatives for the DistilBERT model.

a multi-dimensional framework to evaluate synthetic-data fidelity and utility. The findings demonstrate that synthetic data are not validated by a single similarity metric, but through complementary measures of representation-level discriminability and

cross-domain task transfer. These findings indicate that sequence-aware algorithms can model task-relevant aspects of relational dynamics, provided they are engineered to evaluate longitudinal behavioural clusters rather than isolated syntax.

4.1 The Value of Synthetic Temporal Modelling in Digital Forensics

Bridging the synthetic-to-real transfer gap represents a fundamental shift in how behavioural scientists can approach the study of digital pathology. The multi-dimensional evaluation indicates that artificial generation does not necessarily result in substantial distributional or task-level degradation. Historically, synthetic datasets were identifiable by their repetitive phrasing and lack of emotional depth, rendering them useless for nuanced psychological research. By constraining the Large Language Model strictly within the parameters of the DSM-5 and the Narcissistic Personality Inventory, we forced the algorithm to emulate the pragmatics of human manipulation.

The critical advancement in this framework is the integration of temporal tokens without introducing spurious classification shortcuts. Human psychological abuse is deeply linked to the passage of time. In asynchronous messaging environments, silence is frequently weaponised. Tactics such as stonewalling, calculated unavailability and the strategic withholding of affection rely entirely on chronological pacing to destabilise a victim. By embedding explicit temporal constraints during the generation phase, the resulting dataset successfully captured the alternating cycles of idealisation and devaluation. Methodologically, embedding normative delay tokens into the baseline control dialogues ensured that our models did not succumb to shortcut learning. A classifier trained on metadata presence alone would exhibit a severe performance collapse once those tokens were removed. Instead, our text-only ablation experiment demonstrated that DistilBERT retained a Strong F1-score of 0.71 based purely on linguistic subtext. More importantly, the +0.07 Recall surge observed when reintroducing temporal tokens provides support for the role of temporal structure in the model: while verbal devaluation is recognisable in isolation, capturing the full operational scope of coercive control requires evaluating when a message is sent—and how long silence is weaponised—alongside what is explicitly stated.

In parallel, our synthetic control verification dispels concerns regarding dataset provenance and artificial baseline bias. A common limitation when pairing synthetic data with organic human datasets is that deep learning architectures may inadvertently act as AI-text detectors, separating classes based on underlying language model perplexity rather than target behaviour. However, when evaluated against our purely synthetic normative control set, DistilBERT achieved a comparable F1-score of 0.73. Additionally, our cross-generator validation trials on Llama-3-70B dialogue ($F1 = 0.76$) confirm out-of-domain construct validity. Because the classifier successfully identified implicit devaluation and asynchronous silence in prose generated by an entirely different open-weights architecture, these results provide evidence that performance is not solely dependent on the surface characteristics of a single foundation model.

This methodology provides immediate empirical value. The ‘Forensic Data Gap’ has actively suppressed research into how intimate partner violence and narcissistic

abuse manifest in modern applications. With a validated framework for generating synthetic proxies, researchers can now simulate thousands of longitudinal conversational dyads. This allows for the rapid testing of psychometric detection models without the ethical and legal barriers associated with harvesting organic chat logs. Our evaluation against the UK Legal & Case Vignette Corpus ($N = 40$) directly addresses the Forensic Data Gap. Achieving an F1-score of 0.72 on verified Section 76 Court of Appeal evidence demonstrates that procedural synthetic data generated under strict clinical constraints can serve as generalisable proxies for training models on coercive-control-related discourse. However, we acknowledge that these represent external task-transfer benchmarks rather than direct validations of authentic coercive control, given the domain mismatch between intimate partner dialogues and platform-based or court-admissible texts.

The Semantic Gap in Practice

The performance gap between the transformer architecture and shallow baselines highlights a structural limitation in how current digital safety mechanisms operate. The systemic limitation of the VADER sentiment analysis tool was not a technical anomaly; it was a highly predictable outcome that mirrors the operational shortcomings of basic automated moderation systems. When platforms rely on shallow sentiment tools that consistently miss contextual abuse, they inadvertently reinforce power imbalances. By mathematically scoring manipulative subtext as benign, these algorithms fail to safeguard victims or acknowledge their lived experiences of digital harm.

To understand this failure, one must evaluate the mathematical assumptions underlying lexicon-based architectures. Systems like VADER treat language as an unordered collection of sentiment-weighted tokens. They assign toxicity scores based on the presence of historically aggressive words. However, human coercive control is rarely lexical; it is profoundly contextual. A perpetrator engaging in narcissistic gaslighting typically avoids explicit profanity or standard hate speech. Instead, they weaponise common, everyday vocabulary to manipulate their target.

During our testing phase, when VADER was thresholded to detect negative manipulative intent, it yielded an abysmal Recall of 0.14, actively categorising the vast majority of coercive control as positive or neutral sentiment. This occurs because abusive individuals frequently use faux-polite, overly formal or passive-aggressive vocabulary to maintain plausible deniability. Phrases such as “I am only doing this because I care about your mental health” or “You are clearly confused again, let me help you remember” contain lexically positive tokens like “care,” “health,” and “help.” VADER mathematically registers these words as supportive, entirely missing the predatory intent embedded in the pragmatics of the exchange.

This mechanism actively creates and enforces the “semantic gap.” By failing to recognise the underlying intent of the communication, lexically driven algorithms routinely validate the abuser’s methods. When content moderation algorithms or basic sentiment filters review these interactions, they flag them as benign, leaving victims unprotected. Similarly, our benchmarking of off-the-shelf Llama-3-8B-Instruct ($F1 = 0.57$) demonstrates that unconstrained, zero-shot Large Language Models cannot

autonomously replace domain-specific sequence classifiers. Without fine-tuned attention weights calibrated to relational history and chronemic silence, general-purpose LLMs routinely misclassify faux-polite devaluation as benign dialogue.

4.2 Triage Efficacy and Behavioural Error Analysis

Evaluating the performance of the DistilBERT transformer requires contextualising the metrics within their intended operational environment. While the model exhibited exceptional sequence comprehension, yielding an F1-Score of 0.75, a rigorous examination of the Precision (0.65) and Recall (0.88) trade-off provides deep insight into the boundaries of Artificial Intelligence in understanding human nuance.

The Precision score of 0.65 indicates that the model occasionally misclassified normative relational conflict as abusive behaviour. From a purely statistical standpoint, one might view this as a limitation. However, within the context of cyber-psychology and digital forensics, this metric highlights a fundamental truth: human communication is highly ambiguous. The structural syntax of an emotionally heated argument between two healthy individuals can temporarily mirror the linguistic patterns of devaluation. Algorithms do not experience the social world; they cannot feel the historical interpersonal friction or the mutual understanding that defines a healthy relationship experiencing temporary strain.

Because of this inherent ambiguity, treating artificial intelligence as an autonomous adjudicator of human psychological states is both empirically flawed and ethically dangerous. The transformer’s performance validates its intended application as an investigative digital conversation analysis pipeline. Within a forensic framework, the system is designed to function strictly as a triage tool to assist human investigators, rather than making autonomous clinical diagnoses.

In forensic triage, the absolute priority is the minimisation of False Negatives. Failing to identify a genuine instance of coercive control carries severe safety implications for potential victims. The model’s high Recall rate of 0.88 ensures that the vast majority of manipulative behavioural clusters are captured, flagged and elevated for human review.

However, it is critical to acknowledge the operational cost of the 0.65 Precision score. A roughly 35% false-positive rate means the system occasionally flags highly emotional normative arguments as implicit control. In a real-world digital forensics environment, this false-positive rate could overwhelm human analysts with normative relational disputes. Consequently, future deployments of this framework must incorporate adjustable classification thresholds (as visualised previously in the Precision-Recall curve figure 5), allowing investigative units to calibrate the sensitivity of the model to balance precision and recall based on their specific caseload capacity and operational resources.

An analysis of a live test failure demonstrates why human-in-the-loop oversight remains necessary. As illustrated in Figure 7, the model occasionally struggled to contextualise abstract statements of implicit grandiosity masked as faux-concern. During the testing phase, the model generated a false negative for the utterance: “I took care of it myself; I know how easily overwhelmed you get with these complex tasks.”

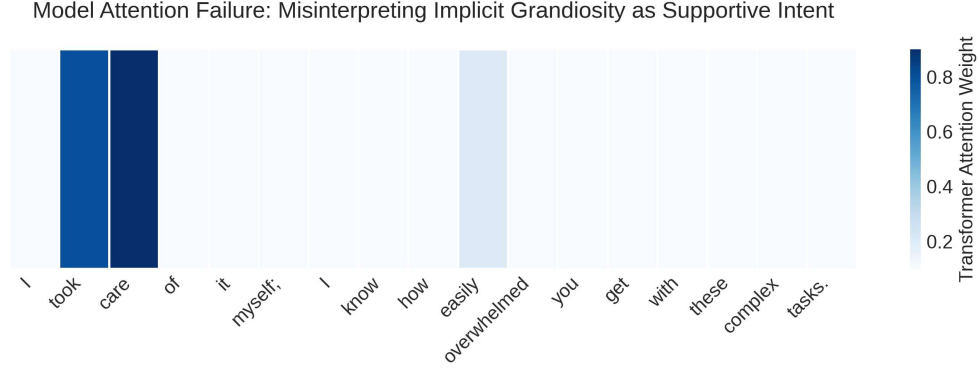


Fig. 7 Attention-map conceptualisation of implicit manipulative words versus structural markers of entitlement. Colour bar intensity (0.2 to 0.8) denotes pooled last-layer self-attention weights directed to the [CLS] classification token.

Within a clinical context, this statement is a textbook manifestation of devaluation and grandiosity, asserting superiority under the guise of assisting a supposedly incapable partner. However, the model classified this input as benign, supportive text. The AI evaluates the lexically positive syntax (“took care of it,” “help”), but it fundamentally misunderstands the condescending social contract that the speaker is violating. It lacks genuine social cognition. This inherent limitation is precisely why we mandated the rigorous human-in-the-loop expert validation phase ($\kappa = 0.84$) during dataset construction; human clinicians intuitively grasp these contextual social violations in ways that computational architectures currently cannot. Consequently, the objective of computational forensic analysis must not be to replace human psychological assessment, but to augment it with highly precise analytical tools capable of mapping longitudinal patterns at scale. By acknowledging the boundaries of synthetic comprehension, researchers can deploy these sequence-aware models responsibly, ensuring that human oversight remains the ultimate authority in the evaluation of digital behaviour. It is critical to note that while our architecture demonstrated strong out-of-sample robustness across verified Section 76 court transcripts ($F1 = 0.72$), this framework represents an investigative decision-support tool, not a clinical diagnostic instrument. Establishing broader external validity across multilingual and culturally divergent organic datasets remains necessary, underscoring why sequence-aware transformers must serve strictly as human-in-the-loop triage aids rather than autonomous adjudicators.

4.3 Practical Implications for Digital Forensics and Law Enforcement

Beyond the theoretical contribution to natural language processing and cyberpsychology, the proposed framework offers direct operational utility for digital forensics units and law enforcement agencies handling domestic abuse and cyber-stalking investigations. Investigators are routinely overwhelmed by vast digital evidence corpora

containing millions of unstructured chat logs, where identifying sustained patterns of coercive control manually is resource-intensive and prone to oversight.

In a future deployment scenario, the fine-tuned sequence-aware transformer could potentially serve as an automated investigative triage layer, allowing forensic analysts to rapidly filter and score multi-message conversation logs. The model’s high recall rate (88%) ensures that subtle manipulative clusters and asynchronous stonewalling tactics are flagged and elevated for human review, effectively mitigating the risk of critical evidence being dismissed due to the “semantic gap” inherent in legacy keyword filters. Furthermore, because the framework relies on privacy-preserving synthetic data generation rather than sensitive real-world victim logs, an approach that has recently proven highly effective for sensitive clinical and prognostic modeling^[29], investigative units can retrain, audit and fine-tune localised models without violating strict statutory data protection regulations such as GDPR.

4.4 Limitations and Future Work

While the proposed methodology successfully models Narcissistic Personality Disorder markers within English-language dyads, several limitations require addressing. First, the expert behavioural assessment ($\kappa = 0.84$) inherently establishes constraint adherence and construct alignment rather than independent ecological validation, as the experts evaluated the presence of the exact clinical criteria used to prompt the generative model. Furthermore, although our synthetic corpus ($N = 275$) and external task-transfer legal benchmark ($N = 40$) establish construct validity, future iterations should scale procedural generation and expert-curated archives to larger datasets (e.g., 1,500–2,000 dyads) to establish wider population-level norms.

Second, the current constraint parameters focus exclusively on a single dimension of the Dark Triad. To establish a comprehensive model of digital abuse, future iterations must scale the generative loop to incorporate the calculated deception characteristic of Machiavellianism and the impulsive antagonism associated with subclinical psychopathy. Additionally, procedural generation via Large Language Models is inherently sensitive to prompt phrasing; minor variations in system instructions or temperature settings can alter the perceived severity and subtlety of the generated dialogue. Future research should systematically audit prompt sensitivity across multiple foundation models.

Additionally, psychological abuse is heavily influenced by cultural and linguistic norms. The weaponisation of silence and the structural syntax of implicit control vary significantly across different conversational cultures; what constitutes an ‘unacceptable delay’ or manipulative ghosting in one demographic may be entirely normative in another. Validating the temporal encoding framework across diverse cultural digital chronemics is necessary to ensure the global applicability of the synthetic data. Subsequent research should prioritise multilingual sequence modelling to test how cultural pragmatics alter the expression and algorithmic detection of coercive control. Finally, while transformer models demonstrate high recall, their occasional failure to contextualise abstract social entitlement highlights the ongoing need for human-in-the-loop oversight in intelligent forensic applications.

Conclusion

This study addresses the methodological challenges restricting the empirical investigation of computer-mediated psychological abuse. By translating clinical diagnostic criteria into machine-enforceable rules and incorporating explicit temporal encoding, we introduced a multi-dimensional evaluation of synthetic-data fidelity and utility. Through bidirectional task transfer (TSTR and RTS) and representation-level similarity (MMD and C2ST), we provide evidence that the synthetic corpus preserves task-relevant discriminative structure across synthetic and real-world data distributions. The external transfer evaluation suggests that researchers can generate meaningful proxies for the Forensic Data Gap without compromising user privacy. Furthermore, the comparative classification trials indicate that shallow, lexically driven algorithms struggle to detect implicit manipulation, highlighting the limitations of shallow, lexically driven approaches for detecting contextual psychological abuse. Detecting the NPD-associated behavioural patterns operationalised in this study benefits from sequence-aware transformer architectures capable of modelling shifting relational context and asynchronous communication. Consequently, whilst not establishing immediate deployment readiness, these results provide preliminary evidence that clinically constrained synthetic dialogues can support the development and benchmarking of sequence-aware forensic triage models.

Data availability

The datasets generated and analysed during the current study, including the synthetic conversational corpus, validation metrics and Python source code, are publicly available in the GitHub repository: <https://github.com/Futuredhruv/Procedural-Coercive-Control-Simulation>. The methodology for this study has been pre-registered with the Open Science Framework (OSF) and can be accessed at: <https://osf.io/ah5rv>.

References

- [1] Paulhus, D. L. & Williams, K. M. The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of Research in Personality* **36**, 556–563 (2002). URL <https://www.sciencedirect.com/science/article/pii/S0092656602005056>.
- [2] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders* 5th edn (American Psychiatric Association, Arlington, VA, 2013). URL <https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596>.
- [3] Arabi, S. *Becoming the Narcissist's Nightmare: How to Devalue and Discard the Narcissist While Supplying Yourself* (CreateSpace Independent Publishing Platform, 2016). URL <https://books.google.co.uk/books?id=jM0YvgAACA AJ>.

- [4] Kircaburun, K., Jonason, P. K. & Griffiths, M. D. The dark tetrad traits and problematic social media use: The mediating role of cyberbullying and cyberstalking. *Personality and Individual Differences* **135**, 264–269 (2018). URL <https://www.sciencedirect.com/science/article/pii/S019188691830429X>.
- [5] Tolmie, J., Smith, R. & Wilson, D. Understanding intimate partner violence: Why coercive control requires a social and systemic entrapment framework. *Violence Against Women* **30**, 54–74 (2024). URL <https://doi.org/10.1177/10778012231205585>.
- [6] Lew, Z., Walther, J. B., Pang, A. & Shin, W. Interactivity in online chat: Conversational contingency and response latency in computer-mediated communication. *Journal of Computer-Mediated Communication* **23**, 201–221 (2018). URL <https://doi.org/10.1093/jcmc/zmy009>.
- [7] Thomas, J. O. & Dubar, R. T. Disappearing in the age of hypervisibility: Definition, context, and perceived psychological consequences of social media ghosting. *Psychology of Popular Media* **10**, 291–302 (2021).
- [8] Craker, N. & March, E. The dark side of facebook®: The dark tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences* **102**, 79–84 (2016). URL <https://www.sciencedirect.com/science/article/pii/S0191886916307930>.
- [9] Waseem, Z. & Hovy, D. Andreas, J., Choi, E. & Lazaridou, A. (eds) *Hateful symbols or hateful people? predictive features for hate speech detection on Twitter*. (eds Andreas, J., Choi, E. & Lazaridou, A.) *Proceedings of the NAACL Student Research Workshop*, 88–93 (Association for Computational Linguistics, San Diego, California, 2016). URL <https://aclanthology.org/N16-2013/>.
- [10] Davidson, T., Warmesley, D., Macy, M. & Weber, I. Automated hate speech detection and the problem of offensive language. *Proc. 11th Int. AAAI Conf. Web Soc. Media (ICWSM)* **11**, 512–515 (2017). URL <https://doi.org/10.1609/icwsm.v11i1.14955>.
- [11] Pluta, A. *et al.* Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others’ pain. *Scientific Reports* **13**, 4127 (2023). URL <https://doi.org/10.1038/s41598-023-31146-1>.
- [12] ElSherief, M. *et al.* Latent hatred: A benchmark for understanding implicit hate speech (2021). URL <https://arxiv.org/abs/2109.05322>. arXiv:2109.05322.
- [13] Wen, J. *et al.* Unveiling the implicit toxicity in large language models (2023). Preprint at <https://arxiv.org/abs/2311.17391>.
- [14] Jurgens, D., Hemphill, L. & Chandrasekharan, E. Korhonen, A., Traum, D. & Màrquez, L. (eds) *A just and comprehensive strategy for using NLP to address*

- online abuse. (eds Korhonen, A., Traum, D. & Màrquez, L.) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3658–3666 (Association for Computational Linguistics, Florence, Italy, 2019). URL <https://aclanthology.org/P19-1357/>.
- [15] Rogers, M. M., Fisher, C., Ali, P., Allmark, P. & Fontes, L. Technology-facilitated abuse in intimate relationships: A scoping review. *Trauma, Violence, ; Abuse* **24**, 2210–2226 (2022). URL <https://doi.org/10.1177/15248380221090218>.
 - [16] Kern, M. L. *et al.* Gaining insights from social media language: Methodologies and challenges. *Psychological Methods* **21**, 507–525 (2016). URL <https://doi.org/10.1037/met0000091>.
 - [17] Chancellor, S. & De Choudhury, M. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine* **3**, 43 (2020). URL <https://doi.org/10.1038/s41746-020-0233-7>.
 - [18] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Burstein, J., Doran, C. & Solorio, T. (eds) *BERT: Pre-training of deep bidirectional transformers for language understanding*. (eds Burstein, J., Doran, C. & Solorio, T.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019). URL <https://aclanthology.org/N19-1423/>.
 - [19] Hossain, M. M., Hossain, M. S., Mridha, M. F., Safran, M. & Alfarhood, S. Multi task opinion enhanced hybrid BERT model for mental health analysis. *Scientific Reports* **15** (2025). URL <https://doi.org/10.1038/s41598-025-86124-6>.
 - [20] Argyle, L. P. *et al.* Out of one, many: Using language models to simulate human samples. *Political Analysis* **31**, 337–351 (2023). URL <http://dx.doi.org/10.1017/pan.2023.2>.
 - [21] Rashkin, H., Smith, E. M., Li, M. & Boureau, Y.-L. Towards empathetic open-domain conversation models: A new benchmark and dataset. *Proc. 57th Annu. Meet. Assoc. Comput. Linguist.* 5370–5381 (2019). URL <https://doi.org/10.18653/v1/P19-1534>.
 - [22] Raskin, R. & Terry, H. A principal-components analysis of the narcissistic personality inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology* **54**, 890–902 (1988). URL <https://psycnet.apa.org/doi/10.1037/0022-3514.54.5.890>.
 - [23] Danescu-Niculescu-Mizil, C. & Lee, L. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs (2011). Preprint at <https://arxiv.org/abs/1106.3077>.

- [24] McCarthy, P. M. & Jarvis, S. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* **42**, 381–392 (2010). URL <https://doi.org/10.3758/BRM.42.2.381>.
- [25] Reimers, N. & Gurevych, I. Inui, K., Jiang, J., Ng, V. & Wan, X. (eds) *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. (eds Inui, K., Jiang, J., Ng, V. & Wan, X.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992 (Association for Computational Linguistics, Hong Kong, China, 2019). URL <https://aclanthology.org/D19-1410/>.
- [26] Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach (2019). Preprint at <https://arxiv.org/abs/1907.11692>.
- [27] Joachims, T. Nédellec, C. & Rouveirol, C. (eds) *Text categorization with support vector machines: Learning with many relevant features*. (eds Nédellec, C. & Rouveirol, C.) *Machine Learning: ECML-98*, 137–142 (Springer Berlin Heidelberg, Berlin, Heidelberg, 1998). URL <https://doi.org/10.1007/BFb0026683>.
- [28] Sanh, V., Debut, L., Chaumond, J. & Wolf, T. Distilbert, a distilled version of BERT: Smaller, faster, cheaper and lighter (2019). Preprint at <https://arxiv.org/abs/1910.01108>.
- [29] Qian, Z. *et al.* Synthetic data for privacy-preserving clinical risk prediction. *Scientific Reports* **14**, 22894 (2024). URL <https://doi.org/10.1038/s41598-024-72894-y>.

Declaration of Generative AI and AI-Assisted Technologies in the Research Process

During the preparation of this work, the authors used OpenAI’s GPT-4 (gpt-4-0613) and Meta’s LLaMA-3 large language models in order to procedurally generate the synthetic conversational datasets and holdout corpora required for methodological testing. These tools were heavily constrained by clinical parameters to simulate behavioural markers. After using these tools, the authors extensively reviewed and validated the generated content via expert behavioural assessment and take full responsibility for the data and methodologies outlined in the published article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Author contributions

Dhruv B. Patel: Conceptualisation, Methodology, Software, Writing - Original draft preparation. **Abhishek Kaushik:** Supervision, Validation, Writing - Review & Editing. **Dr. Anju P. Johnson:** Supervision, Validation, Writing - Review & Editing.

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [LaYexSourcefile.pdf](#)