

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

Dhruv Patel, Abhishek Kaushik and Anju P. Johnson

Appendix A Generative Constraints and System Prompts

To ensure the procedural generation framework remained strictly anchored to clinical psychometrics and avoided producing caricatures of abuse, specific system prompts and temporal constraints were utilised within the Large Language Model (LLM) generation loop.

A.1. Primary System Prompt

The following zero-shot instruction set was provided to the LLM to govern the synthesis of the Narcissistic Personality Disorder (NPD) dataset. The variables for the specific trait and clinical anchor were injected dynamically per generation batch:

"You are a computational psycholinguist building a dataset of implicit coercive control. Your task is to generate distinct, multi-turn text message exchanges demonstrating the Dark Triad trait: [INSERT TRAIT].

CRITICAL CONSTRAINTS:

1. **AVOID CARICATURES:** Do not use overt hostility, shouting or obvious insults. The abuse must be highly subtle, passive-aggressive or shrouded in faux-concern (e.g., 'I only say this because I worry about your mental state').
2. **AMBIGUITY:** To an outside observer, the text should look almost normal. The manipulation is purely subtextual.
3. **TEMPORAL TOKENS:** You **MUST** separate messages with explicit time-delay tokens to simulate asynchronous control (e.g., ignoring someone for hours). Use formatting like `¡DELAY: 2_MINS¿` or `¡DELAY: 14_HOURS¿`.
4. **FORMAT:** Output exactly the requested number of distinct exchanges. Separate each exchange with '—'. Do not include conversational filler.

ANCHOR BEHAVIOUR: [INSERT DSM-5 / NPI ANCHOR]"

A.2. Examples of Clinical Anchors

During Phase 3 of the procedural generation (Devaluation and Implicit Control), specific clinical anchors were inserted into the prompt to isolate distinct behavioural clusters. Examples include:

- **Implicit Entitlement:** Complaining about a mild inconvenience as if it is a severe injustice, expecting the partner to fix it.
- **Calculated Devaluation:** Giving a backhanded compliment that targets the partner's known insecurity.
- **Withholding (Lack of Empathy):** Ignoring a direct, emotional question, implementing a `<DELAY: 8_HOURS>` and replying with something entirely unrelated.