

# LURE: a Phase-Labelled Synthetic Corpus of Online-Grooming Exchanges

Emanuele Florindi, *Member, IEEE*, Valentina Franzoni, *Senior Member, IEEE*, Ludovico Guercio, Alfredo Milani, *Senior Member, IEEE*, and Mattia Polticchia, *Member, IEEE*

**Abstract**—Phase-annotated conversational data is essential for studying the progression of online grooming of minors and for developing early warning tools. However, the only publicly available source (i.e., Perverted-Justice archive, which contains a subset coded using ChatCoder) is small, outdated, and frozen. We present a synthetic corpus of groomer-victim exchanges, generated using a retrieval-augmented pipeline running on institutional hardware. This pipeline uses an openly available, eight-billion-parameter, refusal-ablated language model. The corpus was created using the five stages of the responsible-generation protocol for synthetic datasets: ontology-driven design based on the ChatCoder luring-communication scheme (eight communicative tags mapped onto four grooming phases); multi-model prompt testing; human-in-the-loop supervision; semantic and feature-level validation; and modality-aware documentation. The released corpus contains 2,121 unique exchanges across seven tags and four phases, each accompanied by a Plutchik emotion field. Validation quantifies the integrity and lexical profile of the corpus, as well as the reproduction of real seed lines, in-domain label recoverability (tag accuracy: 0.85; phase accuracy: 0.93) and its transfer to real ChatCoder offender lines, which were coded automatically at the eight-tag level (31,640 lines) and by human coders at the three-construct level (10,876 posts). Accuracy was measured for lexical and sentence-embedding features, while a real-versus-synthetic discriminator was able to separate the two domains with 91% accuracy. The scope of the corpus includes controlled, phase-conditioned experimentation, pipeline development and augmentation alongside real data and awareness material. The corpus, its datasheet and the audit code are released with vetted access.

**Index Terms**—Child safety, corpus validation, digital forensics, large language models, online grooming, responsible AI, retrieval-augmented generation, synthetic data generation.

## I. INTRODUCTION

ONLINE grooming is the gradual process by which an adult establishes trust and emotional dependence with a child in order to facilitate sexual contact or obtain material [1], [2]. This phenomenon has grown along with social networks, messaging apps, and online gaming. The average

age of victims is dropping, with pre-adolescents being the most exposed group [3]–[6]. Its handling is being debated at the European level, where institutional client-side scanning has been assessed as risky [7] and is being reconsidered [8]; scanning by law enforcement based on well-founded suspicion or via a tool chosen and managed by the user remains the more proportionate approach. Both require models that are trained using phase-annotated conversational data. Grooming progresses through identifiable stages of communication (contact, building trust, isolation from protective adults, desensitisation to sexual content and approach), which can be studied and recognised in chat text [9], [10]. The computational work on grooming relies on a single public source: the Perverted-Justice (PJ) archive of conversations between convicted offenders and adult volunteer decoys. This archive was used in the PAN 2012 shared task and in most subsequent detection studies [11]–[13]. The archive is in English, and the only portion with message-level communicative annotation is its ChatCoder subset (288 transcripts coded automatically with a 475-phrase dictionary and 55 coded by trained human coders) [10]. Data protection laws mean that most research groups cannot access new real data on sexual communication with children [14].

The use of Large Language Models (LLMs) to generate labelled text is an established response to data scarcity in sensitive domains [15], [16], and grooming is a natural target: the label of interest (i.e., the grooming phase) is imposed at generation time and no minors are involved. The domain also encompasses the questions that responsible generation must answer. Aligned models refuse the task, while refusal-ablated open models comply but offer no domain knowledge. The most informative phases are the most explicit, and every artefact is dual-use. The protocol of [17] organises the generation of synthetic datasets under these exact conditions into five stages: ontology-driven design; multi-model prompt testing; iterative human-in-the-loop supervision; semantic and feature-level validation; and modality-aware documentation. This process is illustrated using a Plutchik-based emotion corpus and has since been applied to the PLUMED corpus [18]. This paper uses the same protocol for online grooming and provides details of each stage, along with the measurements needed to make the resulting resource usable by others. Figure 1 summarises the pipeline and the five stages.

Our contributions are:

- This is a retrieval-augmented generation (RAG) pipeline for phase- and tag-labelled grooming exchanges. It runs on institutional infrastructure and uses openly available,

Preprint. This work received no project funding; it was supported by the EMORE Lab, University of Perugia, Italy.

V. Franzoni and L. Guercio are with the EMORE Lab, Department of Mathematics and Computer Science, University of Perugia, 06123 Perugia, Italy. V. Franzoni is also with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (corresponding author, e-mail: valentina.franzoni@unipg.it).

E. Florindi and M. Polticchia are with the LESDIT PhD School, Learning Science and Digital Technologies, University of Modena and Reggio Emilia, Reggio Emilia, Italy.

A. Milani is with the Department of Human Sciences, Link Campus University, Rome, Italy.

ORCID: E. Florindi 0009-0002-5420-0326; V. Franzoni 0000-0002-2972-7188; A. Milani 0000-0003-4534-1805; M. Polticchia 0009-0005-6175-9832.

refusal-ablated checkpoints, which are served by Ollama and orchestrated with LangChain. The pipeline is anchored to curated, de-identified seed documents, which are built from the ChatCoder-coded portion of PJ (Sections III–IV).

- A corpus of 2,121 unique groomer-victim exchanges (LURE v1.1, *Luring-theory Utterances for Research on grooming Exchanges*) has been released. Each exchange carries a ChatCoder communicative tag, the corresponding grooming phase, and a Plutchik emotion field. A datasheet is also available [19] and flagged v1.2 (Section VI).
- A validation suite covering integrity, lexical profile, seed-line reproduction and in-domain label recoverability. External validation is also provided against real ChatCoder data at two annotation levels: automatic eight-tag coding (silver) and human three-construct coding (gold). Cross-domain transfer is covered by lexical and sentence-embedding features, as well as a real versus synthetic discriminator. (Section VII).

These measurements will provide a documented scope of use and protocol refinements for future synthetic corpora in sensitive domains. They will also inform the governance framework for producing and sharing such data. (Sections VIII–X).

## II. RELATED WORK

Research on the detection of online grooming is shaped by the tension between the need for early-warning models and the limited availability of up-to-date, ethically shareable conversational data. While cybersecurity work has focused on preserving access and channels (e.g., age and ageing in biometrics [20], encrypted-traffic analysis [21], intrusion detection [22]), while tools for minor protection recognise abusive content once abuse has begun [23], [24]; grooming mainly takes place on social networks and in online games, where remote interaction enables the crime [6].

### A. Grooming and abuse detection in text

Automated grooming detection has largely been approached as a form of supervised text classification using PJ-derived corpora, incorporating lexical, stylometric and, more recently, Transformer features [10]–[12], [25]. Communication scholars proposed stage models of luring [9]; the ChatCoder project categorised them into eight utterance-level categories and created a coding dictionary. This was applied automatically to 288 PJ transcripts and, for a smaller set, by trained human coders at the level of the theory’s constructs [10]. More recent work has shifted towards early detection settings and privacy-preserving training [13], [26], and has begun to explore the use of LLMs as estimators of grooming risk [27]. Cyberbullying and the detection of predatory behaviour in gaming chats share the same data constraints [28], [29], and grooming through gaming platforms follows recognisable, platform-specific patterns [30], [31]. Discourse analysis describes the language used for grooming in computer-mediated settings [32], [33]. The corpus presented here adds a phase- and tag-labelled synthetic resource documented against the ChatCoder material.

Three studies delimit the space in which LURE sits, and a direct comparison clarifies what it contributes. The survey of Borj *et al.* [12] maps the field from the data side: it reviews the corpora on which our grooming detection study is built and the techniques applied to them, and documents that the available resources carry a conversation- or author-level decision (i.e., predator or not) over PJ, so that the communicative progression inside a conversation remains outside the label space. LURE is positioned at that point: it supplies the message-level layer the survey reports as scarce, with a tag and a phase on every exchange, and it reports where its distributions sit with respect to the real ChatCoder coding. Cheong *et al.* [28] work with real chats from a children’s social game, which gives their study the platform realism and the age profile that PJ lacks; those logs were obtained under an agreement with the operator and stay with the research team, which is the condition that makes such corpora difficult to reuse. Besides not being labeled for grooming content, such data should be a good starting point for children chats on games. LURE addresses the same reuse problem from the other side: the exchanges are generated rather than collected, so the resource can be shared under a data-use agreement, and the external validation in Section VII-C states how far the generated text stands from the real material. Murnion *et al.* [29] assemble 126,091 in-game messages from 14,930 match replays and score them for cyberbullying, demonstrating the scale that automatic harvesting of a public game channel reaches on a harm that is observable in open chat; grooming is pursued in private channels and is identified only after a judicial process, which is the reason its corpora are counted in hundreds of conversations. LURE offers a complementary route to volume for this second setting, with generation under human supervision and a release documented for controlled benchmarking, focusing on grooming instead of cyberbullying.

### B. Responsible generation of synthetic corpora

Data generation driven by LLM with retrieval augmentation is a standard tool in low-resource and privacy-sensitive domains [15], [16], [34]. The properties on which a synthetic corpus must be measured are identified by surveys. These properties include diversity relative to the generator’s prompt, distributional fidelity to the target phenomenon, and transfer of what is learned on synthetic text to real text [35], [36]. Here, retrieval is used to guide creative generation towards real examples, adding the requirement to measure the reproduction of retrieved material. Data statements and datasheets provide the documentation framework [19], [37]; the followed protocol [17] organises these requirements into five stages, which are reflected in the structure of Sections III–VII.

### C. Emotion annotation

Grooming is an emotional process, and affective computing provides text-based methods for recognising emotions [38], [39], built on categorical models such as Ekman’s basic emotions [40] or Plutchik’s wheel [41], and on Plutchik-aligned resources such as the NRC lexicon [42]. The corpus contains a Plutchik emotional field for each exchange, which was

produced using a transformer classifier that was transferred and learnt using a Plutchik-based corpus [18]. The emotional field is part of the released schema and will be analysed in other studies.

### III. STAGE 1: ONTOLOGY-DRIVEN DESIGN

#### A. Source material

The PJ archive contains 622 one-to-one conversations between offenders and adult decoys posing as minors on platforms such as Yahoo and MeetMe. The community copy used here has had the usernames, timestamps, and moderator comments removed. These conversations took place between 2004 and 2016 [25]. The ChatCoder project distilled a luring-communication scheme with three theoretical constructs from this archive: deceptive trust development, grooming, (communicative desensitisation and reframing), and isolation/approach. Eight utterance-level tags were also identified (see Table I) [9], [10]. Two coded resources distributed with the project serve as real references in Section VII-C: 288 transcripts that were coded automatically by the ChatCoder dictionary at the eight-tag level (177,366 offender lines, of which 35,392 contain a dictionary phrase and 31,640 carry exactly one tag) and 55 transcripts coded by trained human coders at the construct level (11,204 offender posts with a construct code, of which the 10,876 with at least two word tokens are used in the experiments). The dictionary coding is keyword-based. For instance, an offender username containing the substring 'bra' triggered the desensitisation tag on every line of that transcript. The dictionary coding is treated as silver, while the human coding is treated as gold. Each is used at its own granularity.

#### B. Phase scheme and tag-to-phase mapping

The design stage of the protocol establishes the label ontology prior to the creation of any prompts. Stage models of sexual abuse and online luring [1], [43], [44] describe a progression involving contact, friendship, risk assessment, special bonding, trust-testing missions, secrecy and luring, which exploits classic persuasion principles [45]. Generation is organised in five phase contexts: *Targeting and Gaining Trust*, *Fulfilling Needs*, *Isolation*, *Sexualizing the Relationship and Abuse* and *Control*. The released label set consists of four primary phases, with every tag mapping deterministically to one of these phases (see Table I). The control context, whose tags all belong to the other phases, is folded into them. The isolation phase precedes explicit sexual content and is defined by emotional rather than sexual material. This makes it the operational focus of early-warning research.

#### C. Emotion field

Each exchange carries a Plutchik emotion label (i.e., one of the eight primary emotions or *neutral*) assigned in a separate annotation process using the `all- Roberta-large-v1` model, which was trained on a Plutchik-grounded textual corpus [18], [46]. This field is released as a silver-standard annotation and is documented as such in the datasheet.

### IV. STAGE 2: MULTI-MODEL PROMPT TESTING

#### A. Infrastructure and models

Generation ran on the *UniNuvola* computing service of the University of Perugia, on a node with an AMD EPYC 9334 CPU (2 cores allocated), 32 GB of RAM and an NVIDIA A30 GPU (24 GB), with models served by Ollama [47] and orchestrated with LangChain [48] in Python; prompts and outputs stayed on institutional infrastructure. Three openly distributed refusal-ablated checkpoints were tested: *Llama-3.2-3B-Instruct-abliterated* [49], *Meta-Llama-3.1-8B-Instruct-abliterated* (Q4\_K\_M) [50] and *DeepSeek-R1-Distill-Llama-8B-Abliterated* (IQ4\_XS) [51]. Abliteration removes the refusal direction from the residual stream [52], so that these checkpoints comply with the task; domain knowledge about grooming is supplied through the prompt and the retrieval base.

#### B. Knowledge base

The RAG knowledge base consists of five phase documents (*targeting\_and\_gaining\_trust*, *fulfilling\_needs*, *isolation*, *sexualize\_the\_relationship\_and\_abuse*, *control*) with a fixed structure: a natural-language description of the phase and its real-world warning signs; per-tag sections with the tag definition and its manipulative purpose; and de-identified offender lines selected manually from the ChatCoder-coded PJ portion (337 seed lines of at least three tokens in total). Retrieval for each call is restricted to the document matching the target (phase, tag) pair, following the standard embed-index-retrieve flow.

#### C. Prompt architecture

Each call concatenates three components. The *system prompt* provides in-context domain grounding in the form of the two-stage luring model and its six persuasion principles, the seven-stage online sequence, gaming-specific patterns and the phase/tag scheme with its associated output conventions. The *RAG prompt* instructs the model to use the retrieved examples as inspiration for creating semantically similar messages that differ in their surface form. The *question prompt* specifies the task. In dialogue mode, it also receives the accumulated conversation, enabling each phase to build on the previous one. Components that address residual refusal (e.g., the purpose statement and role framing), are described at this level and kept out of the release (see Section X).

#### D. Generation strategies tested

- 1) **Direct dialogue.** A single prompt requests a complete grooming conversation, divided into phases. The outputs were brief and only partially covered the grooming pattern; the 3B model declined the task in most runs (e.g. “I can’t provide information or guidance on illegal or harmful activities, especially those that could endanger a child”).
- 2) **Phase-by-phase dialogue.** The conversation is generated one phase at a time, with feedback on previous phases. Phase adherence and coherence improved, but residual behaviours included abrupt topic shifts, the

TABLE I  
ANNOTATION SCHEME: CHATCODER COMMUNICATIVE TAGS, LURING-THEORY CONSTRUCT, GENERATION PHASE CONTEXT AND RELEASED PHASE LABEL.

| Tag                           | Construct          | Generation context          | Released phase               |
|-------------------------------|--------------------|-----------------------------|------------------------------|
| Activities                    | deceptive trust    | Targeting & Gaining Trust   | Targeting & Gaining Trust    |
| Personal Information          | deceptive trust    | Targeting & Gaining Trust   | Targeting & Gaining Trust    |
| Relationship                  | deceptive trust    | Targeting; Fulfilling Needs | (merged, see Sec. VI)        |
| Compliment                    | deceptive trust    | Fulfilling Needs            | Fulfilling Needs             |
| Reframing                     | grooming           | Isolation                   | Isolation                    |
| Communicative Desensitization | grooming           | Isolation                   | Isolation                    |
| Isolation                     | isolation/approach | Isolation; Control          | Isolation                    |
| Approach                      | isolation/approach | Sexualizing; Control        | Sexualizing the Relationship |

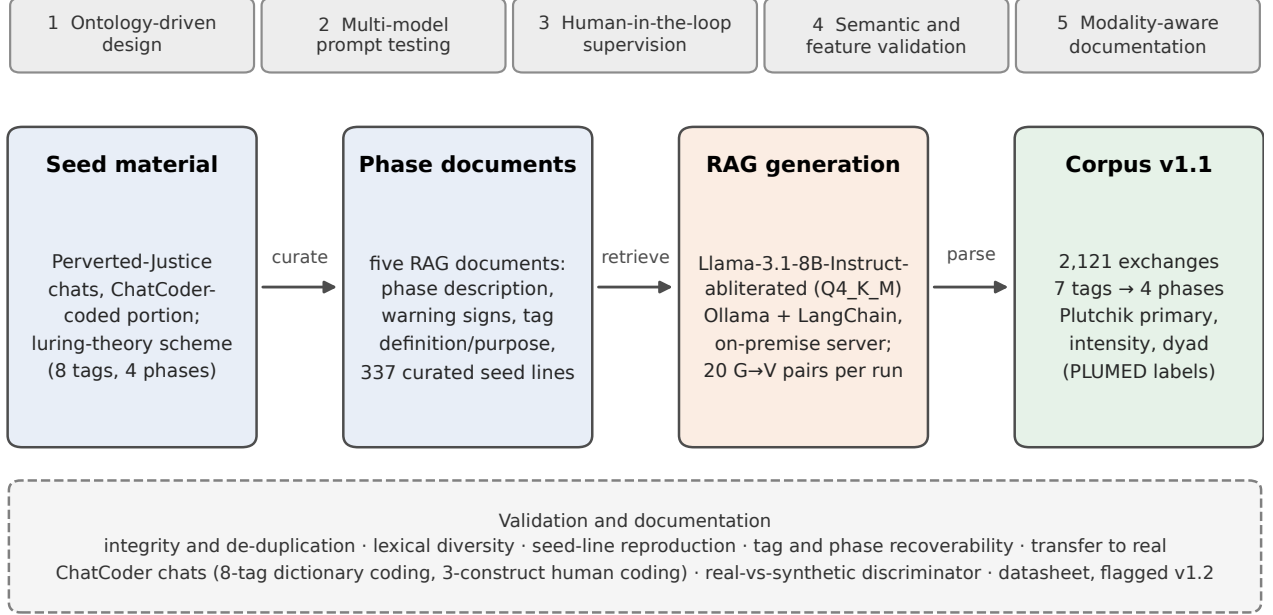


Fig. 1. The generation pipeline and its role in the responsible generation protocol [17] (top strip). Perverted-Justice material coded by ChatCoder [10] is curated into five-phase documents that anchor a locally hosted RAG chain. Example-driven, per-phase, per-tag generation emits labelled groomer–victim pairs that are parsed into the corpus and annotated with the Plutchik emotion field. The lower strip lists the validation and documentation steps of Section VII.

recurrence of covered topics, limited variation in the victim’s voice and sparse output during the explicit phases. The RAG-optimised checkpoints (Llama-3.2 and DeepSeek-R1) tended to produce retrieved seed lines with few modifications in one archived run, six of the 38 generated messages were seed lines. *Llama-3.1-8B-Instruct-abliterated* produced the most stable output and was retained.

- 3) **Example-driven per-tag pairs (adopted).** The model produces 20 groomer-to-victim message pairs in machine-readable format `message|tag|phase` per run for a given tag within a given phase. Retrieval is restricted to the matching document and the corpus is labelled by construction. The task portion of the question prompt reads:

```
Generate 20 samples related to Tag {tag}
for a chat between groomer and victim in an
online platform. For each sample, generate
a message from the groomer, and the reply
from victim in the following format:
groomer message|tag|phase
```

TABLE II  
COMPARISON OF THE THREE GENERATION STRATEGIES ACROSS THE EVALUATED LOCALLY HOSTED REFUSAL-ABLATED MODELS.

| Strategy                                   | Observed quality                           | Residual behaviours                                                           |
|--------------------------------------------|--------------------------------------------|-------------------------------------------------------------------------------|
| S1: direct dialogue                        | Short, partial grooming patterns           | Refusal (3B model); no phase structure                                        |
| S2: phase-by-phase dialogue                | Coherent, phase-adherent (RAG anchoring)   | Topic shifts; repetition; flat victim voice; sparse output on explicit phases |
| S3: example-driven per-tag pairs (adopted) | Phase-consistent, labelled by construction | Limited lexical variation; seed-line reproduction in RAG-optimised models     |

```
victim message|tag|phase
Retrieve context only from the document
that corresponds to phase "{phase}" and tag
"{tag}". [...]
```

Table II summarises the three strategies. Eight per-tag output files were produced and parsed into exchanges, each consisting of one groomer message and its victim reply.

TABLE III  
COMPOSITION OF THE CORPUS BY PHASE AND TAG: EXCHANGES AS GENERATED (v1.0) AND AS RELEASED (v1.1).

| Phase                                            | Tag             | v1.0                  | v1.1         |
|--------------------------------------------------|-----------------|-----------------------|--------------|
| Targeting/Gaining Trust                          | Activities      | 316                   | 307          |
| Targeting/Gaining Trust                          | Personal Info   | 357                   | 357          |
| Targeting/Gaining Trust                          | Relationship    | 330                   | —            |
| Fulfilling Needs                                 | Compliment      | 351                   | 315          |
| Isolation                                        | Reframing       | 194                   | 158          |
| Isolation                                        | Isolation       | 365                   | 355          |
| Isolation                                        | Desensitization | 358                   | 356          |
| Sexualizing the Relationship                     | Approach        | 308                   | 273          |
| Targeting / Fulfilling / Isolation / Sexualizing |                 | 664 / 315 / 869 / 273 |              |
| <b>Total</b>                                     |                 | <b>2,579</b>          | <b>2,121</b> |

### V. STAGE 3: HUMAN-IN-THE-LOOP SUPERVISION

Human supervision was introduced at four points in the pipeline:

- 1) *Seed curation*: the seed lines were selected manually from the ChatCoder-coded section and organised by phase and tag.
- 2) *Document engineering*: early versions of the phase documents only retrieved basic example lists; adding explanatory text at phase and tag level improved the relevance of the retrieved context.
- 3) *Prompt refinement*: the system and question prompts evolved across test runs to fit the models’ context budget while retaining the domain description.
- 4) *Output review*: approximately one quarter of the generated samples were inspected for message plausibility and label consistency. Disagreements were resolved by majority vote among four researchers. The strategy comparison in Table II is based on this inspection.

The release records show where supervision was applied, as well as where sampling was used instead of an exhaustive approach.

### VI. THE RELEASED LURE CORPUS

#### A. Versions

**v1.0** is the corpus as generated and parsed: 2,579 exchanges over eight tag files. Integrity checking (Section VII-A) established that the *Relationship* output file had the same content as the *Compliment* file, and identified 128 further exact duplicates. **v1.1**, the released version, keeps one copy of every exchange: 2,121 unique exchanges over seven tags and four phases (Table III). **v1.2** adds two Boolean columns: *seed\_verbatim* (the groomer message equals a seed line after normalisation; 151 exchanges) and *seed\_4gram\_full* (every 4-gram of the groomer message occurs in the seed set; 162 exchanges; union 167), so that users can select the subset in which no seed line is reproduced.

#### B. Schema and documentation

Each row carries *exchange\_id*; *sentence* (the exchange as “G: ... — V: ...”); *groomer\_text*, *victim\_text*; *tag*, *tag\_label*; *phase*,

TABLE IV  
GLOBAL STATISTICS OF THE CORPUS AS GENERATED (v1.0) AND AS RELEASED (v1.1). DUPLICATES ARE COUNTED ON THE EXCHANGE STRING AFTER LOWER-CASING AND PUNCTUATION REMOVAL; DISTINCT-*n* IS THE RATIO OF UNIQUE TO TOTAL *n*-GRAMS; LENGTHS ARE WORD TOKENS.

| Statistic                                           | v1.0             | v1.1        |
|-----------------------------------------------------|------------------|-------------|
| Exchanges (groomer–victim pairs)                    | 2,579            | 2,121       |
| Tags / phases                                       | 8 / 4            | 7 / 4       |
| Exact duplicate exchanges                           | 458              | 0           |
| Near-duplicate exchanges (normalised)               | 466              | 8           |
| Tokens (exchange level)                             | 40,657           | 37,039      |
| Vocabulary (word types)                             | 1,971            | 1,971       |
| Type–token ratio                                    | 0.043            | 0.053       |
| Distinct-2 / distinct-3                             | 0.25 / 0.45      | 0.30 / 0.56 |
| Mean length groomer / victim                        | 8.6 / 7.1        | 8.6 / 6.9   |
| Exchanges containing emoji (%)                      | 82               | 80          |
| Groomer messages identical to a seed line (%)       | —                | 7.1         |
| Seed material: PJ conversations / ChatCoder phrases | 622 / 475        |             |
| Generation batch size                               | 20 pairs per run |             |

TABLE V  
DISTRIBUTION OF THE PLUTCHIK EMOTION FIELD (EIGHT PRIMARY EMOTIONS PLUS NEUTRAL) IN v1.0 AND v1.1.

| Emotion      | v1.0         | v1.1         | v1.1 (%)     |
|--------------|--------------|--------------|--------------|
| joy          | 659          | 468          | 22.1         |
| trust        | 546          | 405          | 19.1         |
| sadness      | 441          | 404          | 19.0         |
| anticipation | 383          | 340          | 16.0         |
| surprise     | 196          | 172          | 8.1          |
| neutral      | 155          | 151          | 7.1          |
| fear         | 110          | 101          | 4.8          |
| anger        | 56           | 51           | 2.4          |
| disgust      | 33           | 29           | 1.4          |
| <b>Total</b> | <b>2,579</b> | <b>2,121</b> | <b>100.0</b> |

*phase\_label*; *emotion*; token counts; and *source\_id* (originating per-tag file and line). Table IV gives the global statistics, Table V the distribution of the emotion field, and Table VI representative groomer messages curated for responsible disclosure. The datasheet [19] records provenance (fully synthetic; RAG-anchored generation with manually curated seed examples), generator (*Llama-3.1-8B-Instruct-abliterated*), language (English), the supervision points of Section V, label status (tags and phases by construction; emotion as silver annotation), the validation figures of Section VII, known biases (seed-selection subjectivity, generator-distribution artefacts, simplified victim voice), the withheld components and the access regime.

### VII. STAGE 4: VALIDATION

All experiments use TF-IDF over word unigrams and bigrams with class-balanced multinomial logistic regression (lexical), and mean-pooled frozen all-roberta-large-v1 (1,024 dimensions) with the same classifier (semantic), where stated. In-domain figures are based on 5-fold stratified cross-validation (mean  $\pm$  s.d.); real-data folds are grouped by transcript.

TABLE VI  
REPRESENTATIVE SYNTHETIC GROOMER MESSAGES WITH TAG AND PHASE, CURATED FOR RESPONSIBLE DISCLOSURE.

| Generated message                                               | Tag                           | Phase                        |
|-----------------------------------------------------------------|-------------------------------|------------------------------|
| "That's understandable. Do you have any favorite games?"        | Activities                    | Targeting and Gaining Trust  |
| "Are your parents divorced?"                                    | Personal Information          | Targeting and Gaining Trust  |
| "Your skills are insane! You're the best gamer I've ever seen!" | Compliment                    | Fulfilling Needs             |
| "I can be your teacher, and we can learn together"              | Reframing                     | Isolation                    |
| "Don't worry about your mom's calls, I'm here for you"          | Isolation                     | Isolation                    |
| "Dont u think 18+ games r more fun?"                            | Communicative Desensitization | Isolation                    |
| "Can we meet in a private room?"                                | Approach                      | Sexualizing the Relationship |

#### A. Integrity, lexical profile and seed reproduction

**Integrity.** A byte-wise comparison of the eight raw output files, along with normalised matching of exchanges, identified the Relationship/Compliment duplication (330 exchanges), as well as 128 further exact duplicates. v1.1 retains one copy of each. Within v1.1, 209 groomer messages and 313 victim replies occur in more than one exchange with a different partner's response. Both are retained because the unit of the corpus is the exchange.

**Lexical profile.** The released corpus contains 1,971 word types across 37,039 tokens, giving a type-to-token ratio of 0.053. The distinct 1/2/3 ratios are 0.05, 0.30 and 0.56 (see Fig. 2b). On average, groomer messages contain 8.6 tokens, while victim replies contain 6.9. Emojis are present in at least one of the messages in 80% of exchanges. The ten most frequent groomer openings ('Have you ever', 'Do you have', 'Do you like', 'What's your', 'You have such', etc.) account for 470 out of 2,121 messages, which is a characteristic of the generator's register.

**Seed-line reproduction.** After normalisation, 151 groomer messages (7.1%) equal a seed line against the 337 seed lines. Of these, 162 (7.6%) consist entirely of seed 4-grams and 18.5% share at least one 4-gram with a seed line (mean 4-gram containment 0.11). By tag (Fig. 2c), 19% of Communicative Desensitization, 11% of Isolation and 10% of Approach messages equal a seed line, 6% of Reframing, 2% of Compliment, and none of Activities or Personal Information. Victim replies contain no seed line. Against the ChatCoder dictionary reconstructed from the coded transcripts (299 of the 475 phrases were recoverable), 53% of synthetic groomer messages contain a dictionary phrase, compared with 20% of real offender lines; where a phrase is present, its ChatCoder category coincides with the assigned tag in 35% of cases. All exchanges reproducing a seed line are flagged in v1.2.

#### B. In-domain label recoverability

The label recoverability measure indicates how consistently tags and phases are expressed in the corpus text (Table VII, Fig. 3). The lexical model achieves an accuracy of  $0.848 \pm 0.013$  for the seven tags and  $0.925 \pm 0.009$  for the four

TABLE VII  
IN-DOMAIN RECOVERABILITY OF TAGS AND PHASES FROM TEXT (STRATIFIED 5-FOLD, MEAN  $\pm$  S.D.).

| Setting                                   | Accuracy          | Macro-F1          |
|-------------------------------------------|-------------------|-------------------|
| 7 tags, exchange (G+V), lexical           | $0.848 \pm 0.013$ | $0.844 \pm 0.013$ |
| 7 tags, groomer turn only, lexical        | $0.788 \pm 0.019$ | $0.783 \pm 0.021$ |
| 7 tags, victim turn only, lexical         | $0.749 \pm 0.026$ | $0.745 \pm 0.026$ |
| 7 tags, groomer turn, semantic            | $0.768 \pm 0.019$ | 0.748             |
| 4 phases, exchange (G+V), lexical         | $0.925 \pm 0.009$ | $0.919 \pm 0.010$ |
| 4 phases, groomer turn only, lexical      | $0.880 \pm 0.016$ | $0.869 \pm 0.018$ |
| 4 phases, groomer turn, semantic          | $0.829 \pm 0.007$ | 0.819             |
| 4 phases, exchange, BERT fine-tuned       | $0.918 \pm 0.014$ | $0.907 \pm 0.013$ |
| 8 tags, v1.0, random folds, lexical       | $0.660 \pm 0.007$ | $0.671 \pm 0.010$ |
| 8 tags, v1.0, duplicates grouped, lexical | $0.735 \pm 0.011$ | $0.732 \pm 0.012$ |
| Majority class (7 tags / 4 phases)        | 0.168 / 0.410     | —                 |

phases when the entire exchange is considered. Groomer turns alone achieve an accuracy of 0.788/0.880, while victim turns alone achieve an accuracy of 0.749. A fine-tuned bert-base-uncased (max length 64, batch 16, AdamW  $2 \times 10^{-5}$ , 3 epochs) on the same phase folds reaches  $0.918 \pm 0.014$ , with Isolation recall 0.890 and precision 0.928; the semantic model reaches  $0.829 \pm 0.007$  on phases and  $0.768 \pm 0.019$  on tags. Here, lexical and transformer models are equivalent, which is consistent with the idea that per-tag generation produces text that is characteristic of each label in terms of its lexical content. The majority-class references are 0.168 (tags) and 0.410 (phases). The v1.0 rows document the effect of the duplicated file: the eight-tag accuracy is 0.660 with random folds and 0.735 with duplicate-grouped folds; the two identical tags remain merged in both cases.

#### C. External validation against real ChatCoder data

Real lines were lowercased, and the ChatCoder netspeak translation was partially reverted (e.g. the 'question' suffixes were restored to question marks and 'laugh out loud' to 'lol'). Synthetic groomer messages were lowercased and stripped of emojis. The offender/groomer text was compared; the PJ victim side comprised adult decoys. For the semantic experiments, the real silver set was capped at 800 lines per tag (5,230 lines), while the gold set was sampled to comprise 4,000 posts.

**Distributional distance.** The Jensen–Shannon divergence between the unigram distributions of the synthetic and real offender lines per tag ranges from 0.49 bits (Isolation and Communicative Desensitisation) to 0.59 bits (Activities). The divergence between two random halves of the real data is 0.13, and between two halves of the synthetic data it is 0.17 (Fig. 4a). When pooled over tags, the divergence is 0.45. Between 26% (Activities) and 85% (Communicative Desensitisation) of a tag's synthetic vocabulary occurs in the real lines of the same tag, and the mean lengths are comparable (7.8–9.9 vs. 6.4–9.3 tokens).

**Cross-domain transfer.** Table VIII and Fig. 4b report classifiers trained on one domain and tested on the other. With lexical features, a classifier trained on synthetic groomer messages achieves an accuracy of 0.348 on the seven real tags and 0.427 on the four phases (real majority: 0.394 and 0.524; in-domain real: 0.953 and 0.961, with high figures

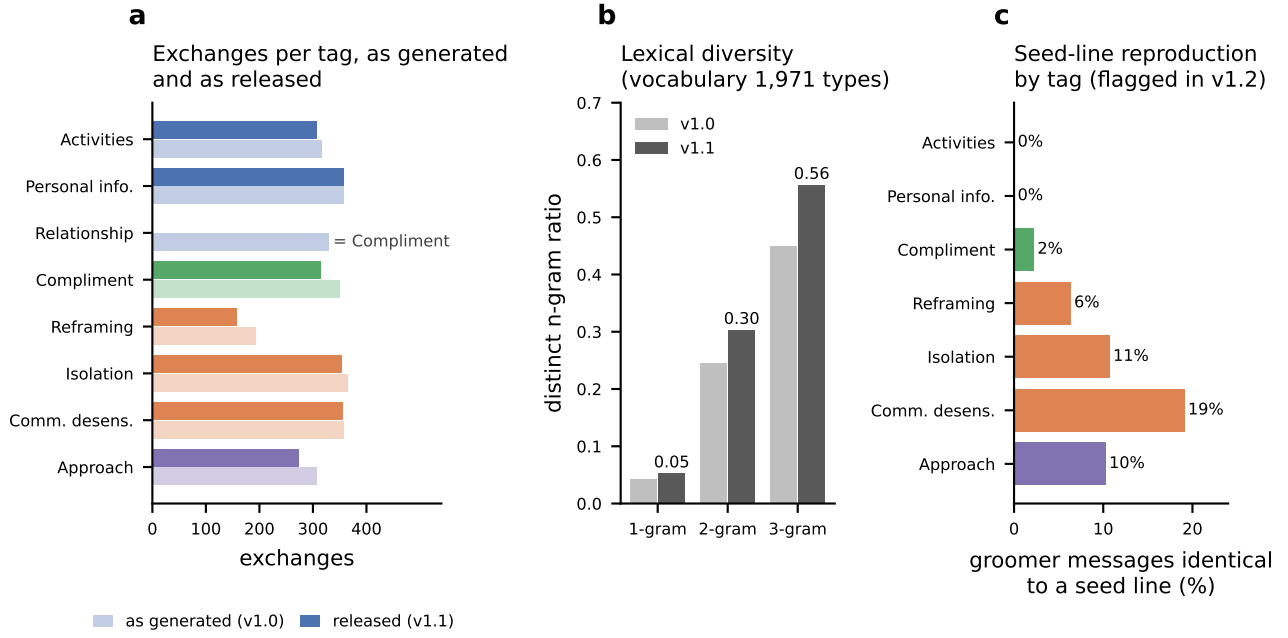


Fig. 2. Integrity, lexical profile and seed reproduction. (a) Exchanges per tag as generated (v1.0) and as released (v1.1); colours follow the phase (blue Targeting, green Fulfilling Needs, orange Isolation, purple Sexualizing). (b) Distinct- $n$  ratios of the two versions. (c) Share of groomer messages identical to a seed line, by tag; these exchanges carry the `seed_verbatim` flag in v1.2.

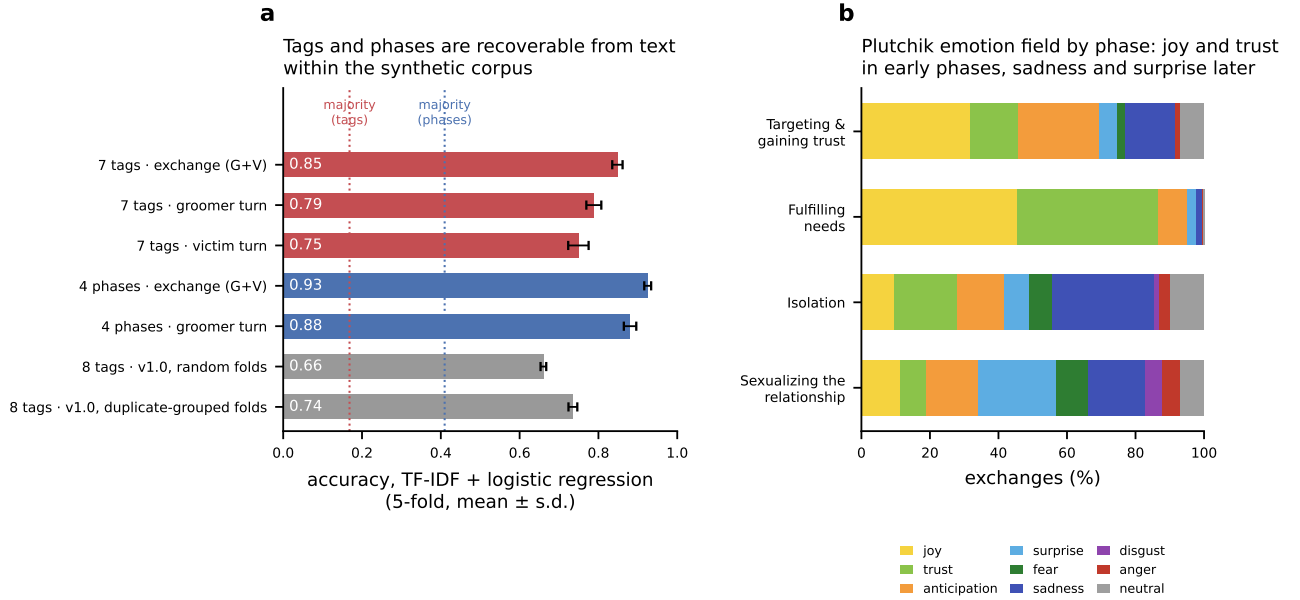


Fig. 3. In-domain validation. (a) Tag and phase recoverability from text with the lexical model (5-fold, mean  $\pm$  s.d.); grey rows are the v1.0 corpus before de-duplication. (b) Distribution of the Plutchik emotion field by phase ( $n = 2,121$  exchanges).

reflecting the lexical origin of the silver labels) and 0.422 on the three gold constructs (majority: 0.437; label-permutation floor: 0.278; in-domain gold:  $0.677 \pm 0.019$ ). The highest per-class recall on real lines is for Isolation (0.59) and Activities (0.66). Using semantic features, the same transfers achieve 0.388 accuracy for tags (majority 0.153 on the tag-capped sample), 0.448 accuracy for phases (majority 0.459), and 0.494 accuracy for constructs (majority 0.432; in-domain gold

$0.734 \pm 0.017$ ). In the reverse direction, real silver lines recover the synthetic tags with 26.8% (lexical) and 41.5% (semantic) accuracy. Real silver labels transfer to gold labels with 50.3% (lexical) and 56.7% (semantic) accuracy, placing construct-level coding of real chat among the more challenging tasks with this material. The confusion matrix on the gold constructs (Fig. 4c) shows 57-65% of real posts of every construct assigned to grooming, the construct corresponding

TABLE VIII

CROSS-DOMAIN TRANSFER BETWEEN THE SYNTHETIC CORPUS AND REAL CHATCODER OFFENDER LINES (CLASSIFIER TRAINED ON THE SOURCE, TESTED ON THE TARGET). SILVER = EIGHT-TAG DICTIONARY CODING OF 288 TRANSCRIPTS (31,640 SINGLE-TAG LINES; 5,230 IN THE SEMANTIC SETTING); GOLD = HUMAN THREE-CONSTRUCT CODING OF 55 TRANSCRIPTS (10,876 POSTS; 4,000 IN THE SEMANTIC SETTING). THE LAST COLUMN IS THE TARGET’S OWN 5-FOLD ACCURACY WITH TRANSCRIPT-GROUPED FOLDS.

| Direction / label set                              | Acc.  | F1    | Maj.  | In-dom. |
|----------------------------------------------------|-------|-------|-------|---------|
| <i>Lexical (TF-IDF + LR)</i>                       |       |       |       |         |
| Synthetic → real, 7 tags (silver)                  | 0.348 | 0.291 | 0.394 | 0.953   |
| Synthetic → real, 4 phases (silver)                | 0.427 | 0.318 | 0.524 | 0.961   |
| Synthetic → real, 3 constructs (gold)              | 0.422 | 0.371 | 0.437 | 0.677   |
| Real (silver) → synthetic, 7 tags                  | 0.268 | 0.222 | 0.168 | 0.788   |
| Real (silver) → synthetic, 4 phases                | 0.422 | 0.309 | 0.410 | 0.880   |
| Real silver → real gold, 3 constructs              | 0.503 | 0.418 | 0.437 | 0.677   |
| <i>Semantic (frozen all-roberta-large-v1 + LR)</i> |       |       |       |         |
| Synthetic → real, 7 tags (silver)                  | 0.388 | 0.375 | 0.153 | 0.869   |
| Synthetic → real, 4 phases (silver)                | 0.448 | 0.367 | 0.459 | —       |
| Synthetic → real, 3 constructs (gold)              | 0.494 | 0.456 | 0.432 | 0.734   |
| Real (silver) → synthetic, 7 tags                  | 0.415 | 0.396 | 0.168 | 0.768   |
| Real silver → real gold, 3 constructs              | 0.567 | 0.531 | 0.432 | 0.734   |

to the largest synthetic phase.

**Real-versus-synthetic discriminator.** A lexical classifier trained to distinguish synthetic grooming messages from a random sample of real offender messages of the same length achieves an area under the curve (AUC) of 0.962 and an accuracy of 0.906 (5-fold). On sentence embeddings, the AUC is 0.944. The words that are most indicative of synthetic content are contracted forms and the generator’s register (e.g. ‘you’re’, ‘ur’, ‘hey’, ‘favorite’, ‘always’, ‘wanna’, ‘such’, ‘anyone’, ‘private’, ‘secret’, ‘game’), while the words that are most indicative of real content are endearments, body words and logistics (e.g. ‘baby’, ‘sweetie’, ‘hun’, ‘bed’, ‘face’, ‘age’, ‘number’, ‘picture’, ‘camera’). The two registers can be separated on lexical and semantic grounds, independently of emoji and spelling. The synthetic corpus is polite, question-driven, and gaming-related, while the 2004–2008 PJ material is driven by endearments and logistics.

## VIII. STAGE 5: DOCUMENTATION, SCOPE OF USE AND PROTOCOL REFINEMENTS

### A. Scope of use

The measurements of Section VII define the corpus. Its tags and phases are consistently expressed in the text (recoverability: 0.85/0.93). It contains no data from minors and no offender lines from the flagged subset of the seed material. Its register is characteristic of the generator and separable from the 2004–2008 PJ offender chat register (discriminator AUC: 0.94–0.96). Labels learned on it transfer to real offender lines with an accuracy of 0.35–0.49, which is above the majority class on the gold constructs with semantic features. The corpus therefore supports controlled experiments in which the phase is the manipulated variable, as well as pipeline, prompt and annotation-tool development. It can also be used to augment real data alongside the transfer gap being measured on the target, and for educational and awareness material where the clarity of the manipulative move is the relevant property. The

datasheet outlines this scope and the transfer figures provide a reference point for any new use.

### B. Protocol refinements

Applying the generation protocol [17] to this corpus yields five refinements for future synthetic corpora in sensitive domains. 1) *Integrity first*: byte-wise comparison of output files and normalised de-duplication precede every other measurement, and per-label counts are reported exactly.

2) *Measured, flagged seed reproduction*: an instruction to paraphrase is complemented by a measurement of reproduction against the retrieval base and by row-level flags; reproduction concentrates in the tags where the generator’s own knowledge is thinnest.

3) *Recoverability as a consistency check*: in-domain accuracy documents label consistency, and a lexical model provides the reference against which transformer results are read.

4) *Real reference at its own granularity*: each available real resource is used with its own label set and label status stated (silver dictionary coding, gold human coding), under both lexical and semantic features.

5) *Run logging*: decoding parameters, seeds and per-model refusal counts are recorded per run; the released code implements this for future generations.

### C. Limitations

The real reference consists of decoy conversations from 2004–2008, which were coded using either a keyword dictionary (silver) or a numeric codebook that was not distributed with the data (gold). The construct mapping was inferred from the co-occurrence of dictionary tags with each code, as well as from the theory. The synthetic corpus is English-only, text-only and static. It was produced using one 8B checkpoint at 4-bit quantisation. A larger generator or one adapted to the domain would change the diversity and transfer figures. The protocol therefore prescribes measuring them again. The emotion field is silver annotation. Seed-reproduction analysis covers the 337 curated seed lines, but cannot reproduce other PJ material memorised during pre-training.

### D. Toward dynamic, role-aware successors

This study has three key findings:

- 1) Emotion-trajectory features across the turns of a conversation, exploiting the emotion field released with the corpus.
- 2) A two-agent simulator in which domain-adapted models play groomer and victim to generate richer synthetic dialogues, as well as serving as an awareness-training tool for parents, educators, and minors.
- 3) Domain-specific fine-tuning to replace prompt-level control of general-purpose checkpoints. The present corpus and its validation figures serve as a reference point for this agenda.

## IX. REPRODUCIBILITY AND ARTIFACT RELEASE

We release, under the vetted-access licence: the labelled corpus (v1.1 and v1.2) with datasheet; the structure and formatting specification of the five phase documents; redacted

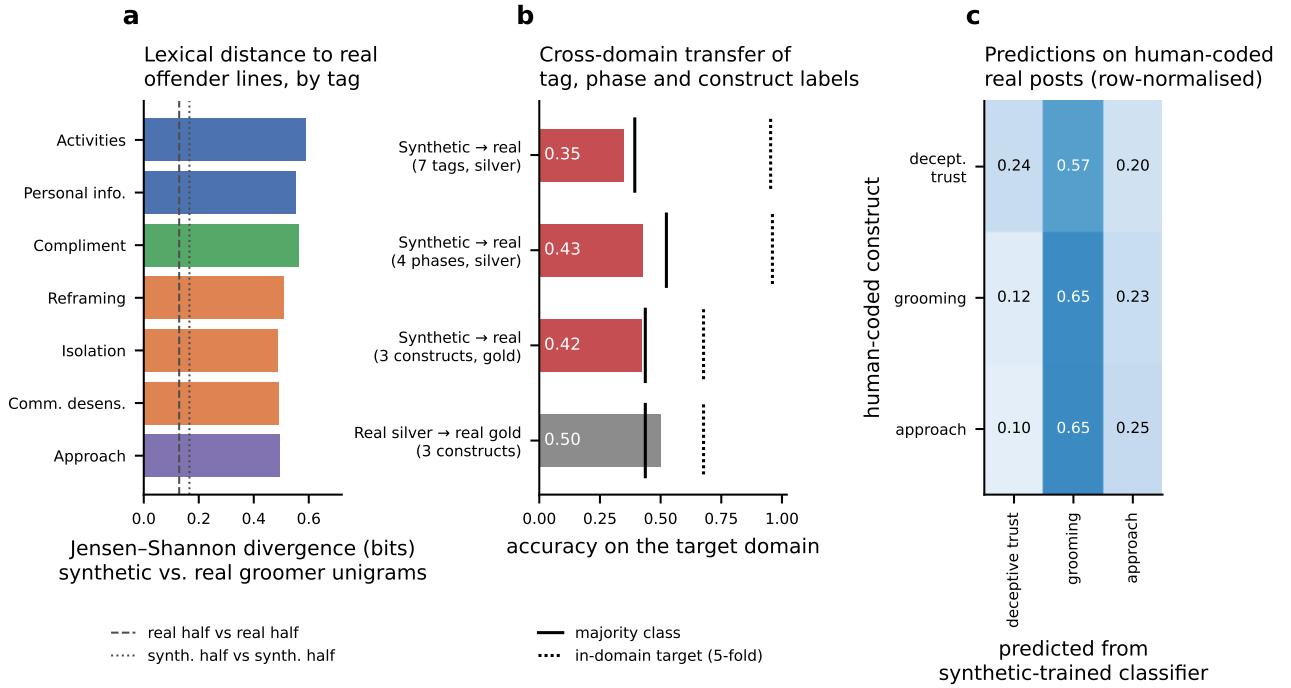


Fig. 4. External validation against real ChatCoder data (lexical features): (a) The Jensen-Shannon divergence between the unigram distributions of real and synthetic offender data per tag. The dashed and dotted lines show the divergence between two random halves of the real and synthetic data, respectively. (b) Accuracy of classifiers trained in one domain and tested in the other (bars), compared with the target majority class (solid tick) and the target’s own in-domain 5-fold accuracy (dotted tick). (c) A row-normalised confusion matrix of the synthetic-trained classifier on human-coded constructs ( $n = 10,876$  posts).

system and question prompt templates; the generation protocol (strategy, batch size, model identifiers and quantisations); the real-reference construction, audit and validation code with the result tables; and the complete classifier configurations, which suffice to reproduce Tables VII–VIII on the released data. Generation ran on the UniNuvola node described in Section IV with Python 3.10, PyTorch 2.2.1, Hugging Face Transformers 4.38.2, CUDA 12.1, Ollama 0.4.6 and LangChain 0.2.2. Random seeds for exact regeneration are not released, as part of the dual-use mitigation that governs the licence; seed documents, prompt components addressing refusal, and environment files are likewise withheld. The ChatCoder resources used as real references were distributed by their authors [10].

## X. ETHICAL CONSIDERATIONS AND RESPONSIBLE DISCLOSURE

This research concerns sexual communication with minors and is classified as dual-use. The generation process was localised and the only available material is the publicly distributed Perverted Justice archive. In this archive, the victims are adult volunteers. The seed documents contain excerpts of real offender messages and remain with the authors. The prompt components that address refusal are described functionally. The corpus is released under a restricted, vetted-access licence to verified researchers and safeguarding organisations, along with the datasheet, audit code and v1.2 flags, under a data-use agreement. Samples of the most explicit tags are included and flagged because they are required for validation purposes. This work was conducted for preventive,

educational and law enforcement support purposes under institutional oversight.

## DATA AVAILABILITY

A partial version of the dataset will be deposited in IEEE DataPort (<https://iee-dataport.org/documents/lure-luring-theory-utterances-research-grooming-exchanges>) and made available in a limited version upon acceptance. The full dataset will be made available to verified researchers and safeguarding organisations upon vetted access, under a data-use agreement. During the review process, the dataset and datasheet can be provided to the editorial office or reviewers via an appropriate channel. Seed materials, random seeds for exact regeneration and components for handling refusals are not released. Requests can be directed to the corresponding author.

## ACKNOWLEDGMENT

This paper builds on the M.Sc. thesis of L. Guercio (University of Perugia) [53]. The authors thank the stakeholders in child-safety and law enforcement who advised on the phase taxonomy and ethical governance. This work did not receive project funding; it is supported by the Emotions and Multidisciplinary-Oriented Research on Ethical and Responsible AI (EMORE) Lab, University of Perugia, Italy. The OA publication is funded by the University of Perugia. The authors declare the use of Large Language Models as a research methodology for generating the synthetic dataset under human supervision, as detailed in Sections IV–V; all content has been

verified by the authors, who assume full responsibility for this work. In line with current editorial policies on the use of artificial intelligence in article preparation, the authors disclose that an AI assistant (Claude Science) was used to support coding for data analysis and for language enhancement (DeepL Write) under the authors' direction; all analyses, results and text were checked by the authors. The authors declare no conflict of interest.

#### AUTHOR CONTRIBUTIONS

Authors are listed in alphabetical order. Conceptualization: V.F.; Methodology: V.F., A.M., L.G.; Software: L.G.; Validation: V.F., L.G., E.F., M.P.; Formal analysis: V.F., L.G.; Investigation: L.G.; Data curation: L.G., M.P.; Visualization: L.G., V.F.; Writing-original draft: V.F., L.G., M.P.; Writing-review and editing: V.F., M.P., E.F.; Supervision: V.F.

All authors have read and agreed to the published version of the manuscript.

#### REFERENCES

- [1] H. Whittle, C. Hamilton-Giachritsis, A. Beech, and G. Collings, "A review of young people's vulnerabilities to online grooming," *Aggression and Violent Behavior*, vol. 18, no. 1, pp. 135–146, 2013.
- [2] J. Wolak, D. Finkelhor, and K. Mitchell, "Internet-initiated sex crimes against minors: Implications for prevention based on findings from a national study," *Journal of Adolescent Health*, vol. 35, no. 5, pp. 424.e11–424.e20, 2004.
- [3] Polizia Postale e delle Comunicazioni, "Resoconto annuale delle attività contro la pedofilia online e i reati contro i minori," Polizia di Stato, Rome, Italy, Tech. Rep., 2024. [Online]. Available: <https://www.poliziadistato.it>
- [4] Internet Watch Foundation, "Annual report 2023," Internet Watch Foundation, Cambridge, U.K., Tech. Rep., 2023. [Online]. Available: <https://www.iwf.org.uk>
- [5] National Center for Missing & Exploited Children, "Cybertipline reports and data," <https://www.missingkids.org/cybertiplinedata>, Alexandria, VA, USA, 2023.
- [6] A. Chan, J. Hallett, and M. Edwards, "Why is johnny getting groomed in roblox?" *IEEE Security & Privacy*, vol. 23, no. 3, pp. 20–30, 2025.
- [7] H. Abelson, R. Anderson, S. M. Bellovin, J. Benaloh, M. Blaze, J. Callas, W. Diffie, S. Landau, P. G. Neumann, R. L. Rivest, J. I. Schiller, B. Schneier, V. Teague, and C. Troncoso, "Bugs in our pockets: the risks of client-side scanning," *Journal of Cybersecurity*, vol. 10, no. 1, p. tyad020, 01 2024.
- [8] European Parliament and Council, "Regulation (EU) 2021/1232 on a temporary derogation for the purpose of combating online child sexual abuse, and the proposed regulation laying down rules to prevent and combat child sexual abuse (COM(2022) 209)," <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32021R1232>, 2021, vERIFY: cite the 2026 legislative act/position actually referred to.
- [9] L. N. Olson, J. L. Daggs, B. L. Ellevold, and T. K. K. Rogers, "Entrapping the innocent: Toward a theory of child sexual predators' luring communication," *Communication Theory*, vol. 17, no. 3, pp. 231–251, 2007.
- [10] A. Kontostathis, L. Edwards, and A. Leatherman, "ChatCoder: Toward the tracking and categorization of internet predators," in *Proceedings of the Text Mining Workshop, Ninth SIAM International Conference on Data Mining (SDM 2009)*, 2009.
- [11] G. Inches and F. Crestani, "Overview of the international sexual predator identification competition at PAN-2012," in *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*, ser. CEUR Workshop Proceedings, vol. 1178, 2012.
- [12] P. R. Borj, K. Raja, and P. Bours, "Online grooming detection: A comprehensive survey of child exploitation in chat logs," *Knowledge-Based Systems*, vol. 259, p. 110039, 2023.
- [13] M. Vogt, U. Leser, and A. Akbik, "Early detection of sexual predators in chats," in *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021, pp. 4985–4999.
- [14] European Union, "Regulation (EU) 2016/679 (general data protection regulation); directive (EU) 2016/680," Official Journal of the European Union, L 119, 2016. [Online]. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679>
- [15] L. Long, R. Wang, R. Xiao, J. Zhao, X. Ding, G. Chen, and H. Wang, "On LLMs-driven synthetic data generation, curation, and evaluation: A survey," in *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 11 065–11 082.
- [16] K. M. Yoo, D. Park, J. Kang, S.-W. Lee, and W. Park, "Gpt3mix: Leveraging large-scale language models for text augmentation," in *Findings of ACL: EMNLP 2021 November 7-11, 2021*, M. M.-F., H. X., S. L., and Y. S.W.-T., Eds. Association for Computational Linguistics (ACL), 2021, Conference paper, p. 2225 – 2239, cited by: 158; All Open Access, Gold Open Access.
- [17] V. Franzoni, "Towards responsible generation of synthetic datasets with large language models," in *Proceedings of the 2025 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 2025, vERIFY co-authors and page numbers against the IEEE Xplore record (document 11676877).
- [18] M. Amici, M. Polticchia, E. Florindi, and V. Franzoni, "A synthetic Plutchik-grounded corpus for fine-grained emotion recognition in text," *Neural Computing and Applications*, 2026, manuscript NCAA-D-26-04319, under review, can be put in Arxiv if acceptance doesn't arrive before this paper.
- [19] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.
- [20] J. Galbally, R. Haraksim, and L. Beslay, "A study of age and ageing in fingerprint biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1351–1365, 2019.
- [21] M. Shen, J. Wu, K. Ye, K. Xu, G. Xiong, and L. Zhu, "Robust detection of malicious encrypted traffic via contrastive learning," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 4228–4242, 2025.
- [22] J. Ghadermazi, S. Hore, A. Shah, and N. D. Bastian, "Gtae-ids: Graph transformer-based autoencoder framework for real-time network intrusion detection," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 4026–4041, 2025.
- [23] A. Jevremovic, M. Veinovic, M. Cabarkapa, M. Krstic, I. Chorbev, I. Dimitrovski, N. Garcia, N. Pombo, and M. Stojmenovic, "Keeping children safe online with limited resources: Analyzing what is seen and heard," *IEEE Access*, vol. 9, pp. 132 723–132 732, 2021.
- [24] G. Mujtaba, S. Ali Khawaja, and K. Dev, "Edgeaiguard: Agentic llms for minor protection in digital spaces," *IEEE Internet of Things Journal*, vol. 12, no. 17, pp. 34 992–35 000, 2025.
- [25] Perverted Justice Foundation, "Perverted-Justice chat-log archive," Archived chat logs, 2007.
- [26] K. Chehbouni, M. De Cock, G. Caporossi, A. Taik, R. Rabbany, and G. Farnadi, "Enhancing privacy in the early detection of sexual predators through federated learning and differential privacy," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 39, no. 27, 2025, pp. 27 887–27 895.
- [27] E. Prosser and M. Edwards, "Helpful or harmful? Exploring the efficacy of large language models for online grooming prevention," in *Proc. European Interdisciplinary Cybersecurity Conference (EICCC 2024)*. New York, NY, USA: ACM, 2024, pp. 1–10.
- [28] Y.-G. Cheong, A. K. Jensen, E. R. Gudnadottir, B.-C. Bae, and J. Togelius, "Detecting predatory behavior in game chats," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 7, no. 3, p. 220 – 232, 2015, cited by: 24.
- [29] S. Murnion, W. J. Buchanan, A. Smales, and G. Russell, "Machine learning and semantic analysis of in-game chat for cyberbullying," *Computers and Security*, vol. 76, p. 197 – 213, 2018, cited by: 78; All Open Access, Green Open Access.
- [30] Radicalisation Awareness Network (RAN C&N), "Digital grooming tactics on video gaming & video gaming adjacent platforms: Threats and opportunities," European Commission, DG Migration and Home Affairs, Tech. Rep., 2021. [Online]. Available: [https://home-affairs.ec.europa.eu/system/files/2021-05/ran\\_c\\_n\\_conclusion\\_paper\\_grooming\\_through\\_gaming\\_15-16032021\\_en.pdf](https://home-affairs.ec.europa.eu/system/files/2021-05/ran_c_n_conclusion_paper_grooming_through_gaming_15-16032021_en.pdf)
- [31] E. Florindi, "Pedofilia e videogiochi: storia ed evoluzione di un fenomeno (Pedophilia and video games: Analysis of a growing phenomenon)," *SSRN*, 2024.
- [32] N. Lorenzo-Dus, C. Izura, and R. Pérez-Tattam, "Understanding grooming discourse in computer-mediated environments," *Discourse, Context & Media*, vol. 12, pp. 40–50, 2016.

- [33] G. M. Winters, L. E. Kaylor, and E. L. Jeglic, "Sexual offenders contacting children online: An examination of transcripts of sexual grooming," *Journal of Sexual Aggression*, vol. 23, no. 1, pp. 62–76, 2017.
- [34] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474.
- [35] Z. Li, H. Zhu, Z. Lu, and M. Yin, "Synthetic data generation with large language models for text classification: Potential and limitations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, vERIFY pages/DOI; arXiv:2310.07849.
- [36] I. Shumailov, Z. Shumailov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "AI models collapse when trained on recursively generated data," *Nature*, vol. 631, pp. 755–759, 2024.
- [37] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587–604, 2018.
- [38] V. Franzoni, G. Biondi, and A. Milani, "Emotional sounds of crowds: Spectrogram-based analysis using deep learning," *Multimedia Tools and Applications*, vol. 79, no. 47–48, pp. 36 063–36 075, 2020.
- [39] A. E. Baia, G. Biondi, V. Franzoni, A. Milani, and V. Poggioni, "Lie to me: Shield your emotions from prying software," *Sensors*, vol. 22, no. 3, p. 967, 2022.
- [40] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992.
- [41] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.
- [42] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.
- [43] D. Finkelhor, *Child Sexual Abuse: New Theory and Research*. New York, NY, USA: Free Press, 1984.
- [44] T. Ward and S. M. Hudson, "Finkelhor's precondition model of child sexual abuse: A critique," *Psychology, Crime & Law*, vol. 7, no. 4, pp. 291–307, 2001.
- [45] R. B. Cialdini, *Influence: The Psychology of Persuasion*. HarperCollins, 2009.
- [46] M. Amici, M. Polticchia, E. Florindi, and V. Franzoni, "A synthetic plutchik-grounded corpus for fine-grained emotion recognition in text," *IEEE DataPort*, 2026. [Online]. Available: <https://ieee-dataport.org/documents/synthetic-plutchik-grounded-corpus-fine-grained-emotion-recognition-text>
- [47] Ollama, "Ollama," <https://ollama.com>, 2024, accessed: 2025-07-09.
- [48] H. Chase, "Langchain," <https://github.com/langchain-ai/langchain>, 2022, accessed: 2025-07-09.
- [49] huihui ai, "Llama-3.2-3b-instruct-abliterated," <https://huggingface.co/huihui-ai/Llama-3.2-3B-Instruct-abliterated>, 2025, hugging Face model card. Accessed: 2025-07-09.
- [50] M. Labonne, "Meta-llama-3.1-8b-instruct-abliterated," <https://huggingface.co/mlabonne/Meta-Llama-3.1-8B-Instruct-abliterated>, 2024, hugging Face model card. Accessed: 2025-07-09.
- [51] huihui ai, "Deepseek-r1-distill-llama-8b-abliterated," <https://huggingface.co/huihui-ai/DeepSeek-R1-Distill-Llama-8B-abliterated>, 2025, hugging Face model card. Accessed: 2025-07-09.
- [52] A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda, "Refusal in language models is mediated by a single direction," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 136 037–136 083.
- [53] L. Guercio, M. Polticchia, E. Florindi, and V. Franzoni, "Towards the prevention of online grooming: Arousal classification in the early phases of online luring," M.Sc. thesis, Università degli Studi di Perugia, Dipartimento di Matematica e Informatica, Perugia, Italy, 2023/2024.
- Mathematics and Computer Science, University of Perugia, where he serves as Data Protection Officer. His research spans cybercrime, artificial intelligence and military professional training. He received his degrees in law and in psychology from the University of Perugia, is a member of scientific and ethical committees in international projects, and a member of the IEEE.
- Valentina Franzoni** is Associate Professor at the Department of Mathematics and Computer Science, University of Perugia, Italy, where she serves as Rector's subdelegate for Artificial Intelligence, with responsibility for affective computing, AI ethics and social impact, and leads the Emotions and Multidisciplinary-Oriented Research on Ethical and Responsible AI (EMORE) Lab. Her research interests include complex networks, artificial intelligence and affective computing. She received her PhD in Engineering for Computer Science from Sapienza University of Rome (DIAG). She is a Senior Member of the IEEE and a member of AAAC, ACM and the CINI Smart Cities & Communities National Laboratory.
- Ludovico Guercio** is a postgraduate of the Department of Mathematics and Computer Science, University of Perugia, Italy, where he received his B.Sc. in Computer Science and M.Sc. in Computer Science–Cybersecurity. His research focuses on artificial intelligence and cybersecurity, in particular AI-based prevention of online child grooming.
- Alfredo Milani** is Associate Professor at Link Campus University, Rome, Italy, where he directs the AI Multicognitive Research Laboratory (AIMLab), affiliated with the CAIRNE European network of AI research. His research interests include evolutionary algorithms, agent planning and generative AI. He received his degree in Information Science from the University of Pisa, Italy, and is a Senior Member of the IEEE.
- Mattia Polticchia** is a PhD candidate in Learning Sciences and Digital Technologies at the University of Modena and Reggio Emilia, Italy. His research interests span game science and serious games, virtual reality and affective computing. He received his B.Sc. in Computer Science and M.Sc. in Computer Science–Artificial Intelligence from the University of Perugia, Italy, summa cum laude, and is a member of the IEEE.

**Emanuele Florindi** is a PhD candidate in Learning Sciences and Digital Technologies at the University of Modena and Reggio Emilia, Italy, previously adjunct professor in law and emerging technologies at the Department of