Aletheia.

Detecting online grooming patterns at scale.

Online enticement reports to NCMEC rose 192 percent in a single year, reaching 546,000 in 2024[1]. Nearly all of the automated detection in production today operates on media files, not on the conversations that produce them. Aletheia is Polycreek's hierarchical transformer for grooming detection in text, trained on 2.5 million conversations across 20 sources and 10+ languages, and validated at F1 0.92. This whitepaper sets out the architecture, the training data, the validation results, and the operational principles that guide its deployment as a nonprofit-priced detection layer for platforms and child-safety organizations.

2.5M

Conversations in Aletheia's training corpus, drawn from 20 distinct sources.

10+

Languages covered. English is the largest single share at roughly 84 percent.

8

Behavioral grooming phases used as discrete classification targets.

SUMMARY

The automated child-safety stack at most large platforms today is built around image hashing and image classification. Both are necessary. Both are blind to the conversation that precedes any image. That conversation is where most grooming actually occurs.

Aletheia is a hierarchical transformer model for grooming detection in text. It segments arbitrarily long conversations, encodes each segment with a pretrained backbone, attends across segments, and produces two outputs at the conversation level: a binary harmful-conversation score, and an attribution of which speaker is the predatory party. It is trained on 2,523,202 conversations drawn from 20 corpora across 10 languages.

This whitepaper documents the scale of the problem, the academic basis for the eight-phase grooming model Aletheia targets, the architecture, the training corpus, and the operational principles that constrain how the model is deployed.

CONTENTS

01

What this paper is about

Most of the automated child-safety stack at large platforms today is looking for known illegal images. Microsoft PhotoDNA, in production since 2009, matches uploaded images against databases of catalogued abuse material[2]. Predictive image classifiers extend that coverage to novel media. Both are necessary. Both leave the same thing uncovered: **the conversation that precedes any image.**

Grooming is conversational. It can take weeks. It can take an afternoon. The signals are linguistic, behavioral, and structural, and they are distributed across many turns rather than concentrated in any single message. Detecting it means analyzing patterns of interaction at the level of full dialogues, in many languages, against contemporary platform vocabularies, with predictable performance and an explicit human-review escalation path.

Aletheia is Polycreek's detection model for that gap. It is a hierarchical transformer that segments arbitrarily long conversations, encodes each segment with a pretrained backbone, attends across segments, and emits two outputs: a binary harmful-conversation score, and an attribution of which speaker is the predatory party. It is now in its third architectural iteration, with validation results that exceed the academic baselines on a corpus three orders of magnitude larger than the field's standard reference.

This paper documents the scale of the problem, the academic literature behind Aletheia's eight-phase classification model, the architecture and training corpus, the validation results, and the operational principles that constrain deployment.

Polycreek is a 501(c)(3) nonprofit. Aletheia exists because the largest trust-and-safety vendors have spent more than a decade investing in image scanning and hash matching, but have not built or deployed a conversational-detection layer at the scale of the threat. Academic research, working from a 2012 corpus of roughly a thousand conversations, has not closed that gap either. Polycreek built Aletheia to do what neither commercial nor academic actors have done. It is offered as a paid API and licensed deployment; because Polycreek is a nonprofit, every dollar funds continued development, training data, research staff, and operations. There are no shareholders.

> Hash matching catches known imagery. Image classifiers catch novel imagery. Neither sees the conversation that produces the imagery. The biggest names in trust and safety have not built that layer. Polycreek built Aletheia for that.

02

The scale of the problem

NCMEC's CyberTipline processed 20.5 million reports of suspected online child sexual exploitation in 2024[1]. The headline figure declined from 36.2 million in 2023, but for a counterintuitive reason: NCMEC introduced a bundling mechanism that consolidates duplicate viral content into single reports. Adjusted for that, the underlying threat continued to rise, and the categories most directly tied to grooming grew the fastest.

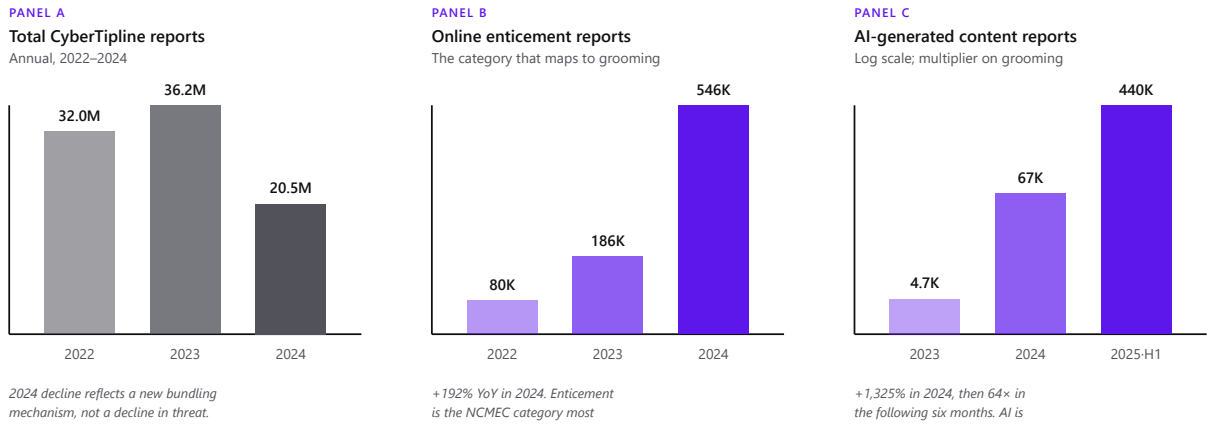


FIGURE 1 The headline number declined while the categories most relevant to grooming detection rose sharply. Sources: NCMEC *2024 in Numbers*[1]; NCMEC mid-2025 update[,3].

Online enticement, the NCMEC category that maps most directly to grooming, rose from 186,000 reports in 2023 to 546,000 in 2024, a 192 percent year-over-year increase[1]. By the first half of 2025, enticement reporting and AI-content reporting had both continued to climb[,3]. Generative AI is a force multiplier for grooming specifically: synthetic personas accelerate trust-building, on-demand explicit imagery accelerates desensitization, and child-simulating chatbots are now being used to rehearse grooming strategies[,22].

546K

online enticement reports to NCMEC in 2024, up from 186K in 2023[1].

+192%

year-over-year growth in enticement, the category most directly tied to grooming.

440K+

AI-generated content reports in H1 2025 alone, against 6,800 in H1 2024[3].

The Internet Watch Foundation, the UK reporting body, assessed 424,047 reports in 2024 and actioned 291,273 webpages, with self-generated content, often produced under coercive grooming, accounting for the majority[4]. These figures describe the visible portion of the problem. Underreporting, end-to-end encryption rollouts, and jurisdictional fragmentation hide the rest.

03

How grooming actually works

Grooming is not an event. It is a process through which an adult builds the relationship that makes exploitation possible. Two decades of research has converged on a recognizable lifecycle, while also documenting that the lifecycle does not run linearly. Aletheia models eight phases as classification targets, derived from O'Connell (2003), Olson et al. (2007), Black et al. (2015), Winters and Jeglic (2017), and the synthesis in Polycreek's internal grooming guidelines[5,6,7,8].

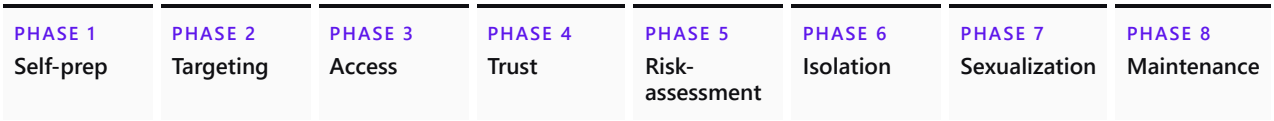


Figure 2. The eight-phase canonical grooming lifecycle used as classification targets in Aletheia. Phases 5 and 6 in particular do not occur as discrete sequential events; risk-assessment runs continuously, and isolation tactics escalate as soon as the offender encounters peer or parental visibility.

What the literature actually shows

Three findings from the literature shape how Aletheia is built. First, grooming does not follow the canonical sequence cleanly. Whittle and colleagues' 2013 review, and subsequent work by Kloess and others, document that real offenders move between phases, regress under resistance, and adapt their language to the victim's responses[9,10]. A classifier trained on a fixed phase sequence will fail against the offender who skips phases or loops back.

Second, the discriminative signal is rarely in any single utterance. Black and colleagues' 2015 linguistic analysis showed that grooming conversations are characterized by an unusually positive emotional tone (around 47 percent positive emotion words, against roughly 26 percent in normal chat) and an elevated rate of affiliation language, but neither marker is independently diagnostic[7]. Lorenzo-Dus and Izura's work on complimenting behavior arrives at a similar conclusion: it is the co-occurrence and sequencing across the dialogue that distinguishes grooming, not any one move[11].

Third, the lifecycle has been compressing. Winters and Jeglic documented that a non-trivial fraction of offenders introduce sexual content within their first conversation[8]. Contemporary attack patterns can collapse the canonical sequence into a single session, particularly on platforms with default-private messaging and weak age assurance. A grooming detector designed for the slow, multi-week pattern will miss the same-day version.

Two patterns are consequential enough to call out separately. Risk-assessment questions probing the child's surveillance environment ("Is your computer in your room? Do your parents check your phone?") are near-pathognomonic; benign adults talking to minors do not systematically probe parental monitoring. Platform migration, the movement of a conversation from a moderated channel to an encrypted or unmonitored one, is a top-tier structural signal even in the absence of explicit content. Both are reflected as structural-flag features in Aletheia's training data.

04

Why text-based detection has been hard

Text-based grooming detection is the least mature of the major child-safety detection categories. The reasons are specific, known, and addressable. Aletheia is built around addressing them.

Dataset staleness

Most published grooming-detection research has trained on the PAN12 corpus or scraped Perverted Justice transcripts[12]. Both predate 2012. PAN12 in particular has been the field's reference benchmark for over a decade. Grooming has not stood still in that time. It has migrated to ephemeral messaging, voice notes, in-game chat, and platforms that did not exist when the PAN12 transcripts were collected. Models trained on PAN12 alone inherit that staleness, both in vocabulary and in interactional structure.

Monolingual coverage

The overwhelming majority of grooming-detection research has been conducted in English. Online exploitation is not. The gap between research-language coverage and threat-language coverage is one of the larger unaddressed problems in the field.

Adversarial fragility

Classifiers built around explicit lexical features can be evaded by offenders who use euphemism, code language, or deliberately neutral phrasing. Bogdanova and colleagues showed in 2014 that high-level features outperform surface lexicon for exactly this reason[13].

The false-positive cost

Accuracy is the wrong metric for the operating regime. At platform scale, even a 1 percent false-positive rate on a billion daily messages produces ten million flagged conversations every day. Each one is either an unjustified intrusion on a normal user or, if ignored due to volume, a degradation of analyst capacity to find the genuine threat. Detection that does not budget for the false-positive cost will not be deployed.

Signal isolation, and what Aletheia does about it

Existing work analyzes text in isolation, with no access to the surrounding behavior: who is talking to whom, across what channels, in what relational structure. Aletheia's design responds directly to each of these problems. The training corpus is multi-source and language-diverse rather than PAN12-only. Adversarial robustness comes from training across many distinct registers of harmful conversation. False-positive cost is managed through a calibrated risk-score output, with explicit thresholds (Safe, Low Risk, Medium Risk, High Risk, Critical) tied to escalation policy. The model also accepts structural-flag inputs alongside conversation text, putting the surrounding behavioral context into the same representation as the language itself.

05

How Aletheia works

Aletheia is a hierarchical transformer that operates at the conversation level. Single-turn classifiers cannot capture the cross-turn patterns the literature identifies as discriminative. Aletheia is built to.

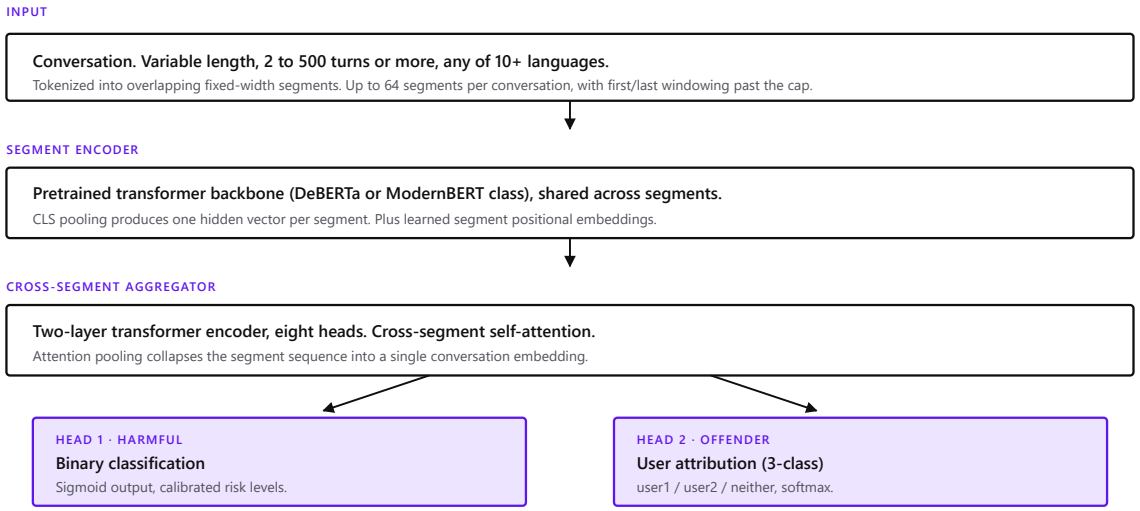


FIGURE 3 Aletheia v3 architecture. Joint training uses a weighted sum of harmful (binary cross-entropy) and offender-attribution (cross-entropy) losses.

Why hierarchical

The fundamental constraint on text-based detection is conversation length. A grooming dialogue can run hundreds of turns; standard transformer encoders have fixed input length (typically 512 or 8,192 tokens). Aletheia handles arbitrary length by segmenting first and aggregating second. Past the maximum segment count of 64, the model retains the first quarter and the last three-quarters of segments, preserving both opening behavior (rapport-forming, age inquiry) and recent behavior (sexualization, meeting logistics).

Why two heads, and what they output

A binary "is this harmful" output is not enough. In any flagged conversation, an analyst needs to know which participant is the predatory party. Aletheia's offender-attribution head emits a softmax over *user1*, *user2*, and *neither*. Joint training (loss weighted 0.6 harmful to 0.4 attribution) forces the model to encode role-specific patterns rather than topic alone, and produces an output that is directly actionable for review queues.

The harmful score is a probability in [0, 1], bucketed into five operational risk levels for downstream routing: *Safe* (< 0.2), *Low Risk* (0.2 to 0.4), *Medium Risk* (0.4 to 0.6), *High Risk* (0.6 to 0.8), and *Critical* (≥ 0.8). Confidence is reported alongside, computed jointly from distance to the decision boundary and the entropy of the user-attribution distribution.

06

What the numbers say

Aletheia has shipped two major architectural iterations and is preparing the third evaluation milestone. The validation numbers are public, the failure modes are documented, and the next eval is in the build.

METRIC	VALUE	SOURCE
PHASE 1 VALIDATION F1	**0.92**	held-out 20% of 409K-conversation training set
PHASE 1 VALIDATION AUC	**0.99**	same
PHASE 1 VALIDATION LOSS	**0.024**	same
PAN12 ACADEMIC BASELINE F1	0.85–0.90	published benchmark range[12]
PHASE 3 ZERO-SHOT DIAGNOSTIC	**10/11 correct (91%)**	hand-curated 11-case test
PHASE 3 VS PHASE 1, PEER FLIRTING	**3/3 fixed**	the original failure mode

Phase 1: the supervised hierarchical classifier

The architecture described in the previous section is what Polycreek calls Phase 1, the supervised iteration. Trained on a held-out 20 percent split of a 409,000-conversation slice of the unified corpus, it produced an F1 of 0.92, AUC of 0.99, and validation loss of 0.024. Published academic baselines on PAN12 reach F1 of roughly 0.85 to 0.90; Phase 1 exceeds that range on a corpus three orders of magnitude larger and substantially more diverse.

Phase 3: the zero-shot iteration that ships today

The current production architecture (Phase 3) moves to a zero-shot approach: a strong base model paired with a custom Polycreek policy derived directly from the eight-phase canonical lifecycle. This was a deliberate response to a Phase 2 failure mode in which a fine-tuned classifier conflated symmetric-register peer flirting with adult-minor predation, the kind of conflation that would generate unacceptable false positives at platform scale.

On an 11-case hand-curated diagnostic spanning peer flirting, fake-peer grooming, and adult-minor predation, Phase 3 returns 10 of 11 correct, including all three peer-flirting cases the supervised classifier had failed on. The single non-match is a peer sexting case the model flags as harmful but Polycreek's scope policy treats as out-of-scope, which is a definitional question rather than a detection failure.

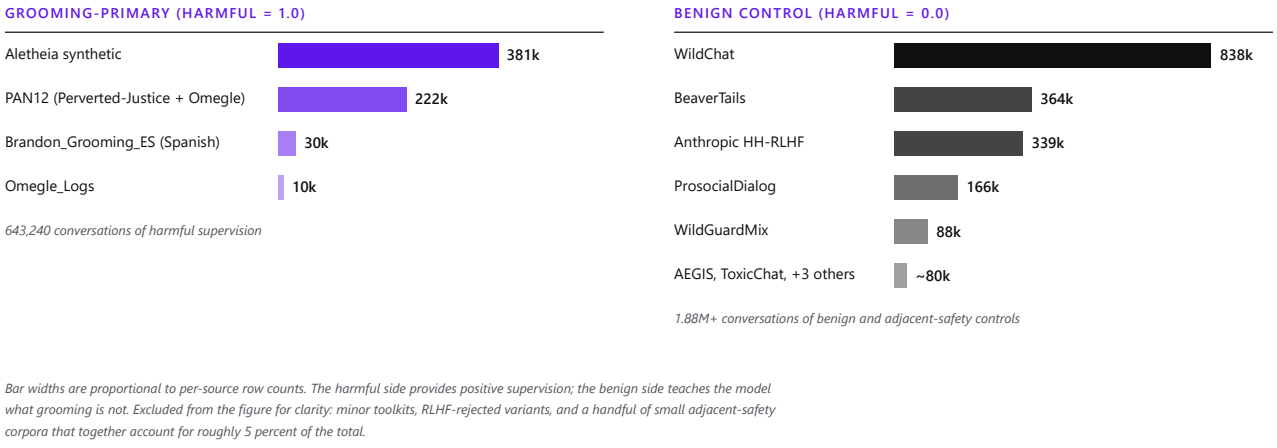
What is in the build

The next evaluation milestone is a 500-case Polycreek-curated test set with five-class labels, designed to yield aggregate precision, recall, and F1 on cases that match Polycreek's actual scope. This replaces the field's reliance on PAN12, a benchmark from 2012 that does not reflect contemporary platforms or attack patterns.

07

The training corpus

Aletheia's training corpus is 2,523,202 conversations drawn from 20 distinct sources. The composition is deliberate: the academic literature has spent more than a decade overfitting to PAN12. Aletheia is trained against a corpus that is more than three orders of magnitude larger and substantially more diverse.



Bar widths are proportional to per-source row counts. The harmful side provides positive supervision; the benign side teaches the model what grooming is not. Excluded from the figure for clarity: minor toolkits, RLHF-rejected variants, and a handful of small adjacent-safety corpora that together account for roughly 5 percent of the total.

FIGURE 4 Top-level composition of Aletheia's training corpus across 20 distinct sources. Source: Polycreek *Aletheia Unified Dataset Build Report*[23].

2.52M	208K	10+	41
total conversations	harmful=1.0 (8.4%)	languages represented	columns of structured annotation

Three points are worth flagging about the design of this corpus. First, the negative controls dominate. RLHF-preference, prosocial-dialog, and chit-chat sources together exceed two million conversations. The model needs to see what benign adult-minor and adult-adult interaction actually looks like, and there is no shortage of it. Second, the supervision is multi-source on the harmful side as well. Synthetic conversations generated under controlled conditions complement PAN12's authentic-but-dated transcripts, with separate Spanish supervision from Brandon_Grooming_ES. Third, the schema includes 13 boolean structural-signal flags (risk-assessment, platform-migration, sextortion, and others) and a trajectory score, which act as auxiliary labels rather than primary supervision. The model learns finer-grained patterns from raw conversation text directly.

08

Operational principles

A conversation-level detection model that is technically capable but operationally undisciplined will not, and should not, be deployed at scale. Five principles constrain how Aletheia is used.

Data minimization

The detection system processes conversation text in-stream. Raw message content is not retained beyond the inference call. Only feature vectors, risk scores, and the structured outputs (harmful score, predatory user, confidence, segment count) are persisted. Full content is accessible to a human reviewer only through a controlled escalation path triggered by a high-risk score.

Purpose limitation

Data processed for child-safety detection is not repurposed for advertising, recommendation, ranking, or user profiling. This is enforced by separate processing pipelines, separate access policies, and separate organizational ownership.

Proportionality

Conversations between adults are not in scope. Conversations involving a known minor and an unrelated adult get behavioral signal analysis. Linguistic analysis is invoked only when behavioral indicators cross a baseline. The result is that most users on a deploying platform are never subject to content-level review.

No autonomous enforcement

Aletheia outputs feed a prioritized human-review queue. No account action, no content removal, and no law-enforcement referral happens on the basis of the model's output alone. The model surfaces signal. Trained analysts decide what it means. This is not a hedge: it is a hard rule, both because the false-positive cost at scale is unacceptable without it and because the legal evidentiary standard for any subsequent enforcement requires human assessment.

Transparency and audit

Platforms operating Aletheia or any equivalent system should disclose it in their terms of service. Independent audits of accuracy, false-positive rates, and data handling should be conducted annually. Polycreek's commitment, as a 501(c)(3), is to operate Aletheia transparently: documenting architecture, training-corpus composition, validation results, and operational principles in artifacts like this one; reporting back to deploying partners on outcomes; and reinvesting subscription revenue into the work, not into shareholder returns, because there are no shareholders.

> The architecture above does not require breaking encryption. Behavioral analysis runs on metadata; conversational analysis runs on what the platform already sees. The framework operates inside existing privacy boundaries, not against them.

09

Closing the gap

For more than a decade, academic grooming detection has been built on PAN12, a 2012 corpus of roughly 1,200 conversations. The largest commercial trust-and-safety vendors have not built a comparable conversational layer at the scale of the underlying threat. Aletheia is Polycreek's answer.

	THE STATUS QUO	ALETHEIA
TRAINING DATA	~1,200 conversations (PAN12)	**2,523,202 conversations across 20 sources**
LANGUAGES	English only	**10+ languages, with multilingual transfer**
OUTPUTS	Binary harmful flag	**Risk score, offender attribution, eight-phase tagging**
CONVERSATION LENGTH	Truncated at single-encoder window	**Hierarchical, arbitrary length**
VOCABULARY ERA	Pre-2012	**Contemporary, validated April 2026**
VALIDATION	F1 0.85–0.90 on PAN12	**F1 0.92, AUC 0.99 on 81K-conversation held-out set**
DELIVERY	Closed academic artifact	**Nonprofit-priced API and licensed deployment**

How Aletheia is delivered

Aletheia is offered as a paid API, with subscription tiers for platforms that need integrated detection in real-time pipelines, and as a licensed deployment for child-safety organizations that need on-prem or air-gapped operation. Pricing is calibrated to use, not to extract margin. Polycreek is a 501(c)(3); every dollar of revenue funds continued development, training-data labor, the research staff who build the model, and the infrastructure that runs it.

Who Aletheia is for

Platforms that already deploy hash matching and image classification, and need a layer that sees the conversation. Child-safety organizations that triage tip-line content. Law-enforcement units that work casework and need automated assistance to prioritize. The biggest names in trust and safety could have built this. They have not. Polycreek built it because someone had to.

The conversation is where most grooming actually happens. Aletheia is one piece of the work needed to make that conversation visible to the systems that already protect children from everything else. The only thing worse than the gap that exists today is the assumption that someone else will close it.

10

References

[1] National Center for Missing & Exploited Children. *2024 in Numbers: NCMEC Releases New CyberTipline Data*. NCMEC, 2025. missingkids.org/blog/2025/ncmec-releases-new-data-2024-in-numbers

[2] Microsoft. *PhotoDNA*. microsoft.com/en-us/photodna. See also Farid, H. "An Overview of Perceptual Hashing." *Journal of Online Trust and Safety*, vol. 1, no. 1, 2021.

[3] National Center for Missing & Exploited Children. *Spike in Online Crimes Against Children a Wake-Up Call*. NCMEC, 2025. missingkids.org/blog/2025/spike-in-online-crimes-against-children-a-wake-up-call

[4] Internet Watch Foundation. *Annual Data and Insights Report 2024*. IWF, 2025. iwf.org.uk/annual-data-insights-report-2024/

[5] O'Connell, R. *A Typology of Cyber Sexploitation and Online Grooming Practices*. University of Central Lancashire, 2003.

[6] Olson, L. N., Daggs, J. L., Ellevold, B. L., & Rogers, T. K. K. "Entrapping the Innocent: Toward a Theory of Child Sexual Predators' Luring Communication." *Communication Theory*, vol. 17, no. 3, pp. 231–251, 2007.

[7] Black, P. J., Wollis, M., Woodworth, M., & Hancock, J. T. "A linguistic analysis of grooming strategies of online child sex offenders." *Child Abuse & Neglect*, vol. 44, pp. 140–149, 2015.

[8] Winters, G. M., & Jeglic, E. L. "Stages of Sexual Grooming: Recognizing Potentially Predatory Behaviors of Child Molesters." *Deviant Behavior*, vol. 38, no. 6, pp. 724–733, 2017.

[9] Whittle, H., Hamilton-Giachritsis, C., Beech, A., & Collings, G. "A review of online grooming: Characteristics and concerns." *Aggression and Violent Behavior*, vol. 18, no. 1, pp. 62–70, 2013.

[10] Kloess, J. A., Beech, A. R., & Harkins, L. "Online Child Sexual Exploitation: Prevalence, Process, and Offender Characteristics." *Trauma, Violence, & Abuse*, vol. 15, no. 2, pp. 126–139, 2014.

[11] Lorenzo-Dus, N., & Izura, C. "'cause ur special': Understanding trust and complimenting behaviour in online grooming discourse." *Journal of Pragmatics*, vol. 112, pp. 68–82, 2017.

[12] Inches, G., & Crestani, F. "Overview of the international sexual predator identification competition at PAN-2012." *CLEF Working Notes*, 2012.

[13] Bogdanova, D., Rosso, P., & Solorio, T. "Exploring High-Level Features for Detecting Cyberpedophilia." *Computer Speech and Language*, vol. 28, no. 1, pp. 108–120, 2014.

[14] Williams, R., Elliott, I. A., & Thomas, S. "Identifying sexual grooming themes used by internet sex offenders." *Deviant Behavior*, vol. 34, no. 2, pp. 135–152, 2013.

[15] Lorenzo-Dus, N., Izura, C., & Perez-Tattam, R. "Understanding grooming discourse in computer-mediated environments." *Discourse, Context & Media*, vol. 12, pp. 40–50, 2016.

[16] Broome, L. J., Izura, C., & Davies, J. "A psycho-linguistic profile of online grooming conversations." *Child Abuse & Neglect*, vol. 109, p. 104647, 2020.

[17] Gupta, A., Kumaraguru, P., & Sureka, A. "Characterizing pedophile conversations on the Internet using online grooming." *arXiv preprint arXiv:1208.4324*, 2012.

[18] Vaswani, A., Shazeer, N., Parmar, N., et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems*, 2017.

[19] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *NAACL*, 2019.

[20] He, P., Liu, X., Gao, J., & Chen, W. "DeBERTa: Decoding-enhanced BERT with Disentangled Attention." *ICLR*, 2021.

[21] Steinebach, M., et al. "An Analysis of PhotoDNA." *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2023.

[22] Internet Watch Foundation. *AI CSAM Report 2026: Harm Without Limits*. IWF, 2026. iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/

[23] Polycreek. *Aletheia Unified Dataset Build Report*. Internal documentation, 2026.