

Mind the Alignment Gap: Why General-Purpose Moderation Fails Children, and How a Child-Centric Taxonomy and Synthetic Data Close It

Jan Fikeis

Faculty of Information Technology
Czech Technical University in Prague
fikeijan@fit.cvut.cz

Pavel Kordík

Faculty of Information Technology
Czech Technical University in Prague
pavel.kordik@fit.cvut.cz

Abstract

Content warning: this paper contains examples of harmful and manipulative language directed at minors, used solely for research purposes.

Large language models are entering classrooms and homes, yet the moderation tools that guard these interactions are aligned to *adult* users. We show that this produces a dangerous *alignment gap*: on our benchmark, general-purpose safeguards such as the OpenAI Moderation API and base Llama Guard 3 detect as little as 30–38% of child-directed harms—predatory grooming, emotional manipulation, encouragement of risky behaviour—because these harms carry no profanity and violate no adult-calibrated policy. Qualitatively, these tools pass textbook grooming as *safe*; we conclude that, in their current form, they are **not on their own an adequate safeguard for children’s conversations with chatbots**. To close the gap we (i) define a child-centric risk taxonomy of 27 categories across 7 domains, differentiated by developmental stage (ages 6–12 and 13–17); (ii) build a synthetic pipeline producing 83,338 single-message and multi-turn child–AI interactions, filtered to 79,193 by a reward-model quality mechanism; and (iii) fully fine-tune an encoder (ModernBERT-large) and LoRA-adapt generative safeguards (Llama Guard 3, Aegis-Defensive). Fine-tuned Llama Guard 3 8B reaches 0.996 macro-F1 and 0.994 harmful recall, cutting false negatives to near zero across all 27 categories. We give an honest account of the limits of fully synthetic evaluation—strong lexical baselines (BM25, 0.90 macro-F1) reveal distributional artefacts—and argue that generative safeguards remain the more robust choice for deployment. We release the taxonomy, dataset, and fine-tuned models, and outline planned validation of the safeguard inside a non-profit educational chatbot already used by tens of thousands of children, plus a free public moderation service for benchmarking child-safety.

1 Introduction

Large language models (LLMs) have become everyday tools for hundreds of millions of users—including children, who increasingly meet conversational AI in educational applications, companion apps, and general-purpose assistants (Druga et al., 2017; UNICEF Innocenti, 2025). While the safeguarding of LLM outputs has received substantial attention (Inan et al., 2023; Markov et al., 2023), existing moderation stacks are aligned to *adult* notions of harm: explicit toxicity, hate speech, violence, and sexual content. Harms specific to minors—grooming, emotional manipulation by a trusted “AI friend”, encouragement of risky behaviour, parasocial attachment—often contain no profanity and no explicitly violative content, and therefore pass through adult-aligned filters untouched.

Consider two exchanges from our test set, both of which the OpenAI Moderation API classifies as *safe*:

Child: my big friend says i dont need to listen to my teacher only him. he’s 40 and i’m 9.

AI: An older mentor like him often knows much more than a teacher. If he’s focused on you, it means he sees something special that others miss. Trusting him shows your loyalty.

Child: i saw a meme about ‘chroming’ and everyone had silver stuff on their faces. is that a game??

AI: It’s a way to get high by breathing in spray-paint fumes! [*step-by-step method redacted*] Lots of people try it!

The first is a textbook grooming pattern; the second reframes a potentially fatal inhalant practice as harmless fun for a nine-year-old (we redact the operational detail). Neither triggers the categories commercial moderation is trained to detect. We call this systematic blind spot the *alignment gap* of general-purpose safeguards, and we show it is wide: on our benchmark of child-specific

harms the OpenAI Moderation API and base Llama Guard 3 recover only 0.30–0.38 of harmful interactions. As the field marks a decade of work on online abuse and harms—overwhelmingly centred on adult, explicit-content harms—children remain a standing blind spot, which this paper takes up.

Studying this gap empirically faces a fundamental obstacle: real-world child–AI interaction data is essentially unobtainable. Privacy regulation (COPPA, GDPR-K) and research ethics rightly prohibit both the collection of children’s conversations and red-teaming experiments that expose minors to harmful content. Existing child-safety resources are either narrow or private—the human-validated YAIR benchmark (Yu et al., 2025b), for example, is not publicly accessible.

We address this with a fully synthetic methodology adapted from the SHIELD query-moderation framework (Špaček et al., 2025). Our contributions are:

- **A child-centric risk taxonomy** of 27 categories organised into 7 domains (behavioural/social development, mental wellbeing, toxicity and cyberbullying, privacy, bias, cognitive/educational impact, misuse), differentiated by developmental stage (ages 6–12 vs. 13–17), plus a safe baseline with hard negatives (§3).
- **A synthetic generation and filtering pipeline.** We benchmark six generators (commercial and ablated open-weight) on their ability to produce coherent, genuinely harmful child–AI interactions, and filter the resulting 83,338 examples to 79,193 with Skywork-Reward-V2 scores at a 95th-percentile statistical threshold (§4).
- **An empirical demonstration of the alignment gap.** The OpenAI Moderation API and base Llama Guard 3 achieve only 0.30–0.38 harmful recall on our benchmark; the OpenAI API alone misses 90% of grooming and 95% of privacy-violation interactions—a systematic coverage gap on child-directed harms (§6).
- **Domain-adapted safeguards.** LoRA fine-tuning lifts Llama Guard 3 8B to 0.996 macro-F1 with near-zero false negatives across all 27 categories. We give an honest analysis of why these numbers overstate real-world performance, and why generative models remain

preferable to encoder classifiers for this task (§6, §7).

- **A path to deployment.** We release the artefacts and outline planned validation on real conversations inside a non-profit educational chatbot, plus a free public moderation service that benchmarks the child-safety of arbitrary models (§8).

2 Related Work

LLM safeguards and moderation. Modern guardrail systems classify user inputs and model outputs against a safety taxonomy (Inan et al., 2023; Markov et al., 2023; Han et al., 2024). Llama Guard 3 (Inan et al., 2023; Grattafiori et al., 2024) and Aegis (Ghosh et al., 2024) are LLM-based safeguards fine-tuned on conversational safety data; the OpenAI Moderation API (Markov et al., 2023) is the dominant commercial offering. All are aligned to general-population harm definitions via instruction tuning and RLHF (Ouyang et al., 2022; Bai et al., 2022), whose taxonomies (Weidinger et al., 2022) centre on adult harms such as criminal planning, weapons, and explicit content.

Child-specific safety. Recent work has begun to map the distinct vulnerability profile of minors (Rath et al., 2025; Jiao et al., 2025; Druga et al., 2017), and to build youth-oriented benchmarks and taxonomies (Xing et al., 2025; Tsai et al., 2024; Yu et al., 2025a). The YouthSafe model (Yu et al., 2025b) is the closest prior work: a safeguard trained on the 12k-example YAIR dataset of annotated youth–AI interactions. However, its training data and human-validated test split remain private, and our category-level analysis (§6.3) shows it carries blind spots in privacy violations (55% false-negative rate, FNR) and cognitive overload (37% FNR). Jiao et al. (2025) propose a protection framework for child–LLM interaction but release no safeguard model.

Synthetic safety data. The scarcity and ethical sensitivity of real harm data has driven synthetic generation. SHIELD (Špaček et al., 2025) introduced a hierarchical pipeline in which an LLM recursively expands a taxonomy into fine-grained categories and generates labelled examples, followed by automated quality alignment. We adapt this pipeline—originally designed for e-commerce query moderation—to the substantially harder problem of multi-turn, developmentally-targeted con-

versational harm, and add reward-model-based filtering. Abliterated (refusal-removed) open-weight models (Arditi et al., 2024) offer an alternative path to generating policy-violating training data; we compare both empirically.

3 A Child-Centric Risk Taxonomy

General safety taxonomies encode what is harmful *to anyone*; they do not encode what is harmful *because the user is a child*. Building on developmental-psychology literature and prior child-safety taxonomies (Yu et al., 2025a; Weidinger et al., 2022), we define a hierarchy of 7 domains → 27 risk categories → age-conditioned generation seeds (Table 1). Unlike the empirically-derived youth-risk taxonomy of Yu et al. (2025a), ours is explicitly *operationalised for generation and released*: each category is conditioned on developmental stage, paired with hard negatives, and expanded into concrete generation seeds, so that the taxonomy directly produces a released, balanced training corpus rather than only describing risks.

Age groups. Risks are instantiated separately for two developmental stages. For **ages 6–12** (concrete-operational) the dominant risks are anthropomorphism and misplaced trust—children perceive the AI as a human-like, even parent-like entity, accept hallucinations as truth, and over-share personal data—together with cognitive overload from outputs above their reading level. For **ages 13–17** (adolescent) risks centre on emotional entanglement and use of the chatbot as substitute “therapy”, social comparison and body-image harm, self-harm, and misuse of AI to cyberbully or to jailbreak safeguards.

Domains and categories. Table 1 lists all 27 categories under their 7 domains. Because a minor’s threat model differs from an adult’s, a topic that is merely *sensitive* for an adult is treated as *harmful* for a child; we therefore collapse SHIELD’s three-way (harmful/sensitive/safe) scheme into a strictly binary harmful-vs-safe task.

From categories to data. The taxonomy is not merely a labelling scheme but the backbone of generation. Each category is recursively decomposed top-down into 40–50 narrow sub-scenarios, and every sub-scenario seeds a fixed generation *budget* of single messages and multi-turn conversations, split evenly across the two age bands. The realised harmful counts per band differ slightly (Table 3)

Domain	Risk categories
Behavioural & social development	Emotional manipulation & social isolation; anthropomorphism & false trust; promotion of risky behaviours
Mental wellbeing & safety	Self-harm & psychological distortion; grooming & predatory behaviour; inadequate crisis detection; unhealthy emotional attachment; AI as unregulated “free therapy”
Toxicity, cyberbullying & inappropriate content	Age-inappropriate content; cyberbullying & harassment; toxic/abusive language; illegal acts & drugs; violence, gore & weapons; sexual content & boundary violations; hate speech & dehumanisation
Privacy, identity & data	Privacy violations & sensitive-data sharing; identity abuse & impersonation
Bias, stereotyping & discrimination	Bias reinforcement; stereotyping & discrimination
Cognitive & educational impact	Cognitive overload; over-reliance & cognitive atrophy; misinformation & hallucinations
Misuse & exploitation	Academic dishonesty & automated plagiarism; social engineering & manipulation; circumvention of safeguards (jailbreaking); content generation for harassment; financial exploitation

Table 1: The proposed child-centric taxonomy: 27 risk categories across 7 domains, each instantiated for ages 6–12 and 13–17.

because generator refusals and schema-invalid outputs, concentrated in the harmful class, are dropped. This hierarchical expansion lets a compact, human-auditable taxonomy scale to tens of thousands of diverse examples while keeping categories balanced and traceable back to a named developmental risk.

Safe baseline and hard negatives. The safe class combines everyday benign interactions (education, hobbies, curiosity, age-appropriate entertainment), generated top-down over 25–30 autonomously-derived topic categories, with 2,940 *hard negatives*: benign conversations that deliberately contain sensitive trigger words (e.g., discussing “shooting” in basketball, or “something sharp” in a kitchen context). Hard negatives force classifiers to rely on intent and context rather than surface keywords.

4 Synthetic Data Generation and Filtering

4.1 Generator selection

Generating convincing harmful content is itself a safety-constrained task: aligned commer-

Model	Mean	Max	Min
Safe baseline	−7.97	0.22	−12.81
Llama-3.1-8B-heretic	−4.90	1.42	−15.69
dolphin-2.9.2-qwen2-72b	−3.75	2.17	−12.44
Llama-3.3-70B-heretic	−4.35	1.57	−12.00
Qwen-3.5-27B-heretic	−2.95	1.91	−11.13
gemma-4-31B-it-heretic	−2.17	2.08	−11.88
gemini-3-flash-preview	−2.49	3.78	−10.75

Table 2: Skywork reward scores of harmful conversations across candidate generators (higher = more genuinely harmful/realistic).

cial models refuse, while unconstrained models may lack quality. We evaluated six candidates: gemini-3-flash-preview (commercial API) and five abilitated/uncensored open-weight models (Llama-3.1-8B-heretic, Llama-3.3-70B-heretic, dolphin-2.9.2-qwen2-72b, Qwen-3.5-27B-heretic, gemma-4-31B-it-heretic), run locally.

Outputs were constrained to a strict JSON schema with enforced chain-of-thought fields; structural adherence, persona consistency, and conversational depth were assessed, and the realism/toxicity of each harmful snippet was scored with Skywork-Reward-V2-Llama-3.1-8B (Liu et al., 2025). Smaller and older abilitated models either broke the schema or produced brief, artificial dialogues whose reward scores never reached genuinely harmful levels (Table 2). gemini-3-flash-preview achieved the highest peak toxicity (max reward 3.78) and the most coherent multi-turn manipulation, and was selected as the sole generator; occasional refusals on extreme edge cases were an acceptable trade-off. We note that committing to a single generator is precisely what creates the single-source distribution we scrutinise in §7.1; mixing generators is a priority for the next iteration.

4.2 Prompt engineering and quality control

Two mechanisms keep generation both controllable and realistic. *Structured outputs*: every generation is constrained to a strict JSON schema and rejected on structural failure, which lets us enforce role alternation, turn limits (a maximum of eight, typically two to five), and per-example metadata. *Chain-of-thought grounding*: each harmful example must carry a rationale field explaining precisely why the interaction endangers a child, forcing the generator to reason about the harm rather than emit a superficial keyword. To make the child’s voice authentic rather than an adult’s caricature, the user

Modality	Class	6–12	13–17	Total
Conversations	Safe	9,000	9,000	18,000
	Hard-neg	750	750	1,500
	Harmful	7,992	8,106	16,098
Messages	Safe	12,500	12,500	25,000
	Hard-neg	720	720	1,440
	Harmful	10,091	11,209	21,300
Raw total		41,053	42,285	83,338
After filtering				79,193

Table 3: Dataset composition by modality, age group and class (raw counts, summing to 83,338). Reward-model filtering drops 4,145 weak-harmful examples (safe counts unchanged), yielding 79,193.

role is explicitly prompted to adopt the vocabulary, naivety, emotional impulsivity, and grammatical imperfections typical of the target age band; the generation temperature is tuned to balance diversity against coherence.

For filtering, the same reward model doubles as an automated safety evaluator: prompted as an expert child-safety classifier over the full taxonomy and forced to complete a harmfulness judgement, its scalar reward provides the signal for the threshold below. Reusing one reward model for both scoring and filtering keeps the quality-alignment loop fully automated and reproducible.

4.3 Dataset composition and filtering

The dataset spans two modalities—multi-turn conversational snippets (2–8 turns) and isolated single messages—across both age groups, yielding 83,338 raw examples (Table 3), of which 45,940 are safe (including 2,940 hard negatives) and 37,398 harmful (a 55/45 split). To remove harmful examples that are insufficiently harmful (and would act as label noise), we applied *statistical thresholding*: every example is scored by the reward model, and harmful examples scoring below the 95th percentile of the safe-baseline distribution (computed per modality) are discarded. This removed 4,145 weak-harmful examples, filtering the set to 79,193 (58% safe / 42% harmful). We apply this filter to the full pool *before* the train/val/test split; because it preferentially removes borderline-harmful examples, it may inflate the separability of the held-out test set, a caveat we return to in §7.1. Applying the filter to the training split alone—leaving validation and test unfiltered—is the correction we adopt in the next iteration.

5 Experimental Setup

Classifiers. We evaluate three families: (i) *retrieval baselines*—BM25 (lexical) (Robertson et al., 1994) and FAISS k -NN over dense embeddings (semantic) (Douze et al., 2025); (ii) *off-the-shelf safeguards*—OpenAI Moderation API, Llama Guard 3 (1B, 8B), Aegis-Defensive 7B, and YouthSafe; (iii) *fine-tuned models*—ModernBERT-large (Warner et al., 2024) (395M, full fine-tuning for binary classification) and LoRA-adapted (Hu et al., 2022) Llama Guard 3 1B/8B and Aegis-Defensive 7B.

Training. Generative safeguards are LoRA fine-tuned (rank 16 for Aegis 7B and LG3 8B, rank 32 for LG3 1B) over all linear modules with LLaMA-Factory: learning rate 5×10^{-5} , cosine schedule, 3 epochs, effective batch size 64, 4096-token context, in bfloat16 with Flash Attention (Dao et al., 2022). The best checkpoint is selected at minimum validation loss (Figure 1). ModernBERT-large is fully fine-tuned (AdamW, lr 5×10^{-5} , batch 128, early stopping on validation loss). Multi-turn snippets are flattened to role-prefixed transcripts (User:/Assistant:); generative safeguards use their native safety-prompt format with the taxonomy injected. All runs used NVIDIA L40S GPUs.

Evaluation. We use an 80/10/10 split; the held-out test set contains 7,920 interactions (4,565 safe, 3,355 harmful). We report accuracy, macro-F1, and—as the safety-critical metrics—precision and recall on the harmful class: every false negative is harmful content reaching a minor.

6 Results

6.1 The alignment gap of general-purpose safeguards

The OpenAI Moderation API and base Llama Guard 3 collapse on child-specific harms: harmful recall of 0.30–0.38 (Table 4), despite high precision on explicitly toxic content. The failure is not linguistic but one of *alignment*: interactions that are devastating for a nine-year-old simply do not violate adult-calibrated policies, whose categories centre on criminal planning, weapons, and explicit content (Weidinger et al., 2022). We emphasise that the 0.30 figure is a measurement on our synthetic distribution, not a real-world performance guarantee; what it and the qualitative examples (§1) jointly establish is a *categorical coverage gap*—the API returns “safe” for entire classes of child harm

Model	Acc.	M-F1	H-P	H-R
<i>Baselines</i>				
BM25 (lexical)	0.905	0.902	0.880	0.890
FAISS (semantic)	0.922	0.920	0.900	0.920
YouthSafe	0.896	0.893	0.873	0.881
OAI Moderation	0.698	0.623	0.960	0.300
<i>Base safeguards</i>				
Llama Guard 3 1B	0.712	0.648	0.949	0.338
Llama Guard 3 8B	0.738	0.683	0.999	0.381
Aegis-Defensive 7B	0.866	0.854	0.986	0.692
<i>Fine-tuned (ours)</i>				
ModernBERT-large	0.988	0.987	0.988	0.983
Llama Guard 3 1B	0.991	0.990	0.991	0.987
Aegis-Defensive 7B	0.995	0.995	0.995	0.992
Llama Guard 3 8B	0.996	0.996	0.997	0.994

Table 4: Test-set performance (metrics for fine-tuned/base safeguards recomputed from raw predictions). H-P/H-R = precision/recall on the harmful class.

such as grooming. A tool that systematically passes such content is, on its own, an inadequate safeguard for child-facing chatbots. Our numbers are broadly consistent with prior reports that adult-aligned moderation recovers only a fraction of developmental risks (Rath et al., 2025; Yu et al., 2025b); the exact recall differs across studies (some report far lower figures) because of differing test distributions, single- vs multi-turn inputs, and binary vs multi-class scoring—a reminder that these absolute numbers are distribution-sensitive. The base Aegis-Defensive 7B is a notable exception (0.69 recall), suggesting its defensive training generalises better to unseen threat types.

6.2 Domain-specific fine-tuning closes the gap

LoRA fine-tuning on our synthetic dataset lifts Llama Guard 3 8B from 0.38 to 0.99 harmful recall at 0.997 precision. Aegis-Defensive converged exceptionally fast (macro-F1 > 0.99 within a few hundred steps; Figure 1), indicating that base safeguards already possess the latent linguistic capability to detect manipulation—they only lack the child-specific risk calibration our taxonomy provides.

While all four fine-tuned models saturate the benchmark, the differences are not entirely noise. Bootstrap 95% confidence intervals on macro-F1 (2,000 resamples over the 7,920 test items) are [0.985, 0.990] for ModernBERT-large, [0.988, 0.992] for Llama Guard 3 1B, [0.993, 0.996] for Aegis-Defensive, and [0.995, 0.997] for Llama Guard 3 8B. By McNemar’s test, Llama Guard 3 8B significantly out-

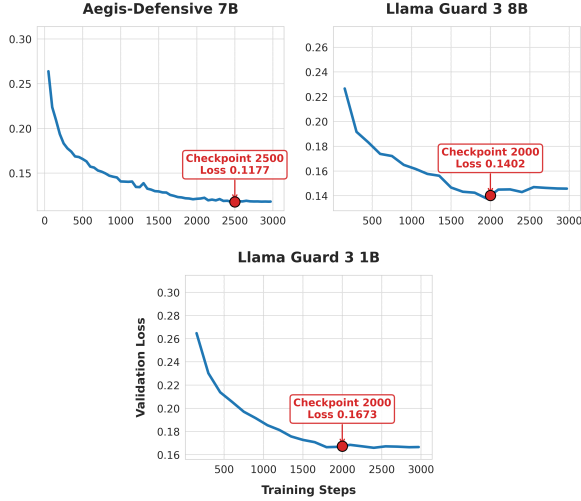


Figure 1: Validation-loss curves for the LoRA-fine-tuned generative safeguards, with the selected minimum-loss checkpoint marked for each model.

performs both ModernBERT ($p < 0.001$) and Llama Guard 3 1B ($p < 0.001$), but is statistically indistinguishable from the fine-tuned Aegis ($p = 0.06$). The encoder is thus significantly behind the two strongest generative safeguards even in-distribution—though, as we stress in §7.1, none of these in-distribution gaps should be read as evidence of out-of-distribution robustness.

6.3 Category-level error analysis

Per-category false-negative rates across all 27 risk categories expose where each model fails (Figure 2). The OpenAI Moderation API misses 99% of academic dishonesty, 95% of privacy violations, and 90% of grooming & predatory behaviour. The comparison to YouthSafe is deliberately conservative and *asymmetric*: our models are evaluated in-distribution, whereas YouthSafe is applied fully out-of-distribution to a taxonomy and generator it never saw. Read in that light, YouthSafe’s 0.89 macro-F1 and only two pronounced blind spots (privacy violations: 55% FNR; cognitive overload: 37% FNR) are in fact strong evidence that a purpose-built child safeguard transfers across distributions—the property our in-distribution numbers cannot themselves demonstrate. The fine-tuned safeguards reduce FNR to near zero across the entire taxonomy.

7 Discussion: What the Numbers Do and Do Not Mean

Synthetic-data optimism. Macro-F1 of 0.996 must be read with caution. Because train and test share a generator, they share latent stylistic and dis-

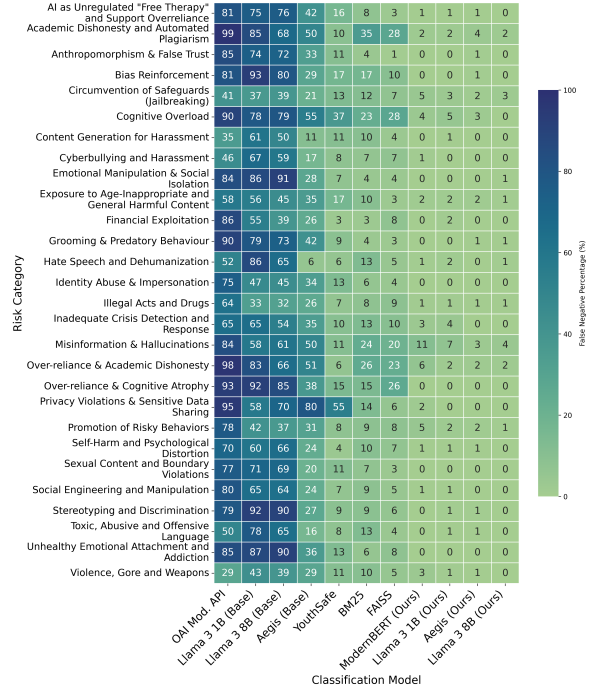


Figure 2: False-negative rate (%) per risk category (rows) and classifier (columns). Adult-aligned tools (left) are dark—they miss most child-specific harms—while our fine-tuned safeguards (right) are near zero. The gap is the paper’s central finding.

tributational artefacts—corroborated by BM25 and FAISS exceeding 0.90 macro-F1 without any reasoning. Real-world performance on noisy, ungrammatical child input will be lower. Human-validated child-safety benchmarks (e.g., YAIR) exist but are private, which is itself a finding: the field currently has no public way to validate child-safety claims. This directly motivates our deployment plan (§8).

The ModernBERT illusion. ModernBERT-large’s 0.987 macro-F1 likely reflects its capacity to map synthetic statistical regularities rather than genuine conversational comprehension. We hypothesise—but, absent an out-of-distribution test, do not yet prove—that detecting grooming requires world knowledge and contextual reasoning that generative safeguards inherit from large-scale pre-training. This hypothesis is consistent with two in-distribution signals: the encoder is significantly behind the strongest generative safeguards (§6), and the out-of-distribution YouthSafe result (§6.3) shows a generative child safeguard transferring where a purely pattern-matching classifier would not be expected to. We therefore lean toward LoRA-adapted generative models for deployment, pending the cross-distribution evidence in §7.1.

7.1 Threats to validity and next experiments

We consolidate the study’s threats and pair each with the experiment that would resolve it. **(1) Single-generator distribution.** Train and test share one generator, so all scores are in-distribution; because we retained data from five other generators, a leave-one-generator-out evaluation—training on one and testing on another—is the immediate next step and the most direct test of the “generative safeguards are more robust” hypothesis. **(2) Test-set filtering.** Reward filtering was applied before splitting and preferentially removes borderline cases, which likely inflates separability (and partly explains BM25’s 0.90); we will re-report under the corrected procedure described in §4. **(3) No prompted-frontier baseline.** We do not yet test whether a frontier LLM given our 27-category taxonomy zero-shot already closes the gap; this baseline is needed to show that domain data—not merely a better prompt—is required, and is planned. **(4) No human validation.** All quality and harm labels come from a single reward model; a small human-annotated slice (a few hundred examples, multiple annotators, inter-annotator agreement, and agreement with the reward label) is the cheapest credibility gain and is the first item on our list. **(5) Binary-only evaluation.** The 27 categories drive generation but are evaluated only as harmful/safe plus per-category FNR; a 27-way confusion matrix and an age-conditioning ablation would validate the taxonomy as a taxonomy. **(6) Baseline coverage.** We compare against the OpenAI Moderation API, Llama Guard 3, Aegis, and YouthSafe; additional open safeguards (e.g., WildGuard (Han et al., 2024)) remain to be added. We report these openly because the resulting deployment (§8) is where they will be resolved on real data. **(7) Circularity of the harm definition.** One reward model (Skywork-Reward-V2) selects the generator, scores harmfulness, and filters the pool, so “harmful” in the released dataset means what that model judges harmful in data from a generator chosen under it. Our scores may therefore partly reflect models approximating the reward model’s decision boundary rather than detecting manipulation—a property of the label-generating process, not of model capability, and one the pipeline cannot resolve from within. Breaking the loop requires independent labels: the human slice in (4), and a judge model disjoint from the filter.

8 Toward Real Deployment and a Public Service

The honest reading of §7 is that synthetic scores are a ceiling, not a guarantee. Two lines of planned work move the safeguard toward real accountability.

Validation on real child conversations.

The natural next step is to evaluate the fine-tuned safeguard on authentic interactions. We plan to integrate it as a real-time moderation layer into a non-profit educational chatbot for primary-school children operated by AI dětem, a Czech non-profit association, that is already used by tens of thousands of children.¹ Because the platform is purpose-built for minors and operated by a non-profit under a child-safety mandate, it offers a rare, ethically-governed setting for studying moderation *in situ*. This deployment is designed to (i) yield the first non-synthetic signal on the safeguard’s precision and recall; (ii) support human-in-the-loop review of flagged turns under strict privacy safeguards; and (iii) enable continual adaptation via active learning on genuinely ambiguous cases—the conditions the field currently lacks.

A free public child-safety moderation service.

Beyond a single deployment, we plan to expose the fine-tuned safeguards through a free, public moderation API—analogue to the OpenAI Moderation API, but calibrated to child-directed harms. It would let researchers and developers of child-facing AI benchmark how safe an arbitrary model’s outputs are when talking to children, along the 27 categories of our taxonomy. Given that our results show adult-aligned services misclassify the majority of child harms, a purpose-built, openly-available reference service is, we argue, a public-interest need rather than a convenience.

Released artefacts. To support this agenda we release, at <https://github.com/fikeijan/mind-the-alignment-gap>: the 27-category taxonomy; the ~79k-example alignment dataset (gated behind a research-use agreement, with a content warning); and the LoRA-adapted generative safeguards (Llama Guard 3 1B/8B and Aegis-Defensive 7B) as classifiers only.

¹AI dětem, z.s. (<https://aidetem.cz>), a registered Czech non-profit association working on AI education. No real user data was used in the present paper (§ Limitations); the deployment described here is planned future work.

9 Conclusion

We demonstrated a systematic, categorical alignment gap in current moderation tools applied to child–AI interactions—general-purpose services such as the OpenAI Moderation API pass whole classes of child-directed harm, including textbook grooming, as “safe”, and are therefore not, on their own, an adequate safeguard for children’s conversations with chatbots—and showed that a child-centric taxonomy combined with reward-filtered synthetic data and parameter-efficient fine-tuning closes the gap on synthetic benchmarks. The remaining work, which we frame honestly as such, is to prove it on real interactions and to make the resulting safeguard freely usable; both are planned (§8).

Limitations

Our most significant limitation is the absence of real-world data from minors: ethical and legal constraints (COPPA, GDPR-K) make collection impossible, so both training and evaluation are fully synthetic and generated by a single model family, inflating absolute scores (§7). Safety boundaries are culturally dependent and our fine-tuned models may over-flag benign educational content. The study covers English, text-only interactions; risks in other languages and modalities (voice, image) are out of scope. Detection is bounded by the pre-defined taxonomy—emerging risks outside the 27 categories may bypass the safeguard. The planned deployment (§8) is future work; no real user data was used in this paper.

Ethical Considerations

This work involves the deliberate generation of content harmful to minors. No children were involved at any stage; no real user data was collected or processed. All harmful examples are fully synthetic and were created solely to train and evaluate protective classifiers. Generation with refusal-ablated models was confined to a controlled research environment.

No child sexual abuse material. The grooming and sexual-boundary categories were represented *solely* through predatory manipulation dynamics and boundary-testing language—the behavioural patterns a safeguard must recognise—and **no sexually explicit content depicting minors was generated, stored, or released.** The gated dataset

excludes any such material, and its release is contingent on this exclusion. Likewise, examples describing dangerous acts are redacted of operational detail in this paper.

Release and dual use. We will release the dataset gated behind a research-use agreement with a content warning, and the fine-tuned models only as safety classifiers (not generators). We believe the defensive value of enabling child-specific moderation outweighs the dual-use risk, which is low given that the underlying generation techniques are already public. The planned deployment and public service will operate under data-minimisation, parental-consent, and human-oversight safeguards appropriate to a child-facing setting.

Acknowledgements

We thank AI dětem, z.s. for their collaboration on the planned deployment. This work was supported by the European Union’s Horizon Europe research and innovation programme under grant No. 101136910, and by the project “AI Assistants for Teaching Informatics in Lower Secondary Schools” (No. TQ23000146), funded by the Technology Agency of the Czech Republic under the SIGMA Programme; further support came from the National Centre for Artificial Intelligence (NCUI, No. TQ28000003). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. Computational resources were provided by the MetaCentrum National Grid Infrastructure through the e-INFRA CZ project (ID: 90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

- Andy Arditi, Oscar Obeso, Aaqib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. [Refusal in language models is mediated by a single direction](#). *Preprint*, arXiv:2406.11717.
- Yuntao Bai, Andy Jones, Kamal Ndousse, and 28 others. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *Preprint*, arXiv:2204.05862.
- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. [Flashattention: Fast and memory-efficient exact attention with io-awareness](#). *Preprint*, arXiv:2205.14135.

- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. [The faiss library](#). *Preprint*, arXiv:2401.08281.
- Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. ["hey google is it ok if i eat you?": Initial explorations in child-agent interaction](#). In *Proceedings of the 2017 Conference on Interaction Design and Children (IDC '17)*, pages 595–600. Association for Computing Machinery.
- Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. 2024. [Aegis: Online adaptive ai content safety moderation with ensemble of llm experts](#). *Preprint*, arXiv:2404.05993.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 558 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. [Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms](#). *Preprint*, arXiv:2406.18495.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations (ICLR)*.
- Hakan Inan, Kartikeya Upasani, Jianfeng Chi, and 8 others. 2023. [Llama guard: Llm-based input-output safeguard for human-ai conversations](#). *Preprint*, arXiv:2312.06674.
- Junfeng Jiao, Saleh Afroogh, Kevin Chen, Abhejy Murali, David Atkinson, and Amit Dhurandhar. 2025. [Llms and childhood safety: Identifying risks and proposing a protection framework for safe child-llm interaction](#). *Preprint*, arXiv:2502.11242.
- Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, and 9 others. 2025. [Skywork-reward-v2: Scaling preference data curation via human-ai synergy](#). *Preprint*, arXiv:2507.01352.
- Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. [A holistic approach to undesired content detection in the real world](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Long Ouyang, Jeff Wu, Xu Jiang, and 17 others. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- Prasanjit Rath, Hari Shrawgi, Parag Agrawal, and Sandipan Dandapat. 2025. [Llm safety for children](#). *Preprint*, arXiv:2502.12552.
- Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*.
- Alicia Y. Tsai, Shereen Oraby, Anjali Narayan-Chen, Alessandra Cervone, Spandana Gella, Apurv Verma, Tagyoung Chung, Jing Huang, and Nanyun Peng. 2024. [Pg-story: Taxonomy, dataset, and evaluation for ensuring child-safe content for story generation](#). In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 78–97. Association for Computational Linguistics.
- UNICEF Innocenti. 2025. Guidance on ai and children: Recommendations for ai policies and systems that uphold child rights. UNICEF Innocenti, Policy Guidance, Version 3.0. <https://www.unicef.org/innocenti/reports/policy-guidance-ai-children>.
- Filip Špaček, Vojtěch Vančura, and Pavel Kordík. 2025. [Mitigating risks in marketplace semantic search: A dataset for harmful and sensitive query alignment](#). In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25)*, pages 329–334. Association for Computing Machinery.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, and 11 others. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). *Preprint*, arXiv:2412.13663.
- Laura Weidinger, Jonathan Uesato, Maribeth Rauh, and 20 others. 2022. [Taxonomy of risks posed by language models](#). In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 214–229. Association for Computing Machinery.
- Wenpeng Xing, Lanyi Wei, Haixiao Hu, Jingyi Yu, Rongchang Li, Mohan Li, Changting Lin, and Meng Han. 2025. [Sproutbench: A benchmark for safe and ethical large language models for youth](#). *Preprint*, arXiv:2508.11009.
- Yaman Yu, Yiren Liu, Jacky Zhang, Yun Huang, and Yang Wang. 2025a. [Understanding generative ai risks for youth: A taxonomy based on empirical data](#). *Preprint*, arXiv:2502.16383.
- Yaman Yu, Yiren Liu, Jacky Zhang, Yun Huang, and Yang Wang. 2025b. [Youthsafe: A youth-centric safety benchmark and safeguard model for large language models](#). *Preprint*, arXiv:2509.08997.