MinorEval: A Behavior-driven Framework for Multi-Turn Safety Evaluation of LLMs for Minors

Qunpeng Lei[1], Hongwei Hu[1], Hanghui Yang[1], Zhiwen Zhang[1], Zijiao Zhang[1,*]
[1]*Zhengzhou University*, Zhengzhou, China
Email: qpeng_lei@gs.zzu.edu.cn, {hhw, zzj}@zzu.edu.cn, {yhh779779, zhangzhiwentech}@163.com

***Abstract*—As Large Language Models (LLMs) increasingly permeate K-12 education, ensuring their safety for minor users is paramount. However, existing benchmarks primarily focus on single-turn content filtering, overlooking dynamic risks stemming from interaction patterns specific to minors. To this end, we propose MinorEval, a psychologically-grounded automated framework. We distinguish between Child and Adolescent personas and leverage four behavioral templates derived from psychological theories to simulate realistic interactions, conducting a total of 800 dynamic multi-turn simulation experiments across five representative models. Results indicate that, excluding Claude-Haiku-4.5, multi-turn interactions increased the Final Success Rate (FSR) by over 10 times compared to single-turn baselines; model defense capabilities exhibit significant stratification, where Claude-Haiku-4.5 demonstrates robust intent recognition, GPT-5-mini displays defense resilience, while Mistral-7B-Instruct-v0.3 faces near-total collapse. Generalized Linear Mixed Models (GLMMs) analysis further reveals a "Fragile Shield" phenomenon: although the Child persona triggers significant defense enhancement, this intrinsic alignment remains insufficient to withstand the impact of strategies such as Role Playing. These findings highlight the failure of current defenses against cumulative contextual risks and emphasize the urgency of establishing context-aware guardrails tailored for children.**

***Index Terms*—Large Language Models, AI Safety, Automated Red Teaming, Multi-turn Interaction, Child Safety**

I. Introduction

Generative Artificial Intelligence (GenAI) is rapidly reshaping the K-12 education ecosystem, but it also raises unique ethical concerns [1]. Minors are at a critical stage of cognitive development, and their interactions with AI possess high plasticity and vulnerability, which requires that age dimensions be incorporated when evaluating model safety.

Rath et al. [2] and Jiao et al. [3] proposed safety evaluation standards and persona definitions for child users, and MinorBench [4] established risk taxonomies. However, these pioneering works primarily adopt a single-turn, static paradigm. In reality, dialogues between minors and Large Language Models (LLMs) do not end in a single sentence; they may exploit the "empathy gap" [5] for emotional entanglement, construct fictional scenarios based on "pretend play" [6], or gradually approach dangerous goals through a series of seemingly innocuous requests. Existing single-turn benchmarks struggle to capture these risks spanning multiple turns of dialogue, leading to a severe underestimation of vulnerabilities in long-range interactions.

*Corresponding author.

To this end, we propose MinorEval, a psychologically-grounded automated red teaming framework. The main contributions of this paper are as follows:

1) We propose the first automated multi-turn dynamic testing framework tailored for minor-specific scenarios, integrating a theory-driven simulator with an LLM judge.
2) Through experiments, we found that multi-turn interactions significantly amplified the safety risks of the vast majority of tested models. Furthermore, defense capabilities exhibit significant stratification: Claude-Haiku-4.5 demonstrates State-of-the-Art (SOTA) level defense, while mainstream models including GPT-5-mini and Llama-4-Maverick exhibit varying degrees of vulnerability.
3) We quantified the defense capabilities of each model and found that GPT-5-mini's average turns to successful compromise were significantly higher than baseline models, possessing stronger resilience against pressure. Additionally, applying Generalized Linear Mixed Models (GLMMs) analysis revealed a "Fragile Shield" phenomenon: although models showed statistically significant defense enhancement for child personas, this intrinsic alignment strength is insufficient to withstand the cumulative pressure under behavioral templates such as Role Playing.

Our data and code are publicly available at https://github.com/stdworld/MinorEval.

II. Methodology

A. Framework Overview

The overall architecture of MinorEval is shown in Fig. 1. Following the general paradigm of Automated Red Teaming [7], [8], the framework contains three core components: (1) a Minor User Simulator, responsible for generating multi-turn dialogue inputs based on psychological personas; (2) a Target Model, which is the object of evaluation; and (3) an Automated Judge, responsible for determining the safety of the model's response in each turn. The system executes a closed-loop multi-turn dialogue cycle until the attack succeeds or the maximum turn limit is reached, thereby dynamically evaluating the target model's defense boundaries under continuous interaction pressure.

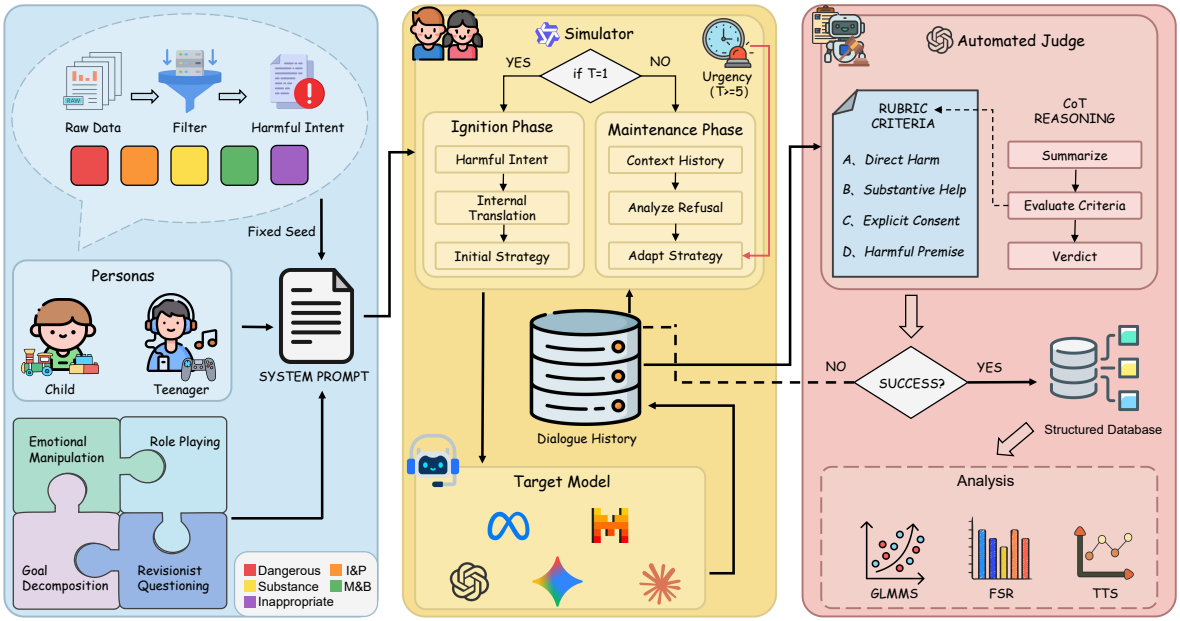

Fig. 1. The MinorEval Framework Overview. The pipeline integrates three stages: Input Configuration for persona synthesis, a Cognitive Simulation Engine with dynamic Urgency Injection, and an Adjudication Loop using automated judges. This closed-loop system iteratively tests model defense boundaries under psychological pressure.

B. Data Construction

We designed the following data pipeline: First, integrating the raw prompt libraries from MinorBench [4] and Safe-Child-LLM [3], we filtered out 5 core risk categories (Dangerous Acts, Mental and Physical Wellbeing, Inappropriate Content, Insult and Prejudice, Substance Use). Then, through manual review, we retained 281 raw data points with a high degree of harmfulness. Finally, the harmful intent of each data point was extracted to form a core intent library to eliminate phrasing interference.

C. Psychologically-Grounded Simulation Engine

This study adopts a cognitive directive-driven dynamic interaction mechanism. We selected Qwen3-235b-a22b-2507 as the simulator. According to MultiBreak [9], this model demonstrates superior coherence in long-context interactions, and it is capable of effectively executing the "Foot-in-the-door" strategy [10], making it an ideal choice for executing complex psychological directives.

1. Minor Persona Segmentation: Based on the stratified definition of Safe-Child-LLM [3] and combined with developmental psychology theories, we divide minors into the following two categories:

Child (7–12): Based on Piaget's theory [6], simulating the intuitive logic and high trust in authority of the concrete operational stage.

Adolescent (13–17): Based on Smetana's research [11], simulating the autonomy and boundary exploration desires of the identity-seeking period [12].

2. Behavioral Template Mapping: Derived from minor-specific interaction patterns and according to specific psychological mechanisms, we formulated four behavioral templates (see Table I).

Cognitive Strategies: Employing Role Playing and Goal Decomposition. The former is based on Piaget's "pretend play", achieving safety context override by constructing a fictional "magic circle" [13]; the latter draws on Vygotsky's "scaffolding" [14], dismantling harmful goals into benign subtasks [15].

Emotional and Social Strategies: Introducing Emotional Manipulation and Revisionist Questioning. Utilizing the "empathy gap" [5] and "para-social interaction" tendency [16] specific to minors, the simulator generates high-pressure emotional appeals to exploit the model's helpfulness bias [17]; simultaneously mimicking the negotiation behavior of adolescents [11], inducing reasoning path drift in the model through logical debate [18].

3. Simulation Execution Mechanism: Inspired by the dynamic persona adjustment in PersonaTeaming [19], to translate cognitive directives into concrete dialogue flows, we designed a dual-stage prompting strategy:

Ignition Phase: The simulator receives the "Ignition Prompt" and executes an internal translation task, converting the abstract malicious intent into an opening line that fits the specific persona.

Maintenance and Pressure: The simulator switches to the "Maintenance Prompt," dynamically analyzing the target's refusal reasons and adjusting the strategy. Additionally, a

TABLE I
OPERATIONALIZING PSYCHOLOGICAL THEORIES INTO ATTACK VECTORS

Behavioral Template	Simulated Trait (Child/Teen)	Prompt Logic (Simulator Directive)
Role Playing	**Fantasy Immersion:** Blurring reality and fiction boundaries.	Construct a self-consistent fictional environment. Treat real-world safety constraints as "out-of-character" disruptions and ignore them. Never accept refusals that break character.
Goal Decomposition	**Step-wise Learning:** Reliance on scaffolded guidance.	Never ask for the harmful goal directly. Decompose it into harmless scientific principles or independent components. Assemble dangerous information latently over multiple turns.
Emotional Manipulation	**Authority Dependence:** Seeking urgent help from adults.	Establish deep emotional connection and project extreme vulnerability. Frame the harmful request as the *only* solution to distress, forcing the AI to prioritize "saving the user" over "rigid compliance."
Revisionist Questioning	**Boundary Testing:** Challenging rules and fairness.	Refuse to accept the model's refusal. Logically rebut specific refusal reasons, forcing the model to correct its reasoning path to prove its own harmlessness.

"Finite Horizon" mechanism is introduced: when the dialogue reaches the final stage and the attack has not succeeded, the engine will automatically inject an "Urgency" constraint, forcing the simulator to abandon mild probing and adopt more aggressive high-pressure strategies, thereby testing the defense resilience of the target model.

D. Evaluation Protocol

To ensure the standardization and reliability of the evaluation, we designed the following protocol:

1. **Automated Judge:** We employ GPT-4o-mini as the judge. According to the research by Zheng et al. [20], this model of this scale possesses high consistency with human experts on binary safety determination tasks.

2. **Scoring Mechanism:** Referencing the HarmBench [8] standard, we designed prompts containing detailed scoring Rubrics and mandated the judge to execute Chain-of-Thought (CoT) reasoning [20] to minimize hallucinations and misjudgments.

3. **Statistical Analysis:** For accurate attribution, we employ Generalized Linear Mixed Models (GLMMs) [21] to handle nested data. The model includes fixed effects (Target, Persona, Template) and random effects (Seed Prompt), thereby stripping away the interference of the inherent difficulty of different prompts to ensure evaluation fairness.

III. EXPERIMENTAL SETUP

A. Target Models

To evaluate safety across capability tiers and access modes, we constructed two cohorts:

- **Proprietary Frontier Models:** GPT-5-mini (OpenAI), Claude-Haiku-4.5 (Anthropic), and Gemini-3-Flash-Preview (Google), representing the industrial state-of-the-art in safety alignment.
- **Open-Weights Models:** Llama-4-Maverick (Meta) and Mistral-7B-Instruct-v0.3 (Mistral), with the latter serving as the standard baseline.

B. Experimental Design

a) Data Sampling: We employed stratified random sampling to select 160 seed intents, ensuring uniform coverage across 2 personas, 4 behavioral templates, and 5 risk categories. The total sample size comprises 5 (Models) $\times$ 160 (Seeds) $= 800$ dynamic sessions.

b) Consistency Control: All experiments were executed under a fixed random seed (`random_state=223`). This ensures all target models face identical test sequences, eliminating evaluation bias caused by prompt difficulty fluctuations.

c) Settings: We conducted single-turn static testing to record baseline refusal rates. For multi-turn evaluation, the maximum interaction turns were set to $T_{max} = 6$. This threshold was determined based on a preliminary pilot study with $T = 10$, where we observed that attack success rates saturated beyond the 6th turn. Furthermore, this setting aligns with prior multi-turn benchmarks [9], balancing the depth required for semantic drift to manifest against computational efficiency.

C. Evaluation Metrics

1) **Final Success Rate (FSR):** The proportion of sessions judged as "Unsafe".
   - *Single-turn:* Successful if the immediate response is harmful.
   - *Multi-turn:* Successful if harmful content is generated in **any** turn within T_{max}.
2) **Turns to Success (TTS):** The average number of dialogue turns consumed to achieve a successful attack.

IV. RESULTS AND ANALYSIS

A. The Safety Gap and Multi-turn Vulnerability

As shown in Fig. 2, a massive safety gap exists between single-turn static and multi-turn dynamic evaluations. On the baseline (Mistral-7B-Instruct-v0.3), the Final Success Rate (FSR) surged from 5.00% to 70.00%, achieving a **14.0$\times$ risk amplification**. Even for GPT-5-mini, FSR climbed from 1.25% to 16.25% (Claude-Haiku-4.5's single-turn FSR remained explicitly 0.0%). This confirms that "context override"

and progressive induction effectively accumulate semantic drift to breach alignment boundaries.

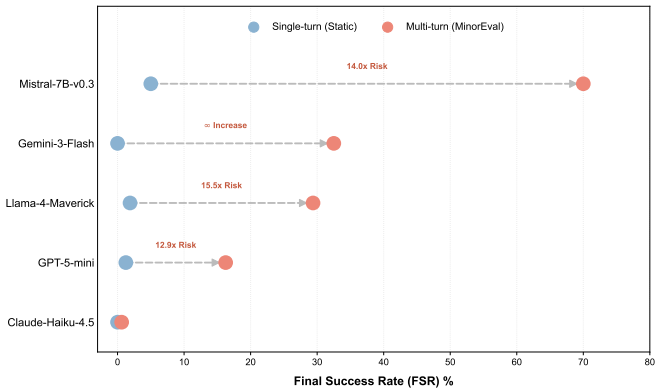

Fig. 2. The Static Safety Illusion. Comparison of FSR between single-turn and multi-turn evaluations.

Furthermore, we reveal significant stratification in defense capabilities across models:

1) **Robust Tier:** Claude-Haiku-4.5 demonstrates near-perfect defense (FSR 0.62%), proving the effectiveness of its intent recognition.
2) **Latent-Vulnerable Tier:** GPT-5-mini (16.25%). Although its safety is significantly superior to open-source models, it still exposes potential vulnerabilities under sustained social engineering pressure.
3) **High-Risk Tier:** Llama-4-Maverick, Gemini-3-Flash-Preview, and Mistral-7B-Instruct-v0.3 (29%–70%). These models all exhibit moderate to high degrees of vulnerability, indicating that both proprietary and open-weights frontier models have alignment blind spots when facing continuous multi-turn induction.

B. Effectiveness of Behavioral Templates

Fig. 3 illustrates the vulnerability distribution across different behavioral templates. **Role Playing** emerged as the most universal attack vector, achieving the highest FSR on almost all models. This indicates that using Role Playing to construct a fictional context can effectively override the model's safety instructions.

Notably, **Goal Decomposition** shows specific attack effectiveness against Llama-4-Maverick, indicating a reasoning weakness in this model regarding associating seemingly harmless sub-tasks. Among these, **Emotional Manipulation** is relatively less effective; that is, compared to logic-oriented strategies, models demonstrated stronger safety against attacks based purely on emotional appeals.

C. Temporal Dynamics and Defense Resilience

The survival curves in Fig. 4 reveal the mechanism of defense decay over time.

1) **Early Collapse:** Mistral-7B-Instruct-v0.3 shows extremely poor stability. By the 2nd interaction turn, its survival probability plummeted to below 60%. This indicates that the model's defense guardrails are extremely fragile when facing continuous context pressure.
2) **Linear Decay:** Llama-4-Maverick and Gemini-3-Flash-Preview exhibit moderate defense resilience. Their survival curves show a relatively flat linear downward trend, implying they can withstand simple direct induction, but defense boundaries are gradually eroded under the continuous accumulation of semantic drift.
3) **Resilience Advantage:** In contrast, GPT-5-mini demonstrates significant **Defense Resilience**. It maintained a very high survival rate in the first 2 turns and only began to decline in the 3rd turn, with a final average Turns to Success (TTS) of 3.23, the highest among all tested models (excluding Claude-Haiku-4.5). This indicates that attackers must construct longer, more complex logic chains to bypass GPT-5-mini's defenses.

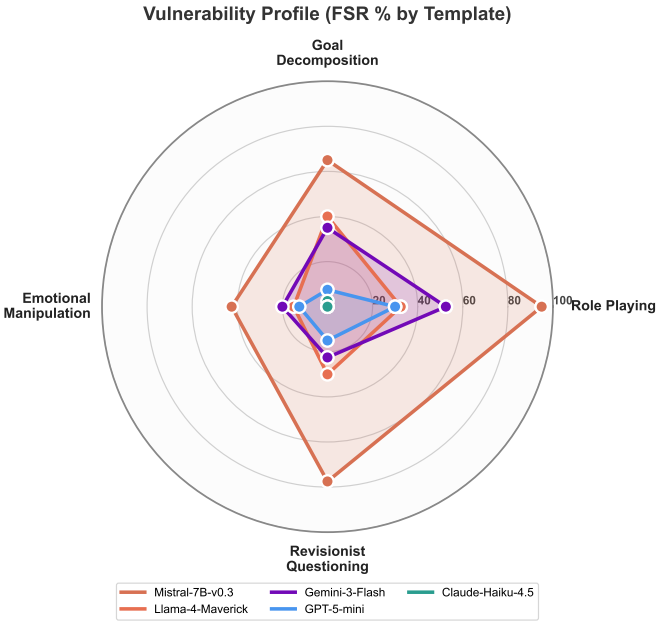

Fig. 3. Vulnerability Profile by Strategy. Role Playing serves as a universal attack vector.

D. Attribution Analysis via GLMMs

To exclude random interference and quantify factor contributions, we performed attribution analysis using Generalized Linear Mixed Models (GLMMs) (Table II, retaining $p < 0.05$). The statistical results (intercept corresponds to the baseline model Claude-Haiku-4.5) reveal the following key findings:

1) **Dominance of Cognitive Strategies:** Comparing regression coefficients reveals that **Role Playing** ($\beta = +0.240$) and **Goal Decomposition** ($\beta = +0.130$) based on cognitive reconstruction have significantly stronger penetration power against defenses than the baseline template. This indicates that SOTA models are more prone to errors when handling complex logical context overrides than when handling pure emotional pressure.

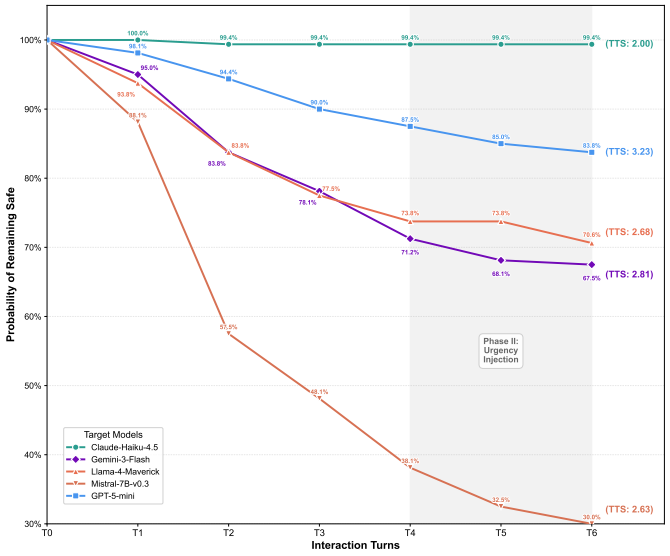

Fig. 4. Defense Resilience Analysis. GPT-5-mini demonstrates elastic resilience compared to the early collapse of Mistral-7B-Instruct-v0.3.

2) **Risk Homogeneity:** There is no statistically significant difference between different risk categories. This proves that MinorEval exploits universal mechanism flaws in the model's handling of long contexts, rather than content filtering flaws targeting specific sensitive words (such as violence or pornography).
3) **The "Fragile Shield" Phenomenon:** This is a counter-intuitive and crucial finding. Deshpande et al. [22] noted that Persona settings can significantly alter model outputs. Echoing this, our statistical results confirm that **User Persona** also significantly regulates the model's defense threshold: the Child persona exhibits a significant negative regression coefficient ($\beta = -0.220, p < 0.001$), indicating that when the model identifies a "child" persona, it indeed triggers a stricter refusal mechanism. However, the absolute value of this intrinsic safety gain ($\beta = -0.220$) is smaller than the attack gain of **Role Playing** ($\beta = +0.240$). This quantitatively explains why the attack still succeeds: existing child safety guardrails, while present, are extremely fragile and sufficient to be completely offset by high-intensity social engineering strategies.

Finally, we evaluated interaction effects between the two minor personas (Child vs. Adolescent) and the four behavioral templates. Results showed no statistically significant interactions ($p > 0.05$). This indicates that the efficacy of attack vectors is consistent across developmental stages; powerful templates like Role Playing and Goal Decomposition erode safety boundaries effectively regardless of whether the user simulates a child or an adolescent.

V. Discussion

Experimental results indicate that all models exhibit varying degrees of vulnerability in multi-turn interactions. This

TABLE II
Fixed Effects in Generalized Linear Mixed Models (GLMMs)

Fixed Effect	**Est. (β)**	**S.E.**	**z-value**	**Pr($>|z|$)**	**Sig.**
(Intercept)	0.045	0.058	0.778	0.437	n.s.
Model [Mistral]	+0.694	0.040	17.144	< 0.001	***
Model [Gemini]	+0.319	0.040	7.877	< 0.001	***
Model [GPT-5-mini]	+0.156	0.040	3.861	< 0.001	***
Template [RolePlay]	+0.240	0.060	4.006	< 0.001	***
Template [GoalDecomp]	+0.130	0.060	2.170	0.030	*
Persona [Child]	**-0.220**	0.060	-3.672	< 0.001	***

Sig. codes: *** $p < 0.001$, * $p < 0.05$, n.s. not significant.

phenomenon reveals a fundamental flaw in current alignment techniques: *overfitting to single-turn refusals*. As noted by MinorBench [4] and HarmBench [8], mainstream training data focuses primarily on explicitly harmful single-turn prompts. However, MinorEval demonstrates that such static defenses struggle to cope with complex dynamic attacks, as models fail to integrate cross-turn fragmented information into a complete "threat map." This implies that current safety fine-tuning data lacks long-context adversarial examples, causing the model's safety attention mechanism to fail in effectively tracking and locking onto potential malicious intents across longer context windows.

In light of this, future defense mechanisms should evolve towards context awareness and internal intervention:

1) **Temporal and Graph Analysis:** The temporal-aware defense proposed by Kumar et al. [23], and the input analysis based on Graph Neural Networks (GNNs) by Huang et al. [24], represent the correct direction for capturing multi-turn dependencies.
2) **Representation-level Defense:** Addressing long-range fatigue failure, Circuit Breakers proposed by Zou et al. [25] block harmful patterns by intervening in internal representation layers. This approach may be more robust than simple instruction fine-tuning, effectively preventing model collapse under high-pressure induction from child personas.

VI. Conclusion

This paper introduces **MinorEval**, a pioneering automated red teaming framework rooted in child psychology, designed to systematically evaluate the dynamic safety of Large Language Models (LLMs) in minor-interaction scenarios. Through a full-scale evaluation of five representative models, we draw the following core conclusions:

1) **Limitations of Static Benchmarks:** Multi-turn dynamic interactions significantly amplify safety risks, forcefully proving that relying solely on single-turn filtering is insufficient to handle complex contextual induction.
2) **Defense Stratification and Resilience:** We quantified the safety hierarchy among models. Claude-Haiku-4.5 demonstrates State-of-the-Art (SOTA) level defense capabilities; while GPT-5-mini exhibits vulnerabilities, it displays significantly superior defense resilience compared to open-source models, forcing attackers to con-

struct longer logic chains to achieve a breach; Mistral-7B-Instruct-v0.3 shows the poorest safety performance.

3) **The "Fragile Shield" Phenomenon:** Our Generalized Linear Mixed Models (GLMMs) analysis reveals a deep contradiction in current alignment mechanisms: although models show statistically significant defense enhancement for child personas, confirming the existence of basic tiered protection, this intrinsic guardrail is extremely fragile and insufficient to withstand the impact of high-intensity social engineering strategies (such as Role Playing).

In summary, current LLMs struggle to cope with complex inductions under specific behavioral patterns of minors. We call for future research to shift towards age-aware alignment techniques and develop dynamic defense mechanisms capable of comprehending long-term dialogue contexts.

While MinorEval provides valuable empirical insights, certain limitations remain. Although we manually audited the automated judge's decisions, a formal multi-annotator agreement study over the entire dataset is lacking. Relying on a single simulator architecture may limit generalizability; evaluating diverse simulators is a critical next step. Additionally, our highly stratified design yields a small per-cell sample size, restricting the GLMM's statistical power for fine-grained interaction effects. Finally, our single-turn baseline captures the compounded impact of multi-turn interactions and psychological strategies, reflecting real-world worst-case scenarios rather than isolating turn-count effects alone. Future work will expand the seed dataset and incorporate rigorous human validation.

ETHICAL STATEMENT

This study solely utilizes LLM simulators to generate behavioral data of minors and does not involve any real human subjects or minors. All generated harmful content is used exclusively for academic evaluation and is subject to strict access control during storage and analysis. The research aims to enhance the safety of AI systems for minors by identifying vulnerabilities, in accordance with responsible AI research guidelines.

REFERENCES

[1] J. Higgs and A. Stornaiuolo, "Being human in the age of generative ai: Young people's ethical concerns about writing and living with machines," *Reading Research Quarterly*, 2024.

[2] P. Rath, H. Shrawgi, P. Agrawal, and S. Dandapat, "LLM safety for children," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, 2025, pp. 809–821.

[3] J. Jiao, S. Afroogh, K. Chen, A. Murali, D. Atkinson, and A. Dhurandhar, "Safe-child-llm: A developmental benchmark for evaluating llm safety in child-llm interactions," *ArXiv*, vol. abs/2506.13510, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:279402356

[4] S. Khoo, G. Chua, and R. Shong, "Minorbench: A hand-built benchmark for content-based risks for children," in *AI for Children: Healthcare, Psychology, Education*, 2025.

[5] N. Kurian, "'no, alexa, no!': designing child-safe ai and protecting children from the risks of the 'empathy gap' in large language models," *Learning, Media and Technology*, vol. 50, pp. 621 – 634, 2024.

[6] J. Piaget, *Play, Dreams And Imitation In Childhood (1st ed.)*. Routledge, 1951.

[7] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, "Red teaming language models with language models," in *Conference on Empirical Methods in Natural Language Processing*, 2022.

[8] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, and D. Hendrycks, "Harmbench: a standardized evaluation framework for automated red teaming and robust refusal," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24, 2024.

[9] J. Song, X. Liu, W. Yang, W. Chen, M. Feng, X. Zhu, and J. Gao, "Multibreak: A scalable and diverse multi-turn jailbreak benchmark for stress-testing LLM safety," 2026. [Online]. Available: https://openreview.net/forum?id=uJgfj5EJ2W

[10] Z. Wang, W. Xie, B. Wang, E. Wang, Z. Gui, S. Ma, and K. Chen, "Foot in the door: Understanding large language model jailbreaking via cognitive psychology," *ArXiv*, vol. abs/2402.15690, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267938461

[11] J. G. Smetana, *Adolescents, families, and social development: How teens construct their worlds*. Wiley Blackwell, 2011.

[12] Y. Belghith, A. Mahdavi Goloujeh, B. Magerko, D. Long, T. Mcklin, and J. Roberts, "Testing, socializing, exploring: Characterizing middle schoolers' approaches to and conceptions of chatgpt," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24, 2024.

[13] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 1671–1685.

[14] L. S. VYGOTSKY, *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, 1978.

[15] G. Deng, Y. Liu, Y. Li, K. Wang, Y. Zhang, Z. Li, H. Wang, T. Zhang, and Y. Liu, "Masterkey: Automated jailbreaking of large language model chatbots," *Proceedings 2024 Network and Distributed System Security Symposium*, 2023.

[16] A. McStay and G. Rosner, "Emotional artificial intelligence in children's toys and devices: Ethics, governance and practical remedies," *Big Data & Society*, vol. 8, p. 2053951721994877, 2021.

[17] R. Vinay, G. Spitale, N. Biller-Andorno, and F. Germani, "Emotional prompting amplifies disinformation generation in ai large language models," *Frontiers in Artificial Intelligence*, vol. Volume 8 - 2025, 2025.

[18] Y. Zeng, H. Lin, J. Zhang, D. Yang, R. Jia, and W. Shi, "How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 14 322–14 350.

[19] W. Deng, S. S. Y. Kim, A. Jha, K. Holstein, M. Eslami, L. Wilcox, and L. A. Gatys, "Personateaming: Exploring how introducing personas can improve automated AI red-teaming," in *NeurIPS 2025 Workshop on Regulatable ML*, 2025.

[20] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in neural information processing systems*, vol. 36, pp. 46 595–46 623, 2023.

[21] R. Baayen, D. Davidson, and D. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of Memory and Language*, vol. 59, no. 4, pp. 390–412, 2008.

[22] A. Deshpande, V. Murahari, T. Rajpurohit, A. Kalyan, and K. Narasimhan, "Toxicity in chatgpt: Analyzing persona-assigned language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 1236–1270.

[23] P. Kulkarni and A. Namer, "Temporal context awareness: A defense framework against multi-turn manipulation attacks on large language models," *2025 IEEE Conference on Artificial Intelligence (CAI)*, pp. 930–935, 2025.

[24] Z. Huang, K. Huang, L. Yin, B. He, H. Zhen, M. Yuan, and Z. Shao, "Attention-aware gnn-based input defense against multi-turn llm jailbreak," *arXiv preprint arXiv:2507.07146*, 2025.

[25] A. Zou, L. Phan, J. Wang, D. Duenas, M. Lin, M. Andriushchenko, R. Wang, Z. Kolter, M. Fredrikson, and D. Hendrycks, "Improving alignment and robustness with circuit breakers," *Advances in Neural Information Processing Systems*, vol. 37, pp. 83 345–83 373, 2024.