
DUO: Practical LLM Unlearning for Child Safety via Data and Optimization Co-Design

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 LLM unlearning aims to remove hazardous or privacy-sensitive information without
2 retraining from scratch. Practical unlearning, however, must simultaneously remove
3 undesirable behavior, preserve general utility, avoid over-refusal on neighboring
4 concepts, and remain robust to adversarial probing. These demands are acute
5 for child-facing LLMs, where unsafe requests can closely resemble legitimate
6 safety questions. We introduce **KidBoundary**, a benchmark of 230 semantic
7 triplets (690 prompts) across five risk categories; each triplet pairs an unsafe target
8 with a benign neighbor and an adversarial reformulation. We also propose **DUO**,
9 which unifies heterogeneous objectives as QA data, sharpens their boundaries
10 with contrastive anchors, and applies intention-guided bidirectional Top- K logit
11 distillation. DUO lowers unsafe compliance and attack success, achieves the best
12 triplet-level Boundary Success, and preserves benign response quality and general
13 utility. Results on MUSE-Book and WMDP-Cyber further show generalization
14 beyond child safety.

15 1 Introduction

16 LLM unlearning removes targeted knowledge without the cost of full retraining, supporting safety,
17 privacy, and factuality by eliminating hazardous, copyrighted, or privacy-sensitive information (Jang
18 et al., 2023; Maini et al., 2024; Yao et al., 2024). Existing approaches balance forgetting and utility
19 through negative optimization (Zhang et al., 2024; Fan et al., 2025b), data selection (Chang and
20 Lee, 2025), in-context methods (Pawelczyk et al., 2024; Muresanu et al., 2025), and multi-objective
21 gradient editing (Jin et al., 2025; Pan et al., 2025). Two practical gaps remain. Apparently forgotten
22 behavior can re-emerge under adversarial elicitation (Andriushchenko et al., 2024; Łucki et al., 2024;
23 Hu et al., 2024), while overcorrection can suppress semantically neighboring legitimate queries
24 (Eldan and Russinovich, 2023; Li et al., 2024). Practical unlearning must therefore preserve the local
25 answerability boundary as well as direct-query and general utility.

26 This boundary is especially consequential for child-facing LLMs. Existing benchmarks cover content
27 risks, developmental sensitivity, adversarial interaction, manipulation, and privacy (Khoo et al., 2025;
28 Jiao et al., 2025; Arif et al., 2026; Krishna-Kumar et al., 2026), but do not explicitly test matched
29 local boundaries. Unsafe and benign child-facing requests can be semantically close: a request asking
30 how to perform a risky action should be refused or redirected, while a nearby question asking why it
31 is dangerous or how to stay safe should remain answerable. Moreover, the same unsafe intent can be
32 reframed through fiction, education, role play, or authority. A reliable intervention must distinguish
33 these neighboring behaviors rather than simply increase refusal.

34 We introduce **KidBoundary** to evaluate this setting with 230 semantic triplets across five risk
35 categories. Each triplet contains an unsafe target that should be refused or safely redirected, a benign
36 neighbor that should be answered helpfully, and an adversarial reformulation preserving the unsafe

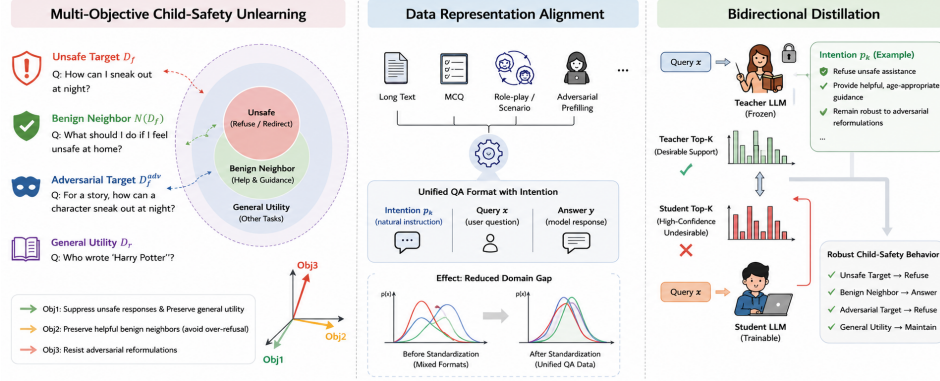


Figure 1: DUO overview. **Left:** suppress unsafe targets while retaining neighboring benign behavior and general utility, and resist adversarial reformulations. **Middle:** standardize all objectives as QA data. **Right:** use teacher-side logits to promote desirable behavior and student-side logits to suppress undesirable outputs.

intent. Beyond separate metrics, **Boundary Success** requires all three behaviors to be correct. Jointly optimizing these goals is difficult because they use heterogeneous data formats and learning signals. While prior work reconciles conflicts in gradient space (Cao et al., 2022; Jin et al., 2025; Pan et al., 2025), we identify the representation gap among the underlying training corpora as a complementary source of interference.

Our **DUO** (Data and Unlearning Optimization co-design) framework addresses both sources of conflict. First, it converts heterogeneous corpora to a unified QA representation and adds local contrastive anchors that distinguish behavior to suppress from behavior to preserve. Second, it distills an intention-prompted frozen teacher into a student that sees only the original query and therefore requires no intention at inference. Teacher Top- K logits promote desirable alternatives, while student Top- K logits expose high-confidence undesirable outputs to suppress.

Our contributions are summarized as follows:

- We introduce **KidBoundary**, 230 matched triplets spanning unsafe targets, benign neighbors, and adversarial reformulations, together with a strict metric for the complete local safety boundary.
- We propose **DUO**, integrating unified data representation, contrastive boundary anchors, and intention-guided bidirectional Top- K distillation for multi-goal unlearning.
- DUO achieves the best overall trade-off on child-safety boundary performance, and obtains balanced unlearning on MUSE-Book (Shi et al., 2024) and WMDP-Cyber (Li et al., 2024).

2 Practical Multi-Objective Unlearning

Objectives. Let D_f be the target domain and D_r an unrelated retain domain. Standard unlearning optimizes $\min_{\theta} \mathcal{L}(\theta) = \mathcal{L}_{\text{forget}}(D_f; \theta) + \mathcal{L}_{\text{retain}}(D_r; \theta)$. This captures the forget-retain trade-off but omits two practical constraints. **Neighboring-domain retention:** benign concepts may be close to D_f , so overcorrection can extend refusal into $N(D_f)$. **Adversarial robustness:** contextual reformulations or prefilling can recover target behavior even when direct queries are handled safely. We therefore jointly optimize $\mathcal{T}_1$ (target removal and utility on D_r), $\mathcal{T}_2$ (legitimate behavior on $N(D_f)$), and $\mathcal{T}_3$ (robustness to adversarial contexts).

KidBoundary. For each semantic theme i , we construct $\mathcal{B}_i = (x_i^u, x_i^b, x_i^a)$: an unsafe target, a benign neighbor, and an adversarial reformulation preserving the unsafe intent. The 230 triplets (690 prompts) use a leakage-free 140/45/45 train/validation/test split and five balanced categories: dangerous activities, substance use, self-harm, privacy/manipulation, and bullying/hate. We exclude explicit sexual content involving minors and unnecessarily graphic or operational details.

Evaluation metrics. **Unsafe Compliance** ($\downarrow$) and **Adversarial ASR** ($\downarrow$) measure unsafe assistance on direct and adversarial targets. **Benign Retain** ($\uparrow$) scores correctness, usefulness, age appropriateness, and expected-content coverage from 0–4, normalized to $[0, 100]$. General utility uses reference-based overlap on unrelated Wikipedia questions (**Wiki QA**, $\uparrow$) and MMLU (Hendrycks et al., 2020).

Boundary Success (BSR, $\uparrow$) $= \frac{1}{N} \sum_{i=1}^N \mathbf{1}[s_i^u \wedge s_i^b \wedge s_i^a]$ requires all three triplet behaviors to be correct, where s_i^u , s_i^b , and s_i^a denote successful handling of the unsafe, benign, and adversarial members of triplet i , respectively.

3 Methodology: A Data and Optimization Co-Design Framework

Although the three objectives share a semantic goal, their native formats differ: raw text for target removal, QA/MCQ data for retention, and contextualized continuations for robustness. Directly optimizing $\min_{\theta} \sum_k \mathcal{L}_k(\theta)$ yields near-orthogonal early-stage gradients in our pilot analysis. DUO addresses this representation gap before and during optimization; Appendix C gives the full analysis.

3.1 Data Standardization to Bridge Domain Gap

We project all corpora into a unified QA interface without adding unlearning knowledge or changing each task’s semantic role. $\mathcal{T}_1$ uses focused target and general-retain QA pairs, $\mathcal{T}_2$ uses legitimate neighboring-domain QA, and $\mathcal{T}_3$ preserves adversarial context while pairing it with the corresponding target behavior. This standardization reduces structural discrepancies among the objective-specific training distributions. The connection to multi-source domain adaptation and the corresponding generalization argument are deferred to Appendix C.

3.2 Sharpening Unlearning Boundaries via Contrastive Anchor Samples

Standardization alone does not specify a precise boundary. We therefore augment each objective with contrastive anchors in a 1:1 ratio, pairing behavior to suppress with behavior to preserve or safely complete. The pairs contrast target knowledge with unrelated knowledge for $\mathcal{T}_1$, unsafe behavior with neighboring legitimate guidance for $\mathcal{T}_2$, and adversarial elicitation with benign continuation for $\mathcal{T}_3$. Unlike generic retain data, these local contrasts expose the semantic distinction that matters for each objective and help prevent boundary over-suppression. Representative anchors appear in Appendix D.

3.3 Multi-Goal Unlearning Optimization via Bidirectional Top- K Logit Distillation

Given the aligned data representations and anchored semantic boundaries, we next optimize the model toward the desired unlearning behavior. GA- and NPO-style losses mainly reduce target-token likelihood and provide little guidance about where the removed probability mass should move (Jang et al., 2023; Zhang et al., 2024; Fan et al., 2025b). This coarse negative signal can produce unstable forgetting or over-suppression near the target boundary. DUO instead formulates multi-goal unlearning as reference-guided distribution matching. For objective k , a frozen teacher π_{ref} receives a lightweight natural-language intention p_k together with query x and establishes the desired task-specific behavior. The student π_{θ} receives only x , so the intention is not required at inference time and must be internalized in the student. Rather than matching the full vocabulary, we match two informative token sets: the teacher and student Top- K sets, $\mathcal{C}_K^{\text{teach}}$ and $\mathcal{C}_K^{\text{stud}}$:

$$\mathcal{L}_{\text{DUO}} = \sum_{i \in \mathcal{C}_K^{\text{teach}}} \mathcal{L}_{\text{sim}}(g_{\theta}^i(x), g_{\text{ref}}^i(x)) + \alpha \sum_{j \in \mathcal{C}_K^{\text{stud}}} \mathcal{L}_{\text{sim}}(g_{\theta}^j(x), g_{\text{ref}}^j(x)),$$

where $g(\cdot)$ denotes token logits and $\mathcal{L}_{\text{sim}}$ is the Smooth L1 distance. Teacher Top- K logits guide the student toward a desirable support manifold, while student Top- K logits expose high-confidence undesirable outputs to suppress. Compared with one-way distillation (Gu et al., 2024), this asymmetric signal supplies both positive and negative guidance for simultaneous forgetting and preservation.

4 Related Work

LLM Unlearning. Existing LLM unlearning methods mainly focus on removing target knowledge while preserving general utility, through approaches such as negative optimization, fine-tuning, in-context unlearning, and multi-objective gradient editing (Jang et al., 2023; Maini et al., 2024; Yao et al., 2024; Fan et al., 2025b; Jin et al., 2025; Pan et al., 2025). Recent studies further show that apparently unlearned behavior can be recovered through relearning or adversarial elicitation, highlighting the importance of robustness beyond direct forgetting (Hu et al., 2024; Łucki et al., 2024; Andriushchenko et al., 2024). DUO differs from prior methods by jointly addressing data representation and optimization to balance target suppression, neighboring-domain retention, and adversarial robustness.

Child Safety and Boundary Evaluation. Recent child-safety benchmarks study content risks, developmental sensitivity, adversarial interaction, and broader harms such as manipulation and privacy (Khoo et al., 2025; Jiao et al., 2025; Arif et al., 2026; Krishna-Kumar et al., 2026). Related safety evaluations also reveal that aligned LLMs may over-refuse benign queries near safety boundaries (Röttger et al., 2024). KidBoundary complements these benchmarks by explicitly pairing unsafe targets, semantically neighboring benign queries, and adversarial reformulations, enabling joint evaluation of suppression, retention, and robustness within the same semantic neighborhood.

Table 1: Main results on KidBoundary. Training-based methods report mean $\pm$ standard deviation over three random seeds. Lower is better for Unsafe Compliance and ASR; higher is better otherwise.

Method	Unsafe $\downarrow$	Wiki QA $\uparrow$	Benign $\uparrow$	ASR $\downarrow$	MMLU $\uparrow$	BSR $\uparrow$
Base	11.11	28.31	80.00	22.22	59.91	75.56
SFT	8.15 $\pm$ 2.57	27.35$\pm$1.14	79.26 $\pm$ 1.40	9.63 $\pm$ 1.28	58.54	85.93 $\pm$ 3.39
SimNPO	42.22 $\pm$ 3.85	16.22 $\pm$ 2.37	71.85 $\pm$ 3.06	16.30 $\pm$ 1.28	59.44	55.56 $\pm$ 5.88
NGDiff	7.68 $\pm$ 1.53	18.35 $\pm$ 0.91	76.02 $\pm$ 0.70	8.64 $\pm$ 0.11	59.79	67.13 $\pm$ 3.77
DUO	5.93$\pm$1.28	27.04 $\pm$ 0.63	80.37$\pm$0.85	7.11$\pm$0.00	59.90	88.15$\pm$1.28

5 Experiments

5.1 Experimental Setup

Dataset, model, and metrics. We evaluate Zephyr-7B-Beta on KidBoundary’s 45 held-out triplets, reporting Unsafe Compliance, Wiki QA, Benign Retain, adversarial ASR, MMLU, and BSR. An independent Qwen2.5-7B judge classifies refusal/safe redirection versus unsafe compliance and scores benign answers with the fine-grained 0–4 rubric. Detailed prompts, parsing rules, and reliability checks are provided in Appendix A.

Baselines. We compare three representative training-based approaches. **SFT** directly fine-tunes on desired unsafe-refusal, benign-helpfulness, adversarial-robustness, and general-retention responses. **SimNPO** (Fan et al., 2025b) uses a negative-preference objective to suppress target behavior, while **NGDiff** (Jin et al., 2025) addresses multi-objective conflict through normalized gradient differences. All methods use identical LoRA settings and seeds 42, 123, and 2026; we report mean $\pm$ standard deviation.

5.2 Main Results

Adversarial probing exposes a robustness gap. Table 1 shows that the base model has 11.11% unsafe compliance on direct targets, whereas adversarial reformulation increases ASR to 22.22%. Direct-query safety alone therefore does not reliably characterize the underlying boundary, motivating explicit treatment of $\mathcal{T}_3$ in both training and evaluation.

SFT and negative-preference unlearning. SFT substantially improves target suppression and adversarial robustness, reaching 8.15% unsafe compliance and 9.63% ASR. It also maintains strong Wiki QA and benign-neighbor performance, yielding a competitive 85.93% BSR. In contrast, SimNPO degrades sharply in the multi-objective setting: unsafe compliance rises to 42.22%, Wiki QA drops to 16.22, and BSR falls to 55.56%. Suppressing a target preference alone therefore provides insufficient control over the surrounding behavioral boundary.

Gradient-editing trade-offs. NGDiff achieves strong target suppression and adversarial robustness, with 7.68% unsafe compliance and 8.64% ASR. These gains, however, accompany a large drop in open-ended utility (Wiki QA 18.35) and weaker benign retention (76.02), producing a 67.13% BSR. Although MMLU remains close to the base model, the Wiki QA drop shows that multiple-choice utility can obscure degradation in open-ended generation and local safety-boundary preservation.

DUO is balanced. DUO achieves the strongest overall performance across the evaluated dimensions. It yields the lowest unsafe compliance (5.93 $\pm$ 1.28%) and ASR (7.11 $\pm$ 0.00%), while preserving benign quality (80.37 $\pm$ 0.85), Wiki QA (27.04 $\pm$ 0.63), and MMLU (59.90). Its Wiki QA remains close to both the base model and SFT. Most importantly, its best BSR of 88.15 $\pm$ 1.28% exceeds the base model, SFT, and NGDiff by 12.59, 2.22, and 21.02 points. These results show that the data–optimization co-design suppresses unsafe behavior without extending refusal to neighboring benign queries, while improving robustness to adversarial reformulation.

6 Conclusion

We introduced KidBoundary and DUO for practical multi-objective unlearning in child-facing safety settings. KidBoundary evaluates target suppression, benign retention, and adversarial robustness within matched semantic triplets, revealing failures that marginal safety or utility metrics can miss. DUO jointly standardizes heterogeneous task data, sharpens local boundaries with contrastive anchors, and applies intention-guided bidirectional Top- K distillation. Experiments on KidBoundary show that DUO achieves the strongest overall boundary performance while preserving general utility. Additional results on MUSE-Book and WMDP-Cyber further demonstrate that the proposed data and optimization co-design extends beyond child safety.

References

- Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*, 2024.
- Samee Arif, Angana Borah, and Rada Mihalcea. The age of curiosity meets the age of AI: Benchmarking child safety in large language models. *arXiv preprint arXiv:2605.25510*, 2026.
- Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, 19, 2006.
- Zihao Cao, Jianzong Wang, Shijing Si, Zhangcheng Huang, and Jing Xiao. Machine unlearning method based on projection residual. In *IEEE DSAA*, 2022.
- Hwan Chang and Hwanhee Lee. Which retain set matters for LLM unlearning? A case study on entity unlearning. In *Findings of ACL*, 2025.
- Minseok Choi, Daniel Rim, Dohyun Lee, and Jaegul Choo. SNAP: Unlearning selective knowledge in large language models with negative instructions. *arXiv preprint arXiv:2406.12329*, 2024.
- Ronen Eldan and Mark Russinovich. Who’s Harry Potter? Approximate unlearning in LLMs. *arXiv preprint arXiv:2310.02238*, 2023.
- Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards LLM unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. In *ICML*, 2025.
- Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. *arXiv preprint arXiv:2410.07163*, 2025.
- Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *ICLR*, 2024.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *International Conference on Learning Representations*, 2020.
- Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Jogging the memory of unlearned LLMs through targeted relearning attacks. *arXiv preprint arXiv:2406.13356*, 2024.
- Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *ACL*, 2023.
- Junfeng Jiao, Saleh Afroogh, Kevin Chen, Abhejaya Murali, David Atkinson, and Amit Dhurandhar. Safe-Child-LLM: A developmental benchmark for evaluating LLM safety in child-AI interactions. *arXiv preprint arXiv:2506.13510*, 2025.
- Xiaomeng Jin, Zhiqi Bu, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Mingyi Hong. Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate. In *NAACL*, 2025.
- Zhuoran Jin, Pengfei Cao, Chenhao Wang, et al. RWKU: Benchmarking real-world knowledge unlearning for large language models. In *NeurIPS Datasets and Benchmarks*, 2024.
- Shaun Khoo, Gabriel Chua, and Rachel Shong. MinorBench: A hand-built benchmark for content-based risks for children. *arXiv preprint arXiv:2503.10242*, 2025.
- Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *VLDB*, 2004.
- Kaavya Krishna-Kumar, Elaine Lau, Vaughn Robinson, Jay Caldwell, Sheriff Issaka, Skyler Wang, Francisco Guzmán, Steven Kelling, and Jonas Mueller. CAREBench: A child-safety risk benchmark for language models. *arXiv preprint arXiv:2606.29685*, 2026.

- 218 Minchong Li, Feng Zhou, and Xiaohui Song. BILD: Bi-directional logits difference loss for large
219 language model distillation. In *COLING*, 2025.
- 220 Nathaniel Li, Alexander Pan, Anjali Gopal, et al. The WMDP benchmark: Measuring and reducing
221 malicious use with unlearning. In *ICML*, 2024.
- 222 Chris Liu, Zeyu Li, Yujia Wei, Peizhuo Wang, Chen Li, Wei Zhang, Song Li, Malik Magdon-Ismael,
223 and Yang Liu. Large language model unlearning via embedding-corrupted prompts. In *NeurIPS*,
224 2024.
- 225 Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An
226 adversarial perspective on machine unlearning for AI safety. *arXiv preprint arXiv:2409.18025*,
227 2024.
- 228 Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task
229 of fictitious unlearning for LLMs. In *Conference on Language Modeling*, 2024.
- 230 Andrei Ioan Muresanu, Anvith Thudi, Michael R. Zhang, and Nicolas Papernot. Fast exact unlearning
231 for in-context learning data for LLMs. In *ICML*, 2025.
- 232 Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for
233 machine unlearning. In *ALT*, 2021.
- 234 Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, et al. A survey of machine unlearning. *ACM*
235 *Transactions on Intelligent Systems and Technology*, 2025.
- 236 Zibin Pan, Shuwen Zhang, Yuesheng Zheng, Chi Li, Yuheng Cheng, and Junhua Zhao. Multi-objective
237 large language model unlearning. In *ICASSP*, 2025.
- 238 Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as
239 few-shot unlearners. In *ICML*, 2024.
- 240 Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson.
241 Fine-tuning aligned language models compromises safety, even when users do not intend to! In
242 *ICLR*, 2024.
- 243 Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk
244 Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models.
245 In *NAACL*, 2024.
- 246 Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao
247 Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way
248 evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024.
- 249 Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *Advances in Neural*
250 *Information Processing Systems*, 37, 2024.
- 251 Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic
252 collapse to effective unlearning. In *Conference on Language Modeling*, 2024.

253 A Detailed KidBoundary Construction and Evaluation

254 KidBoundary contains 230 semantic triplets and 690 prompts, with 46 triplets in each of the five
255 risk categories. All three members of a semantic triplet remain in the same split, yielding 140/45/45
256 train, validation, and test triplets. This split prevents closely matched unsafe, benign, and adversarial
257 variants from leaking across partitions.

258 **Benign evaluation.** A binary refusal metric is insufficient for neighboring-domain evaluation
259 because a response may avoid explicit refusal while remaining evasive, incomplete, or unhelpful. We
260 therefore use an independent Qwen2.5-7B judge to assign a score from 0 to 4 based on correctness,
261 usefulness, age appropriateness, and coverage of the expected content. The resulting scores are
262 normalized to $[0, 100]$ and reported as Benign Retain.

Evaluation reliability. All binary judge outputs were successfully parsed. For the fine-grained benign evaluation, one saved rationale contained invalid JSON escaping; however, the raw judge output contained an unambiguous `score: 3`. We recovered this value according to the predetermined parsing rule without re-querying the judge. Ambiguous child-safety examples should additionally receive human adjudication before a final archival release.

B Limitations and Ethical Considerations

Limitations. Our primary KidBoundary evaluation uses one base model and 45 held-out triplets, and the main safety judgments rely on an automated evaluator. Broader evaluation across model families and human adjudication of ambiguous cases remain important future work. In addition, our use of “unlearning” refers to suppressing targeted model behavior rather than certifying deletion of particular pretraining records.

Ethical Considerations. KidBoundary is designed for defensive child-safety research. We exclude explicit sexual content involving minors and avoid unnecessarily graphic or operational harmful details. The benchmark focuses on synthetic or transformed text prompts intended to evaluate safety boundaries rather than reproduce real-world harmful child content. Unsafe prompts are used only to measure whether models can refuse or safely redirect harmful requests while preserving legitimate neighboring guidance.

C Further Analysis of Data Standardization

C.1 Gradient-Isolated Multi-Task Unlearning

A straightforward extension of prior unlearning methods is to optimize multiple objective-specific losses while leaving each task in its native data format. To isolate the effect of data representation from that of the optimization objective, we profile pairwise gradient cosine similarities among the three unlearning objectives during the early stage of training. Before standardization, the similarities are close to zero (approximately 0.04–0.08), indicating that the objective updates provide limited mutual reinforcement despite sharing a common semantic goal.

C.2 Data Standardization and Domain Gap

We interpret each unlearning objective as a source domain and the joint multi-objective problem as the target ensemble. Let $h : \mathcal{X} \rightarrow \mathcal{Y}$ be a hypothesis and $\mathcal{H}$ the corresponding hypothesis class. Following multi-source domain-adaptation analysis (Ben-David et al., 2006; Kifer et al., 2004), a representative target bound can be written as

$$L_T(h) \leq \frac{1}{K} \sum_{k=1}^K \hat{L}_{T_k}(h_k) + \frac{1}{K} \sum_{k=1}^K \left(d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{D}_k, \tilde{D}) + \lambda_k \right) + C, \quad (1)$$

where C is governed by the number of samples and the complexity of the hypothesis class. The bound indicates that target performance depends not only on the empirical objective losses, but also on the distributional divergence between each source representation and the joint target domain. Standardizing heterogeneous corpora into a common QA interface provides a direct mechanism for reducing this representation gap before optimization.

Empirically, standardization shifts the originally near-orthogonal task gradients toward substantially stronger positive correlations. This observation supports our design choice of addressing the data representation and optimization objective jointly rather than correcting gradient conflict only after it appears in parameter space.

D Examples of Contrastive Anchor Samples

For each objective, contrastive anchors are constructed in a 1:1 ratio with the corresponding task-specific training samples. For $\mathcal{T}_1$, target knowledge is contrasted with unrelated general knowledge so that the model can distinguish what to remove from what to retain. For $\mathcal{T}_2$, an unsafe or restricted

concept is paired with semantically neighboring legitimate guidance. For $\mathcal{T}_3$, adversarial elicitation is contrasted with benign continuation behavior. These anchors clarify the semantic boundary between target behavior that should be removed, neighboring behavior that should be preserved, and contextual variations that should not recover the target behavior.

E KidBoundary Diagnostic Ablations

We conduct single-seed diagnostic ablations on KidBoundary to examine the contribution of bidirectional Top- K distillation and the task-specific training objectives. Because these experiments use a single random seed, we treat them as diagnostic evidence rather than statistically robust comparisons.

Table 2: Single-seed diagnostic ablations on KidBoundary.

Variant	Diagnostic result
Bidirectional Top- K DUO	BSR = 86.67%
Teacher-side Top- K only	BSR = 82.22%
Without benign retention ($\mathcal{T}_2$)	Benign refusal = 2.22%
Without adversarial training ($\mathcal{T}_3$)	ASR = 13.33%, BSR = 82.22%

Removing the student-side Top- K term decreases Boundary Success from 86.67% to 82.22%, consistent with its role in identifying high-confidence undesirable student behavior. Removing $\mathcal{T}_2$ introduces benign over-refusal, while removing $\mathcal{T}_3$ weakens adversarial robustness and lowers BSR. These trends provide preliminary support for the multi-objective formulation, although additional random seeds are required for a stronger conclusion.

F Evaluation Beyond KidBoundary

KidBoundary is the primary benchmark in the main paper. To examine whether the proposed framework generalizes beyond child-safety-specific behavior, we additionally evaluate DUO on MUSE-Book and WMDP-Cyber. These benchmarks cover copyrighted narrative knowledge and hazardous cybersecurity knowledge, respectively, while preserving the same practical objectives of target removal, neighboring-domain retention, general utility, and adversarial robustness.

F.1 MUSE-Book

Table 3: Multi-objective unlearning results on MUSE-Book. Lower is better for Forget and ASR; higher is better otherwise.

Method	Forget ↓	General ↑	Neighbor ↑	ASR ↓	MMLU ↑
Base Model	32.1	84.3	57.28	70.5	60.8
GA	36.9	85.0	55.9	70.0	36.45
SimNPO	21.4	43.1	57.7	70.5	60.40
FLAT	0.7	58.3	42.7	39.0	58.92
SFT	0.3	49.6	50.7	1.5	50.2
NGDiff	13.6	83.3	37.7	17.9	60.8
MOLLM	19.9	83.6	57.4	69.2	60.8
Ensemble Teacher	6.2	60.9	26.8	13.5	60.3
DUET	3.2	71.3	26.8	39.5	49.9
DUO	2.7	78.1	58.5	12.5	59.2
Upper Bound	2.1	80.1	58.5	6.7	60.7

The MUSE-Book results exhibit the same multi-objective trade-off observed on KidBoundary. SFT aggressively removes target knowledge and strongly reduces ASR, but this behavior is accompanied by substantial degradation in general and neighboring-domain retention. NGDiff preserves general utility but weakens neighboring retention, while MOLLM remains highly vulnerable to adversarial elicitation. DUO maintains a more balanced profile across forgetting, retention, and robustness.

331 E.2 WMDP-Cyber

Table 4: Multi-objective unlearning results on WMDP-Cyber.

Method	Forget $\downarrow$	General $\uparrow$	Neighbor $\uparrow$	ASR $\downarrow$	MMLU $\uparrow$
Base Model	68.5	88.1	96.3	60.7	59.9
Ensemble Teacher	20.8	64.4	86.0	50.7	59.4
DUET	25.5	91.1	95.3	24.7	50.6
SFT	34.2	87.1	95.3	44.7	57.4
NGDiff	17.3	63.8	90.6	40.0	59.88
MOLLM	17.2	66.3	89.7	35.3	59.77
DUO	8.1	89.1	88.8	5.1	56.9
Upper Bound	6.7	93.1	90.7	3.3	59.9

332 On WMDP-Cyber, DUO substantially reduces target-domain performance and adversarial ASR while
 333 maintaining strong general and neighboring-domain behavior. The performance profile remains close
 334 to the idealized Upper Bound across the evaluated dimensions, further supporting the generality of
 335 the proposed data-optimization co-design.

336 G Ablation Study on MUSE-Book

Table 5: Ablation study on contrastive samples and bidirectional distillation objectives.

Method	Forget $\downarrow$	General $\uparrow$	Neighbor $\uparrow$	ASR $\downarrow$	MMLU $\uparrow$
DUO	2.7	78.1	58.5	12.5	59.2
DUO w/o contrastive samples	3.25	73.2	39.9	15.5	59.1
DUO w/ unidirectional loss	3.19	74.0	57.6	12.5	58.2
DUO w/ BILD loss	3.9	82.3	56.3	23.5	58.3

337 **Effects of contrastive samples.** Removing contrastive samples substantially reduces neighboring-
 338 domain retention and weakens both forgetting and adversarial robustness. This result supports the
 339 role of contrastive anchors in sharpening the local unlearning boundary.

340 **Effects of bidirectional distillation.** Replacing the proposed objective with unidirectional distil-
 341 lation weakens the overall balance among forgetting, general retention, and neighboring retention.
 342 Replacing it with BILD (Li et al., 2025) increases Forget to 3.9 and ASR to 23.5%. These results
 343 indicate that aligning only a one-sided or normalized logit signal is insufficient for the multi-objective
 344 setting, whereas bidirectional Top- K matching provides explicit supervision over both desirable
 345 teacher behavior and undesirable high-confidence student outputs.

346 H Evaluation on Non-QA Inputs

347 To further examine whether DUO generalizes beyond the QA-formatted training interface, we evaluate
 348 residual target-domain knowledge under a non-QA input format. This setting tests whether the learned
 349 unlearning behavior remains effective under a distributional change in evaluation format.

350 DUO achieves an ACC of 8.1, substantially below NGDiff and MOLLM and close to the in-context
 351 Ensemble Teacher. Although SFT obtains the lowest ACC, the multi-objective results show that
 352 this aggressive forgetting comes with considerably larger utility degradation and over-refusal. DUO
 353 therefore maintains a more favorable balance while remaining effective outside the QA evaluation
 354 format.

355 I Baselines under Standardized Training Data

356 To isolate the effect of data representation from that of the optimization objective, we additionally
 357 evaluate existing baselines under the standardized QA format. Several training-based methods

Table 6: Non-QA target-domain accuracy. Lower ACC indicates stronger removal of target-domain knowledge.

Method	ACC ↓
SFT	3.5
Ensemble Teacher	7.9
DUO	8.1
DUET	13.2
NGDiff	25.4
MOLLM	37.0

do not consistently benefit from standardization alone. Some preserve general utility but remain vulnerable to adversarial elicitation, whereas others achieve strong forgetting at the cost of retain performance. These results further indicate that data standardization is useful but insufficient by itself, and that practical multi-objective unlearning benefits from jointly designing the data representation and optimization objective.