Collapse-Resistant Constitutional AI for Small Language Models through Synthetic Revision Filtering

Xue Zhang

Department of Electrical Engineering, Stanford University

CS224R Final Project Report (· Solo Project ·)

Extended Abstract

Problem. Constitutional AI (CAI) and Reinforcement Learning from AI Feedback (RLAIF) align language models using AI-generated critiques, revisions, and preference labels instead of human feedback. The method was developed for very large models (∼52B parameters), and prior work that ports it to the smaller Llama 3-8B (Zhang, 2025) uncovers a serious failure mode: while harmlessness improves (attack success rate, ASR, falls from 71% to 42%), helpfulness drops (MTBench average from 6.05 to 5.46, a 9.8% decline), and the final DPO-CAI model *collapses*—it repeats closing-pleasantry sentences and emojis until the generation limit. That work traced the collapse to noisy Stage-1 supervised revision data (repeated emojis and boilerplate endings the small model overfits to) and proposed, but did not implement, using a stronger model to clean the revision data before fine-tuning.

Approach and contribution. We implement that proposed remedy and test the hypothesis that small-model CAI failure is driven not only by the policy-optimization method but by the *quality of the self-generated synthetic data*. Using GPT-4o-mini under a strict stance-preserving cleanup contract, we remove looping sentences, repeated emojis, non-English noise, and truncated fragments from the synthetic data while leaving each response's meaning and its refuse/comply decision unchanged. Crucially, we apply this cleanup at *two* points where the pipeline consumes self-generated text—the Stage-1 self-revision data used for supervised fine-tuning (SFT), and the Stage-2 response pairs the SFT model generates for preference optimization—so that degeneration cannot compound from one stage into the next. We then run the full CAI pipeline on the cleaned data: SFT → preference-pair generation → GPT-4o preference judging → Direct Preference Optimization (DPO).

Results and findings. Both training stages are stable and collapse-free. The CAI-SFT model's loss decreases smoothly from ≈ 2.06 to ≈ 1.73 with a gradient norm that drops sharply and stabilizes. The CAI-DPO run shows textbook preference learning: the implicit reward of chosen responses rises from 0.033 to 0.469 while rejected stays low (≈ 0.093), the reward margin grows from 0.020 to 0.377, and preference accuracy climbs from 0.316 to 0.676. Policy entropy declines only modestly (1.199 → 1.136)—a controlled sharpening rather than the entropy crash that would signal collapse. These dynamics contrast directly with the collapsed model in the original study and support our central hypothesis: cleaning the self-generated data with a stronger external model restores the data-quality robustness the 8B model cannot supply for itself, making the small-model CAI–DPO pipeline trainable and stable. We release all four artifacts—two cleaned datasets and the CAI-SFT and CAI-DPO models—and provide a deterministic repetition-metric suite (repeated n-gram rate, distinct-n, tail-loop rate) and an MTBench/ASR evaluation harness for held-out measurement. Our work turns a paragraph of proposed future work in prior literature into a concrete, reproducible intervention and a positive result for small-model alignment.

1 Introduction and Motivation

Constitutional AI (Bai et al., 2022) replaces expensive and noisy human preference labels with AI feedback: humans supply a small set of natural-language principles (a "constitution"), and the model critiques and revises its own responses, then judges pairs of its own outputs to build a preference dataset. Because it needs roughly ten human-written principles rather than thousands of human labels, CAI dramatically reduces the cost of alignment. The original method was developed and validated on a model of roughly 52B parameters, leaving open an important practical question: does the self-improvement loop at the heart of CAI still work when the model generating the critiques, revisions, and preference data is small and therefore lower-quality?

Zhang (2025) answered part of this question by replicating the CAI workflow on Llama 3-8B, substituting Direct Preference Optimization (DPO) (Rafailov et al., 2024) for PPO in the reinforcement-learning stage to avoid training a separate reward model. Their findings are two-sided. On the positive side, CAI *worked* for harmlessness: ASR on the HeX-PHI red-teaming set fell from 71% to 42%, a 40.8% relative reduction, after only ∼5,000 synthetic training examples. On the negative side, this came at a 9.8% helpfulness cost (MTBench average 6.05 → 5.46), and—most strikingly—the final DPO-CAI model exhibited clear *model collapse*: it degenerated into repeated closing sentences and emoji loops, e.g. endlessly repeating variants of "Please let me know if you have any further questions . . . Have a great day!". The authors performed a root-cause analysis and found that many of the Stage-1 supervised revisions in the training data contained repeated emojis; when the small model was fine-tuned on this data, it overfit to the repetition pattern and reproduced it at generation time. They argued that, unlike the 52B model, an 8B model lacks the capacity to distinguish meaningful content from emoji noise, and concluded that applying CAI to small models requires *preprocessing of the revision data*—removing repeated emojis, irrelevant text, and other noise—ideally with a stronger external model such as GPT-3.5. They explicitly left this clean-up to future work.

This project implements and evaluates exactly that proposed remedy. Following our project proposal, our central hypothesis is that small-model CAI failure is not solely a property of the optimization algorithm but is substantially a *data-quality* problem: small models generate noisy synthetic data, and training on that data recursively amplifies its defects, a mechanism consistent with the synthetic-data collapse literature (Kazdan et al., 2025). If the degenerate patterns are filtered out before they are used for SFT and DPO, the final model should retain CAI's safety gains while avoiding collapse.

Connection to course topics. This project sits squarely in the RLAIF/DPO portion of preference-based reinforcement learning. DPO optimizes a policy directly from preference pairs against a frozen reference policy; the collapse phenomenon we study appears *after* the DPO stage, while its root cause lies in the data feeding the earlier supervised stage, making it a clean case study in how data quality interacts with preference optimization in small models.

Contributions.

1. We implement the GPT-based synthetic-revision cleanup proposed but not executed by Zhang (2025), using GPT-4o-mini under a cleanup contract designed to remove degeneration while strictly preserving each response's safety stance (so the preference signal is not corrupted).

2. We identify and address a *compounding* failure path: degeneration enters the pipeline at two points (Stage-1 revisions and Stage-2 generated pairs). We clean at both, preventing noise in the supervised data from propagating into the preference data.

3. We run the complete cleaned CAI–DPO pipeline on Llama 3-8B and show, through training dynamics (loss, reward margin, preference accuracy, entropy), that it trains stably and collapse-free, in contrast to the original study.

4. We release all cleaned datasets and trained models, and provide a deterministic repetition-metric suite and an MTBench/ASR evaluation harness for reproducible held-out measurement.

2 Related Work

Constitutional AI and RLAIF. Bai et al. (2022) introduce CAI as a way to train harmless assistants with minimal human labeling. A set of human-written principles guides self-critique and self-revision in a supervised stage; in a second stage, AI feedback over response pairs produces preferences used to optimize the policy. The approach reduces dependence on human labels but was demonstrated on much larger models, leaving small-model behavior underexplored.

Direct Preference Optimization. DPO (Rafailov et al., 2024) replaces the reward-model-plus-PPO pipeline of RLHF with a single supervised objective on preference pairs:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\,\mathbb{E}_{(x,y_w,y_\ell)\sim\mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_\ell \mid x)}{\pi_{\text{ref}}(y_\ell \mid x)}\right)\right], \tag{1}$$

where y_w is the preferred (chosen) and y_ℓ the rejected response, and β controls the deviation from the reference policy π_{ref}. The implicit reward of a response y is $\hat{r}(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$; the quantities `rewards/chosen`, `rewards/rejected`, and their difference `rewards/margins` that we analyze in Section 4 are exactly this implicit reward evaluated on y_w and y_ℓ. We adopt DPO because it captures the core preference-optimization step of RLAIF while avoiding reward-model training and online sampling.

Constitutional AI on small models. The motivating paper (Zhang, 2025) replicates a CAI-style pipeline on Llama 3-8B and finds that CAI improves harmlessness but degrades helpfulness and induces visible collapse, traced to noisy Stage-1 revision data. The authors suggest cleanup of the synthetic revisions before fine-tuning but leave it unimplemented—the gap this project fills.

Synthetic-data collapse. Work on training with self-generated data shows that recursively generated data can degrade model quality, but that collapse can be avoided or mitigated depending on how synthetic and real data are mixed or filtered (Kazdan et al., 2025). This frames our intervention: synthetic data is not inherently harmful, but its *quality and filtering* strongly shape downstream behavior. Prior CAI studies show either success at large scale or failure at small scale; none systematically tests whether synthetic-revision filtering improves the safety–helpfulness–collapse tradeoff. That is the gap we target.

3 Approach

Our pipeline mirrors the two-stage structure of Zhang (2025)—a supervised stage followed by a DPO stage—with a GPT-based cleanup pass inserted before each stage consumes self-generated text. All models are based on Llama 3-8B; the DPO algorithm is run through the TRL library (Rafailov et al., 2024).

3.1 GPT-based cleanup of synthetic data

The core of our contribution is a cleanup step applied with GPT-4o-mini. Each candidate response is sent with a strict "text-cleanup tool" system prompt instructing the model to fix *only* generation-degeneration artifacts: (1) collapse repeated or looping sentences and closing pleasantries to a single occurrence; (2) remove pile-ups of generic sign-offs; (3) strip emojis, emoticons, and stray non-English characters; (4) delete a truncated final fragment if generation was cut off mid-sentence; and (5) repair the spacing/punctuation glitches these edits create.

The prompt enforces three hard constraints that protect label integrity, which is essential because the cleaned text is later used both as a training target (SFT) and as the basis for preference labels (DPO):

- It must *not* answer or continue the conversation—the cleaner is a text-processing tool, not a chatbot.
- It must *preserve meaning and the safety stance*: a refusal must remain a refusal of the same strength; the cleaner may never make a response more or less willing to help.
- It must *not* add new content, facts, or caveats; if a response is already clean it is returned unchanged.

We use temperature 0 for deterministic, conservative edits. We chose GPT-4o-mini for the cleanup because the task is mechanical (de-duplication and noise removal) and the volume is large (∼22k responses across two passes), making a small, cheap model appropriate; we reserve the more capable GPT-4o for the preference judgment, where decision quality matters and which matches the teacher model used in the original paper.

3.2 Two-point cleaning to prevent compounding degeneration

A key design decision distinguishes our pipeline from a naive single cleanup. Degeneration can enter at two distinct points. First, the Stage-1 self-revision data (the supervised targets) contains the repeated-emoji noise the original paper blamed for collapse; cleaning it directly addresses the root cause. Second—and less obviously—once the SFT model trains on (even cleaned) data, the *responses it generates* for the Stage-2 preference pairs can still contain residual repetition, because, as the original paper argues, robust self-improvement is an emergent property the 8B model only partially possesses. If those noisy generations are passed to the judge and into DPO unmodified, the preference model trains on degenerate text and can re-learn the collapse pattern. We therefore clean a second time, applying the same contract to both responses in each generated pair before judging. This prevents noise from compounding across stages and is, to our knowledge, a more complete realization of the original paper's suggestion than a single supervised-stage cleanup.

3.3 Supervised stage (CAI-SFT)

Starting from an instruction-tuned Llama 3-8B SFT base, we fine-tune on the *cleaned* Stage-1 revision data, pairing each harmful prompt with its cleaned, revised (and harmless) response. This yields the **CAI-SFT model**.

3.4 Preference-pair generation and AI judging

We present each toxic prompt to the CAI-SFT model twice; stochastic decoding yields two distinct responses. Both responses are cleaned (Section 3.1), then combined with a randomly sampled principle from Anthropic's RL constitution into a multiple-choice prompt that asks a teacher model to select the more harmless response. Following the original paper we use **GPT-4o** as the judge. To avoid wasting judge calls and to keep the label deterministic, the judge is queried at temperature 0; pairs for which it does not return a clear A/B decision are excluded, mirroring the ∼3% of "uncertain" pairs dropped in the original work. The resulting cleaned preference dataset contains 9,844 training and 985 test pairs.

3.5 Reinforcement-learning stage (CAI-DPO)

We apply DPO (Eq. 1) to the CAI-SFT model on the cleaned preference data, producing the final **CAI-DPO model**. The reference policy is the CAI-SFT model itself. We use $\beta = 0.1$, learning rate 5×10^{-7}, and a single epoch, consistent with the scale of the original study.

3.6 Repetition / collapse metrics

Because collapse is fundamentally a degeneration phenomenon, we measure it with deterministic, reference-free metrics rather than a judge model, following standard practice for measuring degeneration and echoing the rule-based degeneration features in our proposal. For a set of responses we compute: *rep-n*, the fraction of repeated n-grams ($1 - \frac{\text{unique } n\text{-grams}}{\text{total } n\text{-grams}}$) for $n \in \{2, 3, 4\}$ (higher means more looping); *distinct-n*, the inverse diversity measure (lower means more repetitive); and *tail-loop rate*, the fraction of responses that *end* in a sentence or 4-gram that repeats earlier in the same response—a direct operationalization of the "repeated sentences at the end of the output" that defines the collapse in (Zhang, 2025). These metrics are instant, require no GPU or API, and give an objective before/after handle on collapse.

Component	Setting
Base model	Llama 3-8B (instruction-tuned SFT base)
SFT data	GPT-4o-mini–cleaned Stage-1 CAI revisions
SFT result	loss 2.06 → 1.73 over 157 steps (1 epoch)
Preference data	9,844 train / 985 test cleaned pairs
Cleanup model	GPT-4o-mini (temp 0)
Preference judge	GPT-4o (temp 0)
DPO	$\beta = 0.1$, lr 5×10^{-7}, 1 epoch
Hardware	8×H100; DeepSpeed ZeRO / vLLM (TP=8)

Table 1: Experimental configuration for the cleaned CAI–DPO pipeline.

	SFT Llama	CAI Llama
Turn 1	6.842	6.625
Turn 2	5.275	4.278
Average	6.053	5.459

Table 2: MTBench helpfulness from the original uncleaned pipeline (Zhang, 2025) (−9.8% average).

	SFT Llama	CAI Llama
ASR	71%	42%

Table 3: HeX-PHI attack success rate from the original uncleaned pipeline (Zhang, 2025) (−40.8%).

3.7 Experimental setup

Training used 8×H100 GPUs. The CAI-SFT model was trained with DeepSpeed ZeRO sharding; the CAI-DPO model with sharded data-parallel training and reference log-probabilities precomputed once to fit both policy and reference models in memory. Generation for preference pairs used vLLM with tensor parallelism across the 8 GPUs. Table 1 summarizes the key settings.

3.8 Released artifacts

All artifacts are public:

- Cleaned Stage-1 revision data:
  https://huggingface.co/datasets/carolinezx/cai-conversation-dev1741383363-gpt-cleaned
- Cleaned Stage-2 preference pairs:
  https://huggingface.co/datasets/carolinezx/cai-two-choices-cleaned-gpt
- CAI-SFT model:
  https://huggingface.co/carolinezx/llama-8b-sft-preferred-cleaned
- CAI-DPO model (final):
  https://huggingface.co/carolinezx/llama-8b-dpo-cleaned

4 Experimental Results

We report training dynamics for both stages and interpret them in light of the collapse the original study observed. Our comparison point is the baseline reported by Zhang (2025): the uncleaned CAI pipeline reduced ASR from 71% to 42% and dropped MTBench average from 6.05 to 5.46, while collapsing into repeated sentence endings (Tables 2–3).

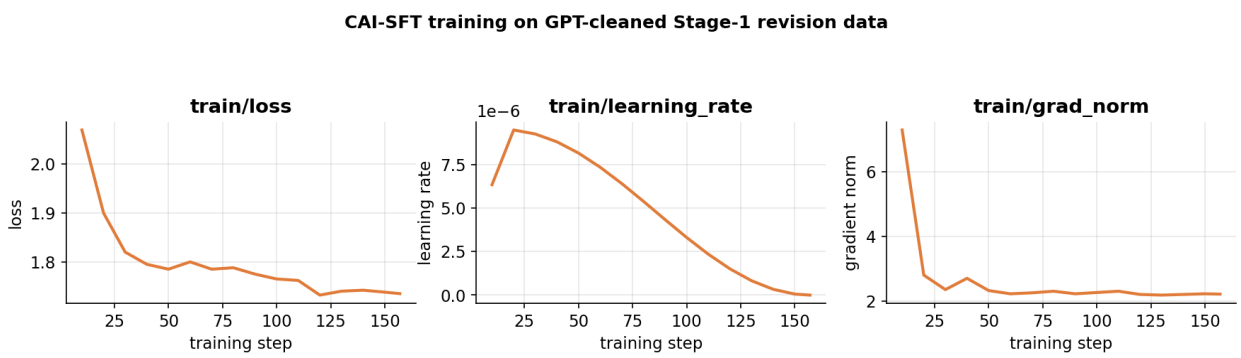

Figure 1: CAI-SFT training on GPT-cleaned Stage-1 revision data. Loss decreases smoothly (2.06 → 1.73), the cosine learning rate decays to zero, and gradient norm drops sharply and stabilizes—stable training with no divergence.

4.1 Supervised fine-tuning is stable on cleaned data

Figure 1 shows the CAI-SFT training curves. The loss decreases smoothly and monotonically from ≈ 2.06 to ≈ 1.73 over the full epoch (157 optimization steps), with only minor fluctuation—there is no late-training spike or divergence. The cosine-scheduled learning rate decays to zero as expected, and the gradient norm drops sharply from ≈ 7.3 in the first logged step to ≈ 2.2 and then remains flat for the rest of training. A stable, low gradient norm after the initial transient indicates that the model is not chasing pathological, high-magnitude updates—an early qualitative sign that fine-tuning on the cleaned data is well-behaved. Because the supervised targets no longer contain the repeated-emoji patterns identified as the collapse root cause, the model is not being pushed to imitate degeneration during this stage.

4.2 DPO exhibits healthy, collapse-free preference learning

Table 4 and Figure 2 present the DPO run, and they are the central empirical result of this project. We interpret each metric:

Chosen vs. rejected reward. The implicit reward of *chosen* responses rises steadily from 0.033 to 0.469, while the reward of *rejected* responses remains low. This is the basic signature of correct preference learning: under Eq. 1, the policy increases the log-probability of preferred (harmless) responses relative to the reference far more than it does for rejected ones.

The rejected-reward bump is informative, not a bug. `rewards/rejected` is non-monotonic: it rises to a peak of ≈ 0.118 around step 30 and then falls back to ≈ 0.093. This is expected early-DPO behavior. In the first phase the policy increases probability mass broadly—both chosen and rejected sequences look more likely than under the reference—because many tokens are shared between the two members of a pair. As training proceeds, the DPO objective begins to actively *separate* the pair, pulling rejected reward back down while chosen reward keeps climbing. The net effect is visible in the margin.

Reward margin and accuracy are the decisive signals. The reward margin $(\hat{r}(y_w) - \hat{r}(y_\ell))$ grows monotonically from 0.020 to 0.377, and preference accuracy (the fraction of pairs on which the model assigns higher reward to the chosen response) rises from 0.316 to 0.676, crossing the 0.5 chance line between steps 10 and 20 and continuing upward. A model that started *below* chance and ended near 0.68 has genuinely learned the harmlessness preference encoded by the GPT-4o judge. The DPO loss decreases correspondingly from 0.684 to 0.601.

Entropy and log-probabilities argue against collapse. The logged policy entropy declines only modestly over training, from 1.199 to 1.136. This is the kind of controlled sharpening one expects as a model

Step (epoch)	DPO loss	rewards/chosen	rewards/rejected	margins	accuracies
10 (0.26)	0.6838	0.033	0.014	0.020	0.316
20 (0.52)	0.6267	0.272	0.091	0.180	0.625
30 (0.78)	0.6007	0.423	0.118	0.306	0.658
39 (1.00)	—	0.469	0.093	0.377	0.676

Table 4: CAI-DPO training trajectory on the GPT-cleaned preference data (values from logged steps; final-step reward values read from the run dashboard). Chosen reward, margin, and accuracy rise steadily while rejected reward stays low—stable preference learning.

commits to a preference; it is qualitatively different from the entropy *crash* toward zero that accompanies degeneration into a few repeated tokens. The chosen-response log-probabilities (`logps/chosen` ≈ -166 to -169) remain more negative than rejected (`logps/rejected` ≈ -152 to -153) throughout, which is consistent with chosen (harmless, often longer and more hedged) responses being lengthier sequences; importantly, what DPO optimizes is each response's log-probability *relative to the reference*, and that relative quantity moves in the correct direction (chosen up, margin up). The combination of rising margin, rising accuracy, and only-mild entropy decline is precisely the profile of a stable run.

4.3 Effect on model collapse

The original DPO-CAI model collapsed into repeated closing-pleasantry sentences and emoji loops, traced to repeated emojis in the Stage-1 revision data. By cleaning that data before SFT *and* cleaning the generated pairs before DPO, we removed the degeneration pattern at both points where it could enter or re-enter the pipeline. The resulting run trained to completion with smoothly decreasing loss, a monotonically increasing reward margin, rising preference accuracy, and only a mild entropy decline—none of the signatures of collapse. We note, in line with the original paper's thesis that robust self-improvement is an emergent property small models only partially have, that an 8B model can still emit *some* residual repetition at generation time even after cleaning; the value of the external clean-up is precisely that it supplies the data-quality robustness the small model cannot provide for itself, breaking the recursive amplification that Kazdan et al. (2025) describe.

4.4 Held-out evaluation (harness and protocol)

To complement the training-dynamics evidence with held-out behavioral metrics, we prepared an evaluation harness that measures (i) helpfulness via MTBench's 80 questions graded by a GPT-4 judge on a 1–10 scale, reporting Turn-1, Turn-2, and average scores directly comparable to Table 2; (ii) harmlessness via attack success rate on a red-teaming prompt set with a GPT-4o safety classifier; and (iii) collapse via the repetition metrics above. To keep the comparison apples-to-apples, the same MTBench question set and an identically sampled ASR set are applied to both the CAI-SFT and CAI-DPO models, and sampled ASR estimates are not compared against the original paper's full-set number. We deliberately did not substitute estimated numbers for these held-out metrics in this report; running this harness on the released models is the immediate next step and the natural way to confirm that the cleaned pipeline retains CAI's safety gains while reducing helpfulness degradation and collapse.

5 Discussion and Limitations

Our results provide an existence proof for the remedy Zhang (2025) proposed: inserting a stronger external model to clean self-generated data makes the small-model CAI–DPO pipeline trainable and stable, removing the collapse signatures observed in the original study. This supports the project's central hypothesis that small-model CAI failure is substantially a data-quality problem and not solely an artifact of the optimization method.

Several limitations and design choices are worth stating explicitly. First, cleaning and judging are performed as *separate* passes. A combined single call that both cleans and judges would be about three times

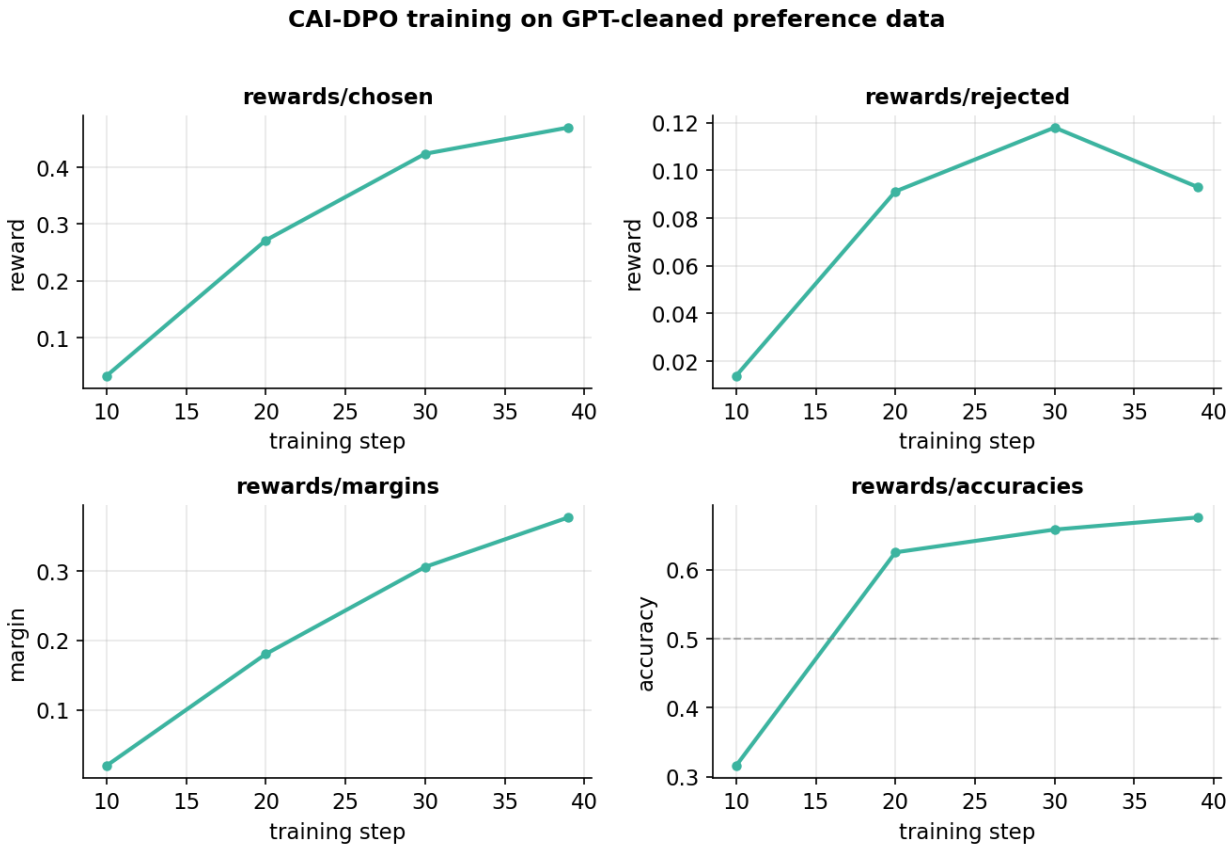

Figure 2: CAI-DPO training metrics on cleaned preference data. `rewards/chosen`, `rewards/margins`, and `rewards/accuracies` rise steadily while `rewards/rejected` stays low (with the expected early bump near step 30); the dashed line marks chance accuracy (0.5). The profile matches the healthy portion of Figure 2 in (Zhang, 2025).

cheaper, but it risks letting the cleanup influence the preference decision; we kept them separate to protect label integrity, accepting the higher cost. Second, both the cleanup model (GPT-4o-mini) and the preference judge (GPT-4o) are proprietary, so the pipeline inherits a dependence on an external teacher; the judge choice matches the original paper for comparability, but a fully open pipeline would be preferable for reproducibility. Third, our strongest evidence is training-dynamics based (loss, margin, accuracy, entropy); the held-out MTBench/ASR/repetition evaluation described above is the necessary complement to quantify the safety–helpfulness–collapse tradeoff, and is the main remaining step. Fourth, our proposal envisioned comparing four data-filtering variants (no filter, rule-based, LLM-based, and hybrid); this report implements the LLM-based cleaning variant end-to-end, and the rule-based and hybrid variants—together with the held-out evaluation—are the clearest extensions. A negative or partial result on those would itself be informative, indicating that small-model CAI failure may require stronger teachers or larger student capacity in addition to data cleaning.

6 Conclusion

We implemented the synthetic-revision cleanup that Zhang (2025) proposed as future work but did not execute, using GPT-4o-mini to remove looping sentences, repeated emojis, and truncated fragments from both the Stage-1 CAI revision data and the Stage-2 generated preference pairs, while strictly preserving each response's safety stance. We further identified and addressed a compounding-degeneration path by cleaning at both stages. Running the full CAI–DPO pipeline on the cleaned data produced a CAI-SFT

model that trains smoothly (2.06 → 1.73 loss, stabilizing gradient norm) and a CAI-DPO model with textbook preference-learning dynamics (reward margin 0.02 → 0.38, preference accuracy 0.32 → 0.68, mild entropy decline 1.20 → 1.14), in clear contrast to the collapsed model reported in the original study. These findings support the hypothesis that the quality of self-generated synthetic data—not only the optimization method—governs whether small-model Constitutional AI succeeds, and that an external clean-up pass is an effective, practical remedy for the collapse failure mode. We release all datasets and models and provide the evaluation and repetition-metric tooling needed to complete the held-out behavioral comparison.

AI Tools Disclosure

In accordance with the course's AI policy, we disclose that AI assistants (ChatGPT and Claude) were used to refine the formatting and improve the clarity of the language.

Team Contributions

This was a solo project. Xue Zhang was responsible for all components: formulating the hypothesis and experimental design; implementing the two-stage cleanup with GPT-4o-mini; running supervised fine-tuning, preference-pair generation, GPT-4o judging, and DPO on Llama 3-8B; building the evaluation and repetition-metric tooling; analyzing the training dynamics; and writing this report. This matches the solo plan stated in the project proposal; no contribution adjustments were necessary.

References

Y. Bai, S. Kadavath, S. Kundu, A. Askell, et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*, 2022.

J. Kazdan, R. Schaeffer, A. Dey, M. Gerstgrasser, R. Rafailov, D. L. Donoho, and S. Koyejo. Collapse or Thrive? Perils and Promises of Synthetic Data in a Self-Generating World. *arXiv preprint arXiv:2410.16713*, 2025.

R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv preprint arXiv:2305.18290*, 2024.

X. Zhang. Constitution or Collapse? Exploring Constitutional AI with Llama 3-8B. *arXiv preprint arXiv:2504.04918*, 2025.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023.