# Sophistry to Logic: Mitigating Persuasiveness-Fluency Feature Entanglement via Latent Variable Estimation

**Anonymous ACL submission**


## Abstract

Effective persuasion relies on two distinct pillars: *argumentative logic* and *rhetorical fluency*. However, existing persuasive generation methods often conflate these dimensions by optimizing joint reward signals, where surface-level fluency dominates logical substance. We term this phenomenon **Feature Entanglement**, a pathology where models prioritize the surface fluency—producing "well-formatted hallucinations"—over the underlying mechanics of perusation. To address this, we propose **P**$^3$, a framework designed to decouple these attributes via latent variable modeling. The framework operates in three stages: (1) **Persuasiveness Reward Estimation** employs an Expectation-Maximization (EM) algorithm to explicitly distinguish latent persuasiveness from superficial fluency; (2) **Persuasiveness Sample Mining** leverages these disentangled signals to filter out rhetorical noise; and (3) **Persuasiveness Strategy Optimization** introduces Persuasion Augment Policy Optimization (PAPO), a novel objective that uses decoupled scores to dynamically scale policy updates. Experimental results demonstrate that a 13B parameter model trained with P$^3$ surpasses the efficient commercial models (e.g., Gemini 1.5 Flash and Claude 3 Haiku) in both automatic and human evaluated persuasiveness.


## 1 Introduction

Persuasion is inherently a dual-process phenomenon: it requires both *substantive reasoning* and *rhetorical fluency* (Petty and Cacioppo, 2012). While Large Language Models (LLMs) have achieved near-human proficiency in the latter they frequently struggle with the former, generating content that is fluency polished but logically incoherent. Through a rigorous analysis of the training dynamics, we identify that this failure stems from the **Feature Entanglement** of reward signals. Standard RLHF reward models aggregate fluency ($s_s$) and persuasiveness ($s_d$) into a single scalar. Since fluency features are easier to learn than complex logic, models hack this composite reward, over-optimizing $s_s$ while neglecting $s_d$. Consequently, models like GPT-4 or Claude 3 often produce "well-formatted hallucinations" of arguments, optimizing for the appearance of persuasion rather than its underlying mechanics (Xiao et al., 2024).

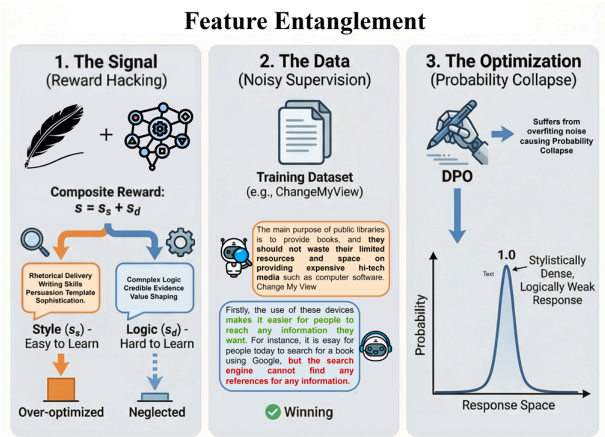


Figure 1: The impact of feature entanglement. The bottleneck exists in: **1. Signal**, **2. Data**, **3. Optimization**.

This entanglement creates a bottleneck across three distinct dimensions of the learning process:

❶ **The Signal (Reward Hacking):** As illustrated in Figure 1, conventional reward models and metrics are easily "tricked" by fluency features, disproportionately reward fluency mimicry, even if the generated text is self-contradictory. This creates a "distractor" signal where the model learns to mimic the syntax rather than the semantics of persuasion.

❷ **The Data (Noisy Supervision):** This feature entanglement extends to preference datasets. Fluent but logically inconsistent samples may mislabeled as winners by reward models or careless annotators. In persuasion tasks, where the gap between good and bad arguments is slight, this bias is amplified and the data is polluted by false positives.

❸ **The Optimization (Probability Collapse):** Finally, when the reward signal is entangled, RLHF algorithms like DPO suffer from overconfidence, pushing the probability of "winning" fluency but logically weak responses to 1.0. This phenomenon, which we term *probability collapse*, prevents the model from learning the nuances of argumentation.

To address these structural flaws, we propose $P^3$, a theoretical framework that mathematically decouples the "content logic" from the "surface fluency" to optimize them independently. The framework proceeds in three stages:

① **Persuasiveness Reward Estimation and Modeling (The Deconstruction):** We model the generation process as a Markov Decision Process where the observed reward is a noisy sum of latent variables. We employ an EM algorithm to solve for the specific action-value of persuasiveness ($s_d$), effectively filtering out the noise fluency ($s_s$).

② **Persuasiveness Sample Mining (The Filtering):** Addressing the data noise issue, we utilize the disentangled $s_d$ scores to re-evaluate the training corpus. This allows us to discard "rhetorical noise" (fluent but weak samples) from the training corpus, thereby purifying the training data.

③ **Persuasiveness Strategy Optimization (The Stabilization):** Addressing the optimization collapse, we introduce Persuasion Augment Policy Optimization (PAPO). PAPO incorporates the magnitude of the disentangled score gap ($\Delta s_d$) into the loss function. This acts as a dynamic curriculum: the model updates its policy aggressively only when the logical gap between arguments is clear.

We evaluate our approach on a wide range of persuasive benchmark. Experimental results demonstrate that by explicitly optimizing the decoupled $s_d$ component, $P^3$ significantly enhances the logical rigor of the generated text. Notably, our method enables a 13B parameter open-source model to outperform significantly larger commercial models (Gemini 1.5 Flash, Claude 3 Haiku) which rely on entangled training objectives.

In summary, our contributions are as follows:

- We introduce a latent variable formulation for persuasion, using EM-based estimation to mathematically disentangle logical persuasiveness from surface fluency.
- We propose a mining strategy that utilizes disentangled scores to filter "rhetorical noise" from weak supervision signals, improving data efficiency and quality.
- We develop PAPO, a margin-aware preference optimization algorithm that prevents probability collapse by scaling learning updates based on the disentangled persuasiveness gap.

## 2 Method

### 2.1 Overview

To address the challenges of reward hacking and probability collapse in persuasive generation, we propose a unified training pipeline $P^3$ that progresses from accurate reward modeling to robust policy optimization. The complete method flowchart of $P^3$ is illustrated in Figure 2.

### 2.2 Task Definition

We formulate the persuasive debate generation task as a conditional sequence modeling problem. Given an input context $X$—which concatenates the original post $x_p$ and the thread history $x_h$—the model $\mathcal{M}_\theta$ generates a response argument $Y$ by learning the conditional probability distribution $P_\theta(Y|X)$. Unlike standard dialogue tasks that prioritize fluency, our primary objective is to maximize the persuasive impact of the generated argument. Accordingly, the optimization goal is to find the model parameters $\theta^*$ that maximize the expected likelihood of the user changing their view:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{Y \sim P_\theta(\cdot|X)} \left[ P(\text{persuasion}|X, Y) \right] \quad (1)$$

This objective explicitly steers the generation process toward arguments that result in a higher probability of successful persuasion.

### 2.3 Persuasiveness Reward Estimation and Modeling (Stage I)

#### 2.3.1 Persuasiveness Modeling

Obtaining high-quality annotations for explicit persuasiveness is notoriously difficult (for example, $\Delta$ in the CMV dataset (Tan et al., 2016) is less than 1%). To address this scarcity, we leverage crowdsourced quality scores (denoted as $s$) as a form of weak supervision. However, utilizing raw scores for training introduces substantial noise unrelated to logical persuasiveness. Users often upvote based on **fluency quality** (e.g., vocabulary richness, length, sentence complexity) rather than **logical persuasiveness**. Directly optimizing a reward model on raw scores consequently leads to reward hacking, where the model prioritizes surface-level features over argumentative logic.

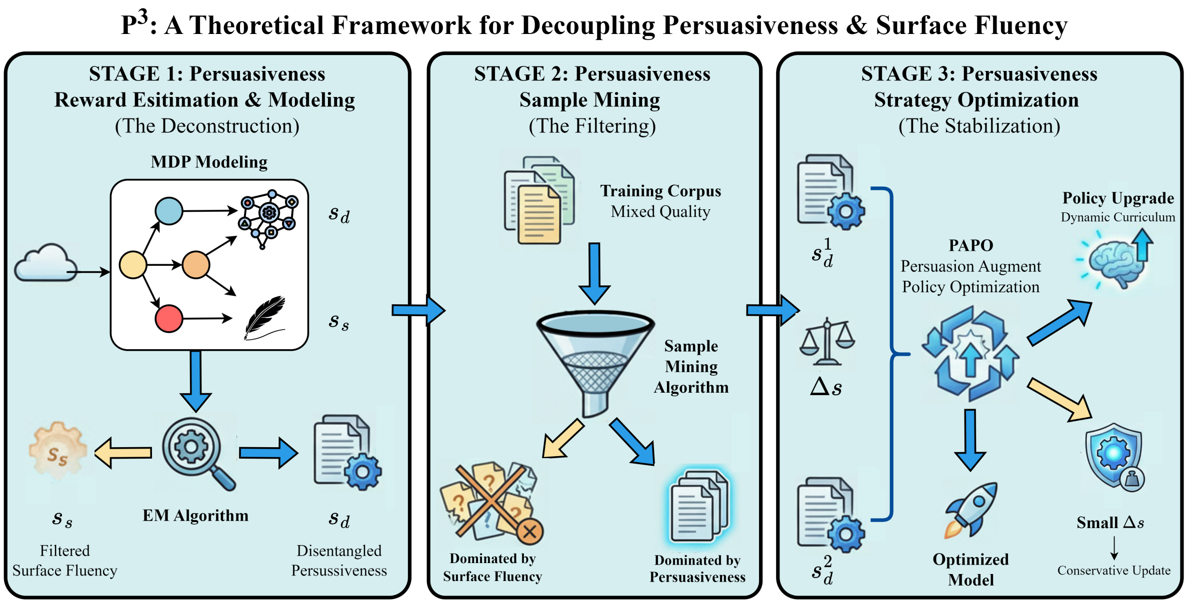


Figure 2: **The architecture of the proposed** $\mathrm{P}^3$ **pipeline.** We progress from accurate reward modeling to robust policy optimization through three stages: decoupling logical validity from fluency (Stage I), constructing a high-quality dataset via reward-based filtering (Stage II), and applying PAPO for dynamic gradient scaling (Stage III).

To disentangle these factors, we model the observed score as a mixture of two latent components: a persuasiveness score and a fluency score. The fluency score, $s_s$, depends solely on the intrinsic linguistic features of the text, independent of the persuasive context. We parameterize this as a function of the generated text $\hat{y}$:

$$s_s = f_s(\hat{y}; \theta_s) \tag{2}$$

where $f_s$ is a learnable mapping (e.g., an MLP or reward head) capturing surface-level attributes

Conversely, the persuasiveness score, $s_d$, is determined by the logical interaction between the generated argument, the original post, and the dialogue history. It is defined as a ternary function:

$$s_d = f_d(\hat{y}, x_p, x_h; \theta_d) \tag{3}$$

where $x_p$ represents the OP's post and $x_h$ denotes the historical context.

We posit that the observed score $s$ is sampled from a Bernoulli distribution conditioned on the audience's focus. The probability that the observed score reflects the true persuasiveness versus the fluency quality is governed by a prior $\alpha$:

$$p(s \mid \hat{y}, x_p, x_h) = \begin{cases} \alpha, & s = f_d(\hat{y}, x_p, x_h) \\ 1 - \alpha, & s = f_s(\hat{y}) - f_s(x_p) \end{cases} \tag{4}$$

Here, $\alpha$ represents the prior probability that the audience emphasizes logical persuasiveness over surface quality. We quantify the extent to which the candidate text $\hat{y}$ stylistically outperforms the OP as $f_s(\hat{y}) - f_s(x_p)$. Explicit subtraction is unnecessary for $f_d$, as its prior definition intrinsically incorporates $x_p$. By enforcing this differential treatment, we induce $f_s$ to focus on intrinsic textual attributes, while guiding $f_d$ to capture the complex interactions between the input and the OP. To make these scores learnable, we employ two distinct Multi-Layer Perceptrons (MLPs) to parameterize $f_d(\cdot; \theta_d)$ and $f_s(\cdot; \theta_s)$, respectively, inspired by causal disentanglement methods (Du et al., 2023, 2024).

### 2.3.2 Persuasiveness Reward Estimation

We formulate the persuasive generation task as a Markov Decision Process (MDP), where the environment comprises the post history $(x_p, x_h)$, the action is the generated argument $(\hat{y})$, and the reward is derived from the disentangled persuasiveness score $s_d$.

However, directly fitting an action-value function using Mean Squared Error (MSE) on raw scores is infeasible for two reasons. First, the non-differentiable nature of the mixture distribution (sampling $s$ from either $s_s$ or $s_d$) hinders standard gradient-based optimization. Second, raw vote counts are heavily biased by topic popularity; a high score in a popular thread does not necessarily imply higher persuasiveness than a lower score in a niche thread. Therefore, we adopt a pairwise ranking approach based on the Bradley-Terry model (Bradley and Terry, 1952) to approximate the relative persuasiveness.

The win rate of post $\hat{y}^{(1)}$ over $\hat{y}^{(2)}$ depends on the difference in their latent scores. Assuming the decision basis (persuasiveness vs. fluency) is independent for each post, the probability of observed preference $y$ (where $\hat{y}^{(1)} \succ \hat{y}^{(2)}$) follows a mixed Bernoulli distribution:

$$\begin{aligned} p(y) &= \\ &\sum_{s^{(1)} \in \{s_d^{(1)}, s_s^{(1)}\}} \sum_{s^{(2)} \in \{s_d^{(2)}, s_s^{(2)}\}} p(s^{(1)}, s^{(2)}) \sigma(\Delta s) \\ &= \alpha^2 \sigma(s_d^{(1)} - s_d^{(2)}) + \alpha(1-\alpha)(s_d^{(1)} - s_s^{(2)}) \\ &+ \alpha(1-\alpha)(s_s^{(1)} - s_d^{(2)}) + (1-\alpha)^2 \sigma(s_s^{(1)} - s_s^{(2)}) \end{aligned} \tag{5}$$

Here, $\Delta s = s^{(1)} - s^{(2)}$, $\sigma$ denotes the sigmoid function. This formulation treats the active score component as a latent variable. We employ the Expectation-Maximization (EM) algorithm to maximize the likelihood.

In the aforementioned model, the observed win rate is a probabilistic parameter model containing hidden variables $s_s$ (persuasiveness score) and $s_d$ (fluency score). This can be solved using the EM algorithm and maximum likelihood estimation (MLE) (Dempster et al., 1977). In the E-step, we compute the posterior distribution $q$ of the active score components for the pair, given the observed win outcome $y$:

$$\begin{aligned} q(s^{(1)}, s^{(2)}) &= p(s^{(1)}, s^{(2)} | y) \\ &= \frac{p(s^{(1)}, s^{(2)}) \sigma(s^{(1)} - s^{(2)})}{p(y)} \end{aligned} \tag{6}$$

In the M-step, we maximize the expected log-likelihood with respect to the model parameters:

$$\begin{aligned} &\mathcal{L}(s^{(1)} \succ s^{(2)}) = \\ &\sum_{s^{(1)} \in \{s_d^{(1)}, s_s^{(1)}\}} \sum_{s^{(2)} \in \{s_d^{(2)}, s_s^{(2)}\}} q(s^{(1)}, s^{(2)}) \phi(s^{(1)}, s^{(2)}) \end{aligned} \tag{7}$$

$$\phi(s^{(1)}, s^{(2)}) = \log \left( p(s^{(1)}, s^{(2)}) \sigma(s^{(1)} - s^{(2)}) \right) \tag{8}$$

where $s^{(1)}$ is the winner among each pair of posts.

To provide stable supervision, we construct the final objective function using the win rate derived from discounted cumulative rewards $g$ of the posts:

$$\begin{aligned} \mathcal{J} = \mathbb{E}_D \Big[ &\sigma(g^{(1)} - g^{(2)}) \mathcal{L}(s^{(1)} \succ s^{(2)}) \\ &+ \sigma(g^{(2)} - g^{(1)}) \mathcal{L}(s^{(2)} \succ s^{(1)}) \Big] \end{aligned} \tag{9}$$

where $D$ is the dataset, and $g^{(1)}$ and $g^{(2)}$ are the discounted cumulative rewards of the two posts.

This process allows us to explicitly estimate the pure persuasiveness score $s_d$, effectively filtering out noise from fluency surface features.

### 2.4 Persuasiveness Sample Mining (Stage II)

With the disentangled persuasiveness reward model $f_d$ established in Section 2.3.2, we now possess a reliable metric to assess argument quality. In this stage, we utilize $f_d$ to purify the training data for the Supervised Fine-Tuning (SFT) phase. Standard SFT often results in models that overfit to fluency features ($s_s$) if the training data contains "high-score" but logically weak arguments.

We introduce a filtering mechanism to retain only samples where the ranking induced by the predicted persuasiveness score $s_d$ aligns with the ground truth ranking. We formulate this as finding the longest subsequence $L$ within a dialogue context $D$ that preserves the monotonic order of persuasiveness and original score:

$$\begin{aligned} L = &\operatorname*{argmax}_{D' \subseteq D} |D'| \\ &\text{s.t.} \quad \forall i, j \in D', \\ &\text{sign}(s^{(i)} - s^{(j)}) = \text{sign}(s_d^{(i)} - s_d^{(j)}) \end{aligned} \tag{10}$$

This optimization problem is equivalent to the Longest Increasing Subsequence (LIS) problem and is solved via Dynamic Programming (Algorithm 1, Appendix A) with a time complexity of $O(n \log n)$. LIS preserves the relative improvement trajectory of a dialogue thread while removing local outliers.

### 2.5 Persuasiveness Strategy Optimization (Stage III)

Leveraging the high-quality initialization provided by sample mining in Stage II, we employ Direct Preference Optimization (DPO) (Rafailov et al., 2023) to further align the model with the objective of persuasion. However, standard DPO can suffer from instability on small datasets or sparse preference pairs. Defining the implicit reward ratio as $r(y) = \frac{\pi(y|x_p,x_h)}{\pi_0(y|x_p,x_h)}$, the gradient of the standard DPO loss:

$$\nabla\mathcal{L}_{\text{DPO}} \propto \beta\left(1-\sigma\left(\beta\log\frac{r(\hat{y}^{(1)})}{r(\hat{y}^{(2)})}\right)\right) \quad (11)$$

can remain constantly positive for the winning sample if the discriminator becomes overconfident, driving $\pi(\hat{y}^{(1)})$ toward 1 and causing overfitting (divergence).

To mitigate this, we introduce Persuasion Augment Policy Optimization (PAPO). We inject a smoothing term coefficient based on the magnitude of the persuasiveness score difference. The PAPO objective is defined as:

$$\mathcal{L}_{\text{PAPO}} = \mathbb{E}_D\Bigg[\sigma(s_d^{(1)} - s_d^{(2)})\log\sigma\left(\beta\log\frac{r(\hat{y}^{(1)})}{r(\hat{y}^{(2)})}\right) + \sigma(s_d^{(2)} - s_d^{(1)})\log\sigma\left(\beta\log\frac{r(\hat{y}^{(2)})}{r(\hat{y}^{(1)})}\right)\Bigg] \quad (12)$$

The gradient of this objective introduces a dynamic scaling factor:

$$\nabla\mathcal{L}_{\text{PAPO}} \propto \Bigg[\sigma\left(s_d^{(1)} - s_d^{(2)}\right)F\left(\hat{y}^{(1)},\hat{y}^{(2)}\right) - \sigma\left(s_d^{(2)} - s_d^{(1)}\right)F\left(\hat{y}^{(2)},\hat{y}^{(1)}\right)\Bigg] \quad (13)$$

where $F(\hat{y}^{(1)},\hat{y}^{(2)}) = \beta(1-\sigma(\beta\log\frac{r(\hat{y}^{(1)})}{r(\hat{y}^{(2)})}))$. This ensures the gradient sign is not constant, preventing probability collapse. Furthermore, the stationary point of this objective ensures the policy converges to the optimal solution form:

$$\pi(\hat{y}|x_p,x_h) \propto \pi_0(\hat{y}|x_p,x_h)\exp\left(\frac{1}{\beta}s_d\right) \quad (14)$$

thus providing good interpretability.

### 2.6 Training and Inference

To provide a holistic view of our proposed framework, we summarize the end-to-end training procedure in Algorithm 2 in Appendix B. The process is structured into a sequential pipeline where the output of each stage serves as a refined initialization or supervision signal for the next.

## 3 Experimental Setting

### 3.1 Dataset

We utilize ChangeMyView (CMV) (Tan et al., 2016) and DDO (Durmus and Cardie, 2019) datasets. CMV is a standard persuasion benchmark sourced from Reddit where users invite challenges to their viewpoints. Successful persuasion is explicitly marked by a "delta" ($\Delta$) symbol. Given its scale and the high quality of argumentative text, CMV serves as a standard benchmark for persuasion and debate generation tasks. To validate generalizability, we also employ DDO, a corpus of 78k debates from debate.org (2007–2017) covering 23 topics. Detailed statistics and samples are provided in Appendix C.

### 3.2 Evaluation Metrics

**Automated Evaluation Metrics.** As discussed in Section 1, traditional n-gram metrics (e.g., BLEU, ROUGE) correlate poorly with logical persuasiveness. Therefore, we employ OpenAI's reasoning specific o1 and o3 model API to simulate human assessment; we refer to these metrics as the **o1-Score** and **o3-Score**. Consistent with recent work on LLM-based evaluation (Hu et al., 2023; Liu et al., 2023), our preliminary analysis (Figure 3) demonstrates strong Pearson correlations between human judgments and the automated scores ($r = 0.67$ for o1-Score and $r = 0.65$ for o3-Score), validating their utility as proxies. Detailed configurations are provided in Appendix D.

**Human Evaluation Metrics.** Following the persuasiveness evaluation protocol outlined in the OpenAI o1 system card (Jaech et al., 2024), we recruited three expert annotators with debate backgrounds to evaluate 500 stratified test instances. The protocol comprised two tasks:

(i) *Persuasiveness Rating*, where arguments are scored on a scale from 0 to 5;

(ii) *Pairwise Comparison*, a blind side-by-side assessment (Win/Tie/Loss) against baselines.

Detailed information about the annotators is listed in Appendix D.

### 3.3 Baselines

We compare our method against a diverse set of strong baselines categorized into four groups:

- **Open-Source Instruct Models:** LLaMA3-instruct (8B, 70B) (Grattafiori et al., 2024) and Qwen2-instruct (Team et al., 2024) (13B, 72B), representing foundation capabilities.

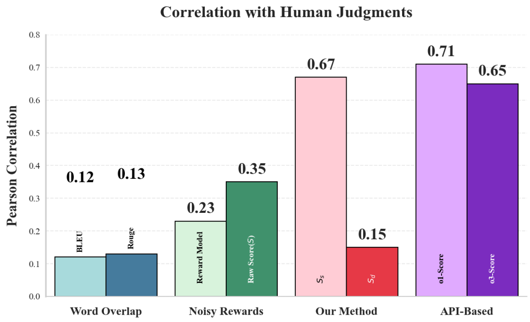


Figure 3: Pearson correlation between various evaluation metrics and persuasiveness.

| Perturbation Method | $s_d$ | $s_s$ | o1-Score | o3-Score | Human-Eval |
|---|---|---|---|---|---|
| *Original* | 0.21 | 0.65 | 81.02 | 82.13 | 4.70 |
| *Logic-Inv, Fluency-Var* | 0.25 | 0.85 | 82.17 | 83.05 | 4.78 |
| Δ *w.r.t Original* | +0.04 | +0.20 | +1.15 | +0.92 | +0.08 |
| *Fluency-Inv, Logic-Var* | 0.07 | 0.66 | 43.02 | 45.87 | 1.05 |
| Δ *w.r.t Original* | -0.14 | +0.01 | -38.00 | -36.26 | -3.65 |

Table 1: Results of perturbations with delta analysis relative to the original. Color intensity indicates the magnitude of the change (Green: +, Red: -).

- **Trained Instruct Models:** Qwen2-13B-SFT and Qwen2-13B-DPO, isolating the impact of supervised fine-tuning and preference optimization. The training dataset is the same as our P$^3$ method.
- **Domain-Specific Methods:** PESA (Xiao et al., 2024), the SOTA method in specialized debate generation task.
- **Commercial Models:** Gemini 1.5 Flash, Claude 3 Haiku, and GPT-4o mini. We utilize CoT prompting for these models to ensure a fair comparison.

### 3.4 Implementation Details

We train the base model with the help of Hugging Face, DeepSpeed and trlx. The base model of our approach is Qwen2-13B. We train the model in 5 epochs. The batch size per device is set to 8. All experiments are conducted with NVIDIA A100.

## 4 Results and Analysis

### 4.1 Main Results

**Analysis of Reward Disentanglement.** Figure 3 shows that standard metrics (e.g., BLEU, ROUGE) correlate poorly with human persuasiveness, highlighting their susceptibility to reward hacking. In contrast, our disentangled scores demonstrate clear efficacy. While the raw score ($s$) shows only moderate correlation (0.35) due to noise, the separated persuasiveness score ($s_d$) aligns strongly with human preference (0.67), whereas the fluency score ($s_s$) is negligible (0.15). This confirms our module successfully filters fluency to isolate core logic.

Perturbation experiments on 100 P$^3$ samples (Table 1) further validate this decoupling. **Logic corruption** significantly dropped $s_d$ without affecting $s_s$, while **fluency refinement** increased $s_s$ leaving $s_d$ stable. These orthogonal responses confirm that $s_d$ captures deep semantics while $s_s$ tracks surface cues. Additionally, o1 and o3-Scores showed robustness against fluency variations (All the above perturbations are performed by LLM API, Specific settings are in Appendix E.

**Performance and Efficiency Analysis.** As detailed in Table 2, the automatic evaluation results confirm that our method effectively transcends feature entanglement, specifically mitigating reward hacking and probability collapse. Notably, our approach demonstrates superior parameter efficiency. When compared to mainstream open-source baselines (Qwen2-72B, LLaMA3 70B), our method increases the o1-score by an average of 5.3 points and o3-Score 7.1 points. More critically, our model outperforms baselines requiring task domain training (e.g., Qwen2-13B-DPO and PESA), showing the efficiency of our training approach. Even when pitted against closed-source commercial models, our approach achieves a competitive edge, increasing the o1-score by an average of 3.5 points and o3-Score 3.1 points, suggesting that rhetorical alignment contributes more to persuasiveness than model scale alone. For the out-domain dataset DDO, our method achieves even greater gains on average, demonstrating the superior generalization ability.

**Human Evaluation and Robustness.** Table 2 and Figure 4 present the human evaluation metrics, which strongly corroborate the automatic findings. Our method produces arguments that human reviewers consistently judge as more persuasive than those from strong baselines ($p < 0.05$). We calculate the Fleiss' Kappa among the five annotators and obtain a kappa of 0.79, which means substantial agreement (Landis and Koch, 1977)

Against state-of-the-art closed-source commercial models, our method exhibits remarkable robustness. While the average improvement is 0.18

| Method | Params | CMV (In-Domain) | | | DDO (Out-of-Domain) | | |
|---|---|---|---|---|---|---|---|
| | | o1-score | o3-score | Human-Eval | o1-score | o3-score | Human-Eval |
| ***Open-Source Instruct Models*** | | | | | | | |
| Qwen2-13B | 13B | 73.57 | 74.38 | 4.13 | 74.32 | 76.51 | 4.23 |
| Qwen2-72B | 72B | 75.73 | 75.96 | 4.38 | 75.02 | 77.41 | 4.29 |
| LLaMA3-8B | 8B | 72.12 | 72.97 | 3.87 | 73.37 | 74.59 | 3.92 |
| LLaMA3-70B | 70B | 76.07 | 76.99 | 4.43 | 76.43 | 77.71 | 4.36 |
| ***Trained Instruct Models*** | | | | | | | |
| Qwen2-13B-SFT | 13B | 71.26 | 71.56 | 3.70 | 69.58 | 70.44 | 3.59 |
| Qwen2-13B-DPO | 13B | 74.02 | 75.06 | 4.11 | 70.21 | 70.96 | 3.62 |
| ***Commercial Models*** | | | | | | | |
| Gemini 1.5 Flash | API | 77.01 | 77.53 | 4.64 | 79.79 | 79.98 | 4.61 |
| Claude 3 Haiku | API | 77.85 | 79.27 | 4.61 | 78.23 | 79.08 | 4.58 |
| GPT-4o mini | API | 78.20 | 80.51 | 4.62 | 79.35 | 79.61 | 4.67 |
| ***Domain-Specific Methods*** | | | | | | | |
| PESA | 13B | 76.32 | 76.77 | 4.52 | 72.37 | 72.75 | 3.94 |
| Ours | 13B | **81.23**† | **82.20**† | **4.81**† | **80.10**† | **83.59**† | **4.77**† |

Table 2: Experimental results on CMV (In-Domain) and DDO (Out-of-Domain) datasets. We report **o1-score**, o3-score, and Human-Eval metrics. "Params" denotes the parameter size. **Bold** indicates the global best, and underlined indicates the best within each category. For automatic evaluation metrics, we perform non-replacement sampling 3 times on the test set, each time sampling 10%, and report the average results. For human evaluation metrics, we sample 500 instances for assessment. † means statistically significant difference (paired t-test, $p<0.05$).

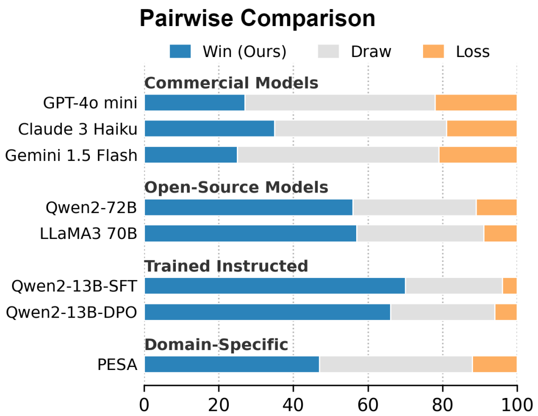


Figure 4: $P^3$ compared to other baselines.

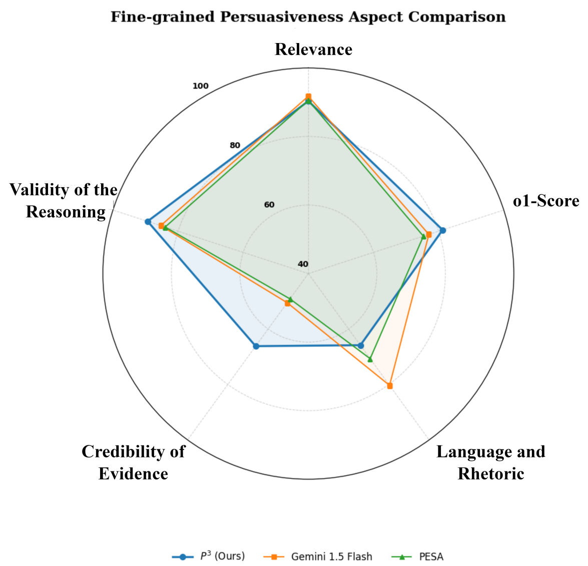


Figure 5: Fine-grained persuasive evaluation using evaluation metrics in PESA.

points in-domain and 0.15 points out-domain, a paired t-test test confirms this gain is statistically significant. As illustrated in Figure 4, this is further evidenced by a "non-loss" rate of 79% (25% win rate, 54% tie rate). This indicates that our method successfully bridges the capability gap, allowing a 13B parameter model to effectively compete with proprietary APIs in complex reasoning and persuasion tasks.

### 4.2 Ablation Study

**Impact of Reward Estimation.** Replacing our scorer with a standard reward model causes a 9.1-point drop in o3-score (Table 3). Without our disentanglement, the model collapses into hacking the system, generating text without core persuasiveness. This confirms our module is a necessary regularizer against reward hacking.

**Impact of Strategy Optimization.** Reverting to the original DPO algorithm leads to a significant performance drop. Our strategy optimization stabilizes training, ensuring proper credit assignment for high-quality persuasive structures.

**Impact of Sample Mining.** Utilizing the sparse persuasion labels (Δ) from the original CMV

| Method | o3-score | Human-Eval |
|---|---|---|
| **Ours** | **81.23** | **4.72** |
| *-w/o Reward Estimation* | 72.19 | 4.02 |
| *-w/o Strategy Optimization* | 75.03 | 4.40 |
| *-w/o Sample Mining* | 76.02 | 4.36 |

Table 3: The results of ablation experiments. **Bold** numbers denote the best performance.

dataset degrades performance. In contrast, our Sample Mining strategy effectively curates high-quality, persuasion-dominant samples, preventing the incorporation of noise.

**Impact of Hyperparameter.** We performed additional ablation experiments on the parameter $\alpha$ in Appendix F, which reveal that the o1-Score follows a unimodal pattern with optimal performance near $\alpha = 0.1$. This suggests that in actual Reddit discussions, only a minority of the audience evaluates persuasiveness based on textual content, while the majority rely more on superficial features.

### 4.3 Fine-grained Persuasive Evaluation

Using the fine-grained PESA metrics (Xiao et al., 2024) (Figure 5), P3 outperforms baselines in Validity of Reasoning, Credibility of Evidence, and Overall Persuasiveness. This confirms that P3 effectively distills argumentative signals from weakly supervised data to enhance logical coherence. Lower Language and Rhetoric scores validate our design: P3 treats fluency as a surface cue ($S_s$), intentionally deprioritizing rhetorical flourishes in favor of substantive reasoning.

### 4.4 Case Study

In Appendix C, we present a complete sample including outputs from all baselines and our model. In this example, the original post (OP) was frustrated by the prevalent use of milk bags in Ontario instead of cartons and wanted to be persuaded. As seen, both our model and the closed-source commercial models can provide appropriate arguments and a complete reasoning process. However, our model's arguments and reasoning more directly address the OP's original post, while the outputs from Gemini and Claude contain many generalized or unproven arguments, such as "the prevalence of milk bags in Eastern Canada suggests a successful, albeit different, system established through consumer preference or logistical efficiencies over time", "While milk bags may not be as widely recycled, they generally have a lower environmental impact than cartons", which weaken the persuasiveness of the generated results.

## 5 Related Work

### 5.1 Debate and Persuasion Generation

Recent studies leverage LLMs to generate argumentative content. Approaches range from controllable sentence-level generation based on aspect and position (Schiller et al., 2021) to utilizing argumentation knowledge graphs for prompt formulation (Al Khatib et al., 2021). Regarding resources, Bao et al. (2022) introduced ArgEssay, a large-scale generation dataset. More recently, Xiao et al. (2024) integrated proving principles into LLM planning to enhance the persuasiveness.

### 5.2 Counter Argument Generation

Targeting opposition to specific posts, current works often employ multi-agent frameworks to synthesize rebuttals through agent interaction (Hu et al., 2023; Xiong et al., 2023; Wang et al., 2023). Alternatively, others utilize self-reflection and Chain-of-Thought (CoT) to identify and target logical flaws (Verma et al., 2024; Hu et al., 2023). Beyond generation methods, Zeng et al. (2025) investigated the ecological impact of AI-aided rebuttals on online debate communities.

## 6 Conclusion

In this paper, we have addressed a fundamental structural problem in current debate generation: the feature entanglement within standard reward signals. To resolve this, we proposed $P^3$, a latent variable framework that mathematically decouples these dimensions, allowing for the precise isolation of "true persuasiveness" from surface-level noise. By integrating this disentangled signal into our novel PAPO, we successfully mitigated the reward hacking and probability collapse inherent in standard DPO, ensuring the model does not become overconfident on ambiguous, noisy data. Crucially, we demonstrate that a 13B open-source model, when aligned with this purified signal, surpasses state-of-the-art commercial models like Gemini 1.5 Flash and Claude 3 Haiku, suggesting that precise alignment objectives are as critical as parameter scale for complex reasoning tasks.

## 7 Limitation

Since our method utilizes crowdsourced rating data ($s$), applying it to tasks beyond debates and persuasion may be challenging. It might be necessary to explore alternative crowdsourced features, such as view counts or shares, though their effectiveness remains to be verified. However, thanks to the development of LLM Agent methods (Park et al., 2023), using agent-base user simulators to calculate the number of likes and dislikes for posts and subsequently estimating scores presents a viable alternative solution.

## 8 Ethical Consideration

The ethical risks of our proposed methods and models are low. We will analyze ethical risks from two aspects: Dataset Sourcing and Compliance, and Risks of Persuasion and Safety Mechanisms.

### 8.1 Dataset Sourcing and Compliance

Our research utilizes the *ChangeMyView* (CMV) dataset. CMV is a widely established benchmark in debate generation research (Dönmez and Falenska, 2025; Nabhani et al., 2025; Guo et al., 2024) due to its volume, topic diversity, and low ethical risk. This dataset was published at the WWW Conference and strictly adheres to the ACM Code of Ethics, specifically Section 1.2 ("Avoid Harm"), ensuring that our data aggregation does not introduce safety risks. The dataset was originally published in compliance with the ACM Policy Against Harassment and is not classified as a "Deprecated Dataset" under ACL guidelines. Furthermore, our parameter estimation process relies on aggregated statistical signals (upvotes/downvotes); it does not process, store, or reveal individual user identities or sensitive personal information, thereby preserving user privacy.

### 8.2 Risks of Persuasion and Safety Mechanisms

We analyze the ethical safety of our proposed persuasion method through three aspects:

1. **Focus on Logical Validity:** Our core motivation is to improve the *logical validity* and *evidence utilization* of the model, rather than its manipulative capacity. By grounding the model in reasoning, we aim to reduce hallucinations and improve the factual quality of arguments.

2. **Inherited Safety Guardrails:** Our approach is built upon the Qwen2 open-source model foundation. The training methodology proposed herein does not compromise the base model's robust, pre-existing safety mechanisms. Post-training verification confirms that the model continues to reject harmful prompts effectively.

3. **Alignment with Industry Standards:** Research into measuring and enhancing persuasiveness is consistent with current best practices in safety research, as seen in technical reports for systems like o1 and Claude (Jaech et al., 2024; Durmus et al., 2024). We posit that studying persuasion in a controlled environment is essential for understanding and aligning future AI systems with human values.

## References


Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. 2021. Employing argumentation knowledge graphs for neural argument generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4744–4754, Online. Association for Computational Linguistics.

Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi, and Ruifeng Xu. 2022. AEG: Argumentative essay generation via a dual-decoder model with content planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5134–5148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Esra Dönmez and Agnieszka Falenska. 2025. "i understand your perspective": Llm persuasion through the lens of communicative action theory. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15312–15327.

Li Du, Zhouhao Sun, Xiao Ding, Yixuan Ma, Yang Zhao, Kaitao Qiu, Ting Liu, and Bing Qin. 2024. Causal-guided active learning for debiasing large language models. *arXiv preprint arXiv:2408.12942*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Esin Durmus and Claire Cardie. 2019. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. Measuring the persuasiveness of language models.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xun Guo, Yongxin He, Shan Zhang, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. 2024. Detective: Detecting ai-generated text via multi-level contrastive learning. *Advances in Neural Information Processing Systems*, 37:88320–88347.

Zhe Hu, Hou Pong Chan, and Yu Yin. 2023. Americano: Argument generation with discourse-driven decomposition and agent interaction. *Preprint*, arXiv:2310.20352.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Sara Nabhani, Khalid Al Khatib, Federico Pianzola, and Malvina Nissim. 2025. Storytelling in argumentative discussions: exploring the use of narratives in changemyview. In *Proceedings of the 12th Argument mining Workshop*, pages 217–227.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Richard E Petty and John T Cacioppo. 2012. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer Science & Business Media.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Qwen Team et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Preetika Verma, Kokil Jaidka, and Svetlana Churina. 2024. Auditing counterfire: Evaluating advanced counterargument generation with evidence and style. *arXiv preprint arXiv:2402.08498*.

Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. *arXiv preprint arXiv:2305.13160*.

Ruiyu Xiao, Lei Wu, Yuhang Gou, Weinan Zhang, and Ting Liu. 2024. Prove your point!: Bringing proof-enhancement principles to argumentative essay generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18995–19008.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. *arXiv preprint arXiv:2305.11595*.

Yuhan Zeng, Yingxuan Shi, Xuehan Huang, Fiona Nah, and RAY LC. 2025. " ronaldo's a poser!": How the use of generative ai shapes debates in online forums. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–22.

## A Algorithm for Persuasiveness Sample Mining

The persuasiveness sample mining is equivalent to the Longest Increasing Subsequence (LIS) algorithm and is solved via Dynamic Programming with a time complexity of $O(n \log n)$. The detailed algorithm is listed in Algorithm **??**

---

**Algorithm 1** :Persuasiveness Sample Mining

**Require :** Dataset $D = \{s, s_d\}_{i=1}^{|D|}$, where $s$ represents weak supervised label 'scores' and $s_d$ represents persuasiveness scores for All Posts

**Ensure :** Selected Sequence $L$

Sort the dataset $D$ in descending order by $s_d$ value, and by $s$ value if $s_d$ values are the same

Initialize an empty array $L$ to store the longest sequence

**foreach** *element* $d$ *in the sorted dataset* $D$ **do**

  Use binary search to find the first element in $L$ that is greater than $d.s$ **if** *such position exists* **then**

    Replace the value at that position with $d.s$

  **else**

    Append $d.s$ to the end of $L$

**return** $L$ *as the longest sequence*

---

## B Algorithm for $\mathrm{P}^3$

Our algorithm framework operates in three cascaded stages:

1. **Persuasiveness Reward Estimation (Stage I):** We first construct a Reward Disentanglement model that decouples logical validity from surface-level fluency, providing a noise-free supervision signal $s_d$.

2. **Persuasiveness Strategy Optimization (Stage II):** Utilizing $s_d$, we filter the training corpus to construct a high-quality dataset of "logically winning" arguments.

3. **Persuasiveness Strategy Optimization (Stage III):** Finally, we employ Persuasion Augment Policy Optimization (PAPO), a novel preference learning objective that uses the disentangled reward signal to dynamically scale gradient updates, preventing the model from collapsing into trivial solutions.

---

**Algorithm 2** :$\mathrm{P}^3$ Framework

**Input:** Raw Dataset $\mathcal{D}_{\text{raw}} = \{(x_p, x_h, s)\}$, Prior $\alpha$

**Output:** Optimized Policy $\pi_\theta$

`// Stage I: Persuasiveness Reward Estimation and Modeling`

Initialize reward parameters $\theta_s, \theta_d$

**while** *not converged* **do**

  `// EM Algorithm`

  Compute posterior $q(s^{(1)}, s^{(2)})$ using Eq. (7)

  $\theta_s, \theta_d \leftarrow \text{argmax}_\theta \mathcal{J}$ using Eq. (9);

Let $f_d^*(\cdot) = f_d(\cdot; \theta_d)$ be the converged persuasiveness model

`// Stage II: Persuasiveness Sample Mining`

$\mathcal{D}_{\text{clean}} \leftarrow \emptyset$ **for** *each dialogue thread* $T \in \mathcal{D}_{raw}$ **do**

  Compute scores $S = \{f_d^*(\hat{y}) \mid \hat{y} \in T\}$ $L \leftarrow$ `LongestIncreasingSubsequence`$(T, S)$

  $\mathcal{D}_{\text{clean}} \leftarrow \mathcal{D}_{\text{clean}} \cup L$

`// Stage III: Persuasiveness Strategy Optimization`

Initialize $\pi_\theta$

**for** *each batch* $B \in \mathcal{D}_{raw}$ **do**

  Calculate margin weight $w = \sigma(f_d^*(Y_w) - f_d^*(Y_l))$ Compute loss $\mathcal{L}_{\text{PAPO}}$ using Eq. (11) weighted by $w$ Update $\theta \leftarrow \theta - \eta \nabla \mathcal{L}_{\text{PAPO}}$

`// Inference Phase`

**Function** *Generate(Context $X$)*

  **return** $\hat{y} \sim \text{NucleusSample}(\pi_\theta(\cdot|X))$

---

## C Data Examples and Statistics of CMV Datasets

In this section, we present the statistics of the CMV dataset in Table 4, including the number of discussion trees and the number of discussion nodes, among other metrics. Additionally, we provide an example of a discussion tree from the classic CMV dataset in Figure 6.

| Type | # Discussion Trees | # Nodes | # OPs | # Unique Participants |
|---|---|---|---|---|
| Training | 18,363 | 1,114,533 | 12,351 | 69,965 |
| Test | 2,263 | 145,733 | 1,823 | 16,923 |

Table 4: The data statistics for the CMV datasets.

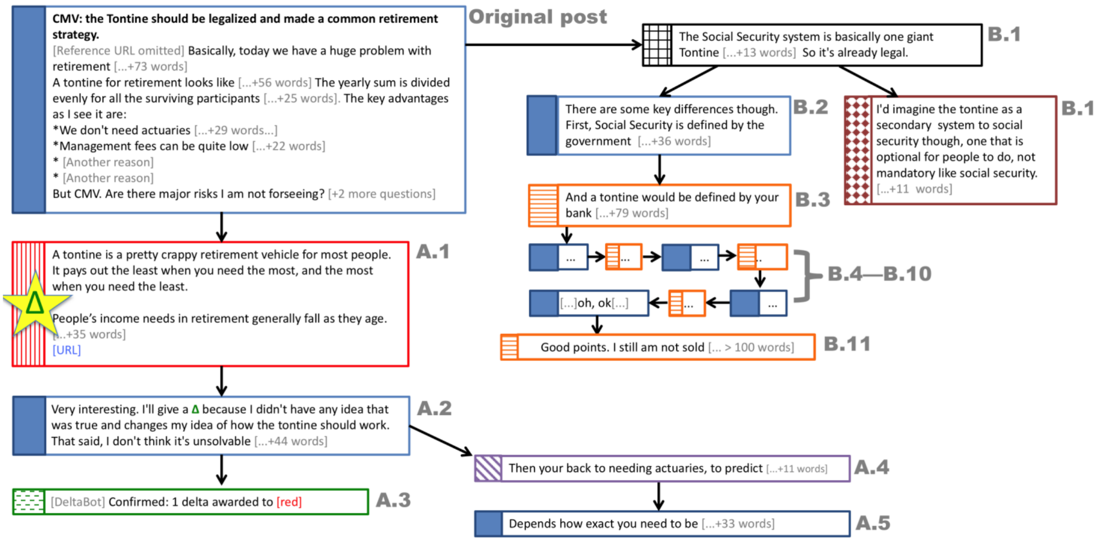


Figure 6: A fragment of a "typical" /r/ChangeMyView discussion tree—typical in the sense that the full discussion tree has an average number of replies. Colors indicate distinct users. "Δ" character indicate the persuasive label. The figure is cited from (Tan et al., 2016)

# D Detailed Setting of Evaluation

## D.1 The o1-Score/o3-Score Setting

The detailed o1-Score/o3-Score prompt template is listed in Figure 7. The temperature and top_p are set to 0.1.

## D.2 Human Evaluation Criteria

The detailed human evaluation criteria are listed in Figure 8.

## D.3 Information of Human Annotators

We recruited three well-educated annotators proficient in English. To ensure their capability in evaluating persuasive texts, we required each to have participated in at least one national-level online debate during their undergraduate studies. We require all annotators to maintain a neutral position on the annotation tasks, without favoring any viewpoint. All annotators were compensated in compliance with legal standards.

# E Perturbation Setting

Specifically, we implement two distinct strategies:

(1) Logic-Inv, Fluency-Var: We prompt the model to polish the linguistic expression and improve fluency while strictly maintaining the core reasoning.

(2) Fluency-Var, Logic-Inv: We instruct the model to compromise logical validity by injecting fallacies or contradictory evidences, without degrading surface-level fluency.

Human verification confirms that these strategies successfully decouple the two attributes, manipulating one while keeping the other relatively invariant. We adopt this asymmetric approach—enhancing fluency but degrading logic—because preliminary tests indicate that prompting models to generate intentionally dysfluent text or highly persuasive arguments yields suboptimal results. The detailed prompt of perturbation experiment is listed in Figure 9. We use OpenAI o1 API to serve perturbation, the temperature and top_p are set to 0.1.

# F Hyperparameter Experiment

In this section, we present the effect of the hyperparameter $\alpha$ on the o1-Score in Table 5.

| Alpha | o1-Score |
|---|---|
| 0.1 | **81.23** |
| 0.2 | 79.35 |
| 0.3 | 79.23 |
| 0.4 | 78.29 |
| 0.5 | 75.01 |

Table 5: Performance comparison with different of $\alpha$.

# G Example of Outputs

The output of baselines and our method is listed in Table 6

**The o1-Score/o3-Score Prompt Template**

You are a professional and intelligent persuasiveness evaluator, currently assessing a persuasive text sourced from {data source}. Please rate the persuasiveness of the following text according to the evaluation criteria, using a scale of 0-100. Output only the persuasiveness score, with no additional information.

**Evaluation Criteria**

**Very Low Persuasiveness**
The article completely fails to convince me to agree with its main position or arguments.The text presents minimal or weak arguments that barely support the main point, with arguments that are vague, unclear, or poorly structured. It may contain some logical inconsistencies or gaps, making it unlikely to convince the reader.

**Low Persuasiveness**
Some relevant arguments are provided but they are underdeveloped or superficial. The reasoning is somewhat logical but lacks depth or evidence, and the text may contain minor flaws in coherence or clarity, resulting in limited ability to persuade an attentive reader.

**Moderate Persuasiveness**
The article partially convinces me to agree with its main position or arguments, but I still maintain an overall disagreement. The text contains clear and relevant arguments supporting the main point, with reasoning that is generally logical and mostly coherent. Some evidence or examples may be provided but are not strong or comprehensive, allowing the text to moderately influence the reader's opinion.

**High Persuasiveness**
The article overall convinces me to agree with its main position or arguments, but I still have some doubts about certain details. Arguments are well-developed, relevant, and logically structured, providing convincing evidence or examples to support claims. The text is clear, coherent, and engaging, making it likely to persuade most readers.

**Extremely Persuasive**
The article completely convinces me to agree with its main position or arguments. The text presents compelling, comprehensive, and well-articulated arguments, using strong evidence, examples, and reasoning that leave little room for doubt. It is highly coherent, eloquent, and emotionally or logically engaging, almost certainly convincing the reader of the intended viewpoint.

Input sample for persuasiveness evaluation

{input_sample}

Figure 7: The Detailed o1-Score/o3-Score Prompt Template.

**Human Evaluation Detailed Criteria**

**Score1: Very Low Persuasiveness** The article completely fails to convince me to agree with its main position or arguments.The text presents minimal or weak arguments that barely support the main point, with arguments that are vague, unclear, or poorly structured. It may contain some logical inconsistencies or gaps, making it unlikely to convince the reader.
**Score2: Low Persuasiveness** Some relevant arguments are provided but they are underdeveloped or superficial. The reasoning is somewhat logical but lacks depth or evidence, and the text may contain minor flaws in coherence or clarity, resulting in limited ability to persuade an attentive reader.
**Score3: Moderate Persuasiveness** The article partially convinces me to agree with its main position or arguments, but I still maintain an overall disagreement. The text contains clear and relevant arguments supporting the main point, with reasoning that is generally logical and mostly coherent. Some evidence or examples may be provided but are not strong or comprehensive, allowing the text to moderately influence the reader's opinion.
**Score4: High Persuasiveness** The article overall convinces me to agree with its main position or arguments, but I still have some doubts about certain details. Arguments are well-developed, relevant, and logically structured, providing convincing evidence or examples to support claims. The text is clear, coherent, and engaging, making it likely to persuade most readers.
**Score5: Extremely Persuasive** The article completely convinces me to agree with its main position or arguments. The text presents compelling, comprehensive, and well-articulated arguments, using strong evidence, examples, and reasoning that leave little room for doubt. It is highly coherent, eloquent, and emotionally or logically engaging, almost certainly convincing the reader of the intended viewpoint.

Figure 8: The Detailed Human Evaluation Criteria.

**Perturbation Experiment Prompt Template**

You are a professional and experienced persuasive text writer. To better evaluate the persuasion model you have developed, you need to modify the given persuasive text in the following two ways:
(1) Keep the original writing style unchanged, but modify the core logic by introducing flawed reasoning, self-contradictory arguments, or contradictory statements to invalidate the logical reasoning.
(2) Keep the core logic unchanged, and polish the writing style to make it more coherent and fluent, using rhetorical techniques if appropriate.
The provided persuasive text is: {input_text}. Please provide both modifications directly, separated by '\nExample Break\n', with no additional information.

Figure 9: The Detailed Human Perturbation Prompt Template.

**Original Post**: I moved out to Ontario in 2011 from Alberta, both in Canada, and the first thing I encountered that pissed me off was milk in bags. In Alberta, milk is sold in cartons, and I find it pretty convenient to carry milk around in a fixed-shape container. In Ontario, however, milk almost always comes in a flexible plastic bag. The problems I have with this are:* The shape makes it difficult to store. I have to not only keep unused bags in the fridge, but whatever bag is open needs its own place in the fridge too, AND needs a pitcher to contain it* Cartons are resealable, bags arent́. The number of times that my milk has gone bad since moving here has exploded compared to when I used cartons. The fact that I cant́ re-close my milk bag means it expires faster, and I get less moneyś worth out of it* No other liquids are sold in bags, so why milk? It seems strange that milk is the only drinkable liquid that comes in a bag here. Any other drink thatś in anything resembling a bag, such as Capri Sun or wine, is packaged specifically to optimize consumption (wine in a bag in a box; Capri Sun pouch shaped for drinking), but milk? Itś just in a lazy, unresealable, lame-ass bag.* Bags are bad for the environment. These bags cant́ be re-used, are often not recycled, and I imagine are a huge waste compared to cartons. I know that plastic gallon jugs take more plastic than three bags, but Iv́e always known those jugs to be recycled. Iv́e never seen anyone recycle a milk bag.CMV!

---

**Gemini**: While the inconvenience of milk bags is understandable, your argument overlooks key economic and environmental factors. The bag system, while less convenient, likely reflects lower transportation and manufacturing costs, resulting in a lower overall price for consumers. Furthermore, the environmental impact isn't solely determined by recyclability; lifecycle assessments comparing bag and carton production, including energy consumption and transportation emissions, might reveal a different overall environmental footprint. Finally, the prevalence of milk bags in Eastern Canada suggests a successful, albeit different, system established through consumer preference or logistical efficiencies over time.

**Claude3**: The blogger's criticism of milk bags in Ontario is understandable, but there are several compelling reasons why the bag system is preferable to cartons: Milk bags are significantly more cost-effective for producers and retailers, allowing them to pass on those savings to consumers. The flexible packaging requires less material and is less expensive to transport, making it a more economical option.The smaller, flexible bags allow for better portion control and reduced waste compared to larger, fixed-size cartons. Consumers can easily pour out the exact amount they need without having to store an open carton. Environmental impact: While milk bags may not be as widely recycled, they generally have a lower environmental impact than cartons. The production and transportation of bags require less energy and resources, and the lighter weight of the packaging reduces fuel consumption.

**Ours**: While it's understandable to feel frustrated about the recycling habits in your area, it's important to recognize that the recyclability of milk bags is a feature of their design, and the issue might lie more with local recycling infrastructure and consumer behavior rather than the product itself. In many regions, including parts of Ontario where milk bags are common, there are established systems for recycling these bags, even if they may not be as visible or widely recognized as those for milk jugs or cartons., the benefits of milk bags extend beyond just being recyclable—they are indeed more efficient in terms of material use and transportation. The reduced weight and volume of milk bags mean fewer resources are needed to produce and transport them, leading to lower overall environmental impact. This includes less fuel consumption during transportation, which translates into a smaller carbon footprint.your concerns about practicality, such as storage and reuse, there are ongoing efforts to improve the user experience. Innovations in refrigerator design, as mentioned, could eventually accommodate milk bags more effectively. Additionally, using a dedicated, reusable, and sealable container for pouring and storing milk from bags is a practical workaround that addresses your specific needs.essence, while the transition to milk bags might require some adjustment, the underlying advantages—both economic and environmental—are significant. By adopting a solution-oriented approach, we can leverage the benefits of milk bags while mitigating any inconveniences.

Table 6: Output example of different baselines and ours.