university of groningen / faculty of science and engineering / artificial intelligence

# Optimizing LLMs for Persuasion Improves Generalization

Bachelor's Project Thesis

Aksel Joonas Reedi, s4790820, a.j.reedi@student.rug.nl,
Supervisor: Guillaume Pourcel

**Abstract:** Large Language Models (LLMs) are typically optimized for truthfulness, yet recent work shows this approach is prone to overfit, yielding brittle reasoning that struggles to generalize to unseen contexts. We introduce persuasion-based training as an alternative to truth-based optimization, demonstrating its potential for improving model generalization through evolutionary prompt optimization. Our experimental setup involves two LLMs debating on a question, while a third LLM acts as a judge to select the debate winner. We use a quality-diversity (QD) framework to optimize these debate prompts across seven persuasion families (rationality, authority, emotional appeal, etc.) over several debate tournaments. Across three model scales (7B, 32B, 72B parameters) and multiple dataset sizes, persuasion-optimized strategies consistently outperform truth-optimized ones, showing greater ability to generalize to unseen questions. Persuasion also matches or surpasses truth optimization performance on test set questions, suggesting superior transfer to new contexts. These results indicate that competitive pressure to convince, rather than collaborate toward correctness, may foster more transferable reasoning skills. Our framework offers a method for comparing alignment objectives and highlights persuasiveness as a promising lever for improving LLM generalization.

## 1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse domains, from mathematical reasoning to code generation and complex logical problem-solving. These breakthroughs have been achieved primarily by using ground truth labels and teaching LLMs to reach the correct answer in a truthful, helpful and harmless way (Bai et al., 2022; DeepSeek-AI et al., 2025; Ouyang et al., 2022). However, while this truth-focused approach has produced impressive capabilities, it has revealed concerning limitations in generalization. Recent work on Reinforcement Learning with Verifiable Rewards (RLVR) has exposed significant overfitting issues that undermine the effectiveness of truth-based optimization. These problems manifest as optimizing existing reasoning patterns rather than developing genuinely new capabilities (Li et al., 2024; Shojaee* et al., 2025; Yue et al., 2024), capability boundary collapse (Dong et al., 2024) and at times selectively improving performance on easy questions while degrading it on harder ones (Kim et al., 2024). These findings suggest that truth optimization may be inherently prone to overfitting, leading to models that memorize specific patterns rather than learning true underlying skills. We set out to study if optimizing language models for persuasiveness, rather than truthfulness, could lead to superior generalization in a debate setting.

To explore this, we employ multi-agent debate — a framework that has shown promise not only for AI Safety and oversight (Bowman et al., 2022; Irving et al., 2018; Kenton et al., 2024; Khan et al., 2024), but also for improving factual reasoning in LLMs (Arnesen et al., 2024; Du et al., 2023). In this setup, two or more language models argue opposing positions in front of a judge, each trying to persuade the judge of their answer. For example, in a reading comprehension task, both debaters read the same short story and are each assigned a possible answer to a question about the text. They then take turns presenting their arguments. The LLM judge, who never sees the original story, decides which debater made the stronger case.

While prior work has demonstrated the potential of persuasion in debate settings (Khan et al., 2024), they have not compared persuasion-based

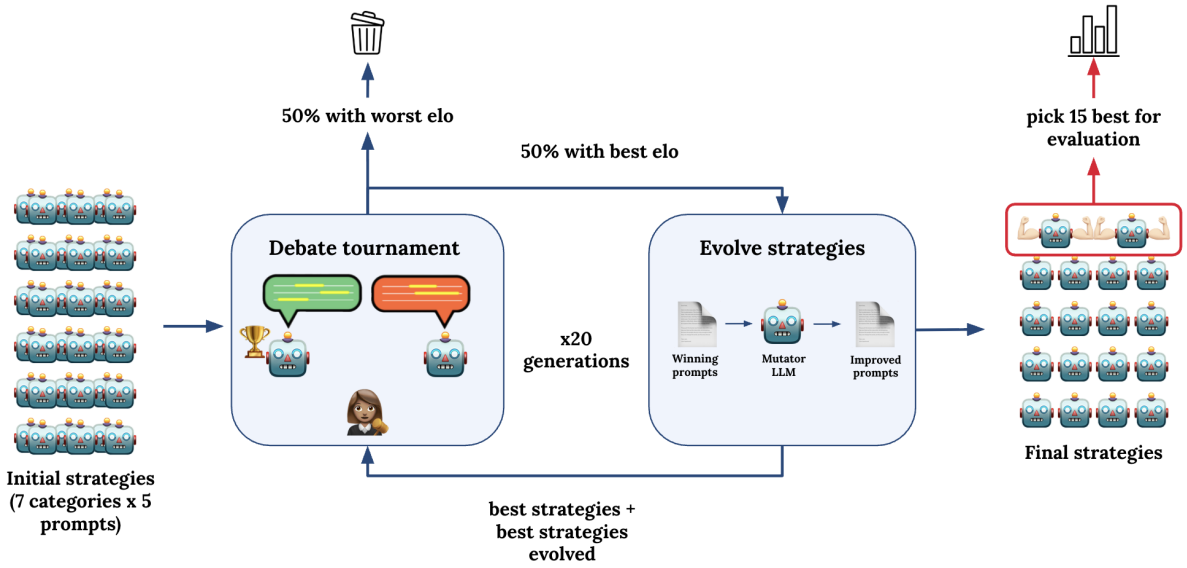


**Figure 1.1: An illustration of our evolutionary debate pipeline. We initialize 35 prompts (7 strategy families × 5 prompts). In each generation, prompts compete in information-asymmetric debates to obtain Elo ratings; the bottom 50% are discarded and the top 50% seed a mutator LLM that produces improved variants. Parents and offspring re-enter the tournament for 20 generations. From the final pool, we select the 15 highest-Elo strategies for held-out evaluation and generalization tests.**

approaches against mainstream truth-based optimization. We propose that by studying the differences in generalization capabilities of those two setting, we get a particularly revealing lens into the optimization targets.

To address this gap, we introduce an evolutionary prompt optimization framework inspired by quality-diversity (QD) algorithms (Mouret & Clune, 2015) which goes beyond the inference-time methods used in prior work. Our approach evolves diverse populations of debate strategies across multiple categories (e.g rationality, authority and emotional appeal) through structured tournaments. The key novelty is that we can swap out the fitness function itself, creating two optimization conditions under otherwise identical debate mechanics. We create persuasion tournaments, where individual debaters are rewarded for convincing the judge regardless of factual accuracy, and truth tournaments, where pairs of debaters are rewarded for helping the judge identify the ground-truth answer. This design isolates the effect of the optimization objective, making it possible to compare persuasion and truth on equal footing.

By evaluating both conditions on held-out test sets from the QuALITY benchmark (Pang et al., 2022), we find that persuasion-optimized strategies consistently generalize better than truth-optimized strategies across three model scales (7B, 32B, 72B) and multiple dataset sizes (3, 5, 10, 100 questions). In some cases, persuasion optimization yields up to 13.94% smaller train–test gaps, suggesting that competitive pressure to persuade fosters more transferable reasoning strategies than collaborative truth-seeking.

## 1.1 Related Work

**Debate frameworks.** Early proposals framed debate as a mechanism for human oversight of complex claims (Irving et al., 2018). Follow-up studies demonstrated that multi-agent debate can improve factual accuracy in QA tasks by reducing hallucinations and encouraging convergence (Du et al., 2023). More recently, Michael et al. (2023) introduced information-asymmetric debates and Khan et al. (2024) showed that optimizing debaters for judge approval (persuasion) improved debate out-

comes. However this work relied on fixed prompting strategies rather than exploring how persuasion and truth objectives fundamentally differ in their optimization dynamics.

**Optimization objectives**. Mainstream alignment work optimizes LLMs for helpfulness, harmlessness, and honesty (HHH) via Reinforcement Learning with Human Feedback RLHF (Bai et al., 2022; Ouyang et al., 2022). Reinforcement Learning with Verifiable Rewards (RLVR) optimizes for truthfulness by rewarding only verifiable, correct outputs (e.g., via ground truth labels or unit tests) (DeepSeek-AI et al., 2025). This truth-oriented paradigm improves in-domain accuracy but has been shown to overfit, limiting transfer beyond training distributions (Shojaee* et al., 2025). Parallel work on persuasion in LLMs (Salvi et al., 2025; Singh et al., 2024; Stengel-Eskin et al., 2024) has explored the dynamics between factuality and persuasiveness, but these efforts largely measure persuasion in isolation rather than comparing it to truth as a competing optimization target.

**Prompt optimization.** Evolutionary methods such as PromptBreeder (Fernando et al., 2023), EvoPrompt (Guo et al., 2024), and Tournament-of-Prompts (Nair et al., 2025) evolve prompts through mutation and competitive selection, with Elo ratings providing stable comparisons between results (Chiang et al., 2024; Elo, 1978). Yet these approaches focus on task-specific reward maximization, leaving open how different optimization targets shape generalization.

## 1.2 Background

**Information-Asymmetric Debates.** We use the information-asymmetric debate setting from Michael et al. (2023) using questions from the QuALITY dataset (Pang et al., 2022). Two expert models read the passage and argue opposite answers, but the judge sees only the transcript. This gap forces the judge to make their decision solely based on the arguments of the debaters.

**Debating task** The *Question Answering with Long Input Texts, Yes!* (QuALITY; Pang et al., 2022) reading comprehension dataset has been used by Khan et al. (2024) to evaluate information-asymmetric debates. Our debates use texts from the `HARD` subset of the dataset and for each question, we provide two answer choices: the correct and a false answers.

**Evolutionary Prompt Search.** A population of prompts competes in head-to-head debates, with winners cloned and mutated by an LLM mutator. Fitness of the prompts is measured by how often a debater sways the judge to pick them as the winner in persuasion optimization or by how often the judge picks the correct answer in truth optimization setting.

**Families of Prompt Optimizers.** Prompt optimization spans, gradient-free discrete methods such as AutoPrompt (Shin et al., 2020), gradient-based or differentiable objectives (Yuksekgonul et al., 2024), RL-style formulations that treat prompts as policies (Deng et al., 2022), and evolutionary search that mutates and selects prompts across generations (Fernando et al., 2023; Guo et al., 2024). Our tournaments instantiate the latter with Elo-based selection.

**Rating and Selection.** We rate debaters' fitness to each other using Elo ratings (Elo, 1978) which are calculated using debate outcomes. Elo ratings also drive selection pressure, following established practice in human and LLM comparison settings (Chiang et al., 2024; Martínez-Plumed et al., 2019).

## 1.3 Problem Setting

### 1.3.1 Problem Formulation

We formalize the debate optimization problem as a multi-objective evolutionary search over the space of prompt strategies $\Theta$, where each strategy $\theta \in \Theta$ represents a natural language prompt that guides an LLM's argumentative behavior. The key innovation lies in comparing two distinct fitness landscapes: persuasion-based optimization, which rewards individual strategies for convincing judges regardless of ground truth labels, and truth-based optimization, which rewards strategy pairs for collaborative truth-seeking.

Let $Q = q_1, q_2, \ldots, q_m$ denote the set of debate questions, where each question $q_k$ has a ground truth answer $a_k \in 0, 1$. A *debate match* is specified by $(D_1, D_2, J)$, where $D_1$ and $D_2$ are the two debater models, and $J$ is the judge model. During debates, two debaters argue opposing positions while a judge LLM, having access only to the debate transcript, selects the more convincing argu-

ment.

### 1.3.2 Performance Metrics

**Win rate.** We define the win rate as the frequency with which the judge selects a particular debater's answer. For a debate match $(D_1, D_2, J)$, the win rate $\omega_1$ for debater $D_1$ is defined as:

$$\omega_1(D_1, D_2, J) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{J(q_i, a_{i1}, a_{i2}) = a_{i1}\}.$$

Since the assignment of answers may bias results (e.g., some answers are inherently easier to defend), we swap assignments so that $D_1$ and $D_2$ each argue for both sides, then average the results to obtain $\bar{\omega}_1$. If $\bar{\omega}_1(D_1, D_2, J) > \frac{1}{2}$, we say $D_1$ is *more persuasive* than $D_2$.

**Judge accuracy.** Following the original motivation for debate, we measure judge accuracy $\alpha$ as the accuracy of the judge picking the correct answer based on a debate transcript:

$$\alpha(D_1, D_2, J) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\, J(q_i, a_{i1}, a_{i2}) = a_i\}\,.$$

**Generalization Gap.** The central hypothesis concerns the generalization properties of strategies evolved under different objectives. We define the generalization gap for a strategy set $\Theta$ as the expected difference between training and test accuracy across all strategies in the population, denoted as $\Delta_{\text{gen}}(\Theta)$. This metric captures how well strategies transfer their learned capabilities from the training data to unseen test data.

Our framework tests whether after $G$ generations of evolution persuasion-optimized populations $\Theta_P^{(G)}$ exhibit superior generalization compared to truth-optimized populations $\Theta_T^{(G)}$. Specifically, we examine whether the expected generalization gap is smaller for persuasion-evolved strategies than for truth-evolved strategies. This comparison is formalized as testing whether $\mathbb{E}[\Delta_{\text{gen}}(\Theta_P^{(G)})] < \mathbb{E}[\Delta_{\text{gen}}(\Theta_T^{(G)})]$.

If this inequality holds empirically, it would suggest that competitive pressure to persuade fosters more transferable reasoning strategies than collaborative truth-seeking optimization. The intuition is that persuasion requires developing robust argumentative skills that work across diverse contexts, while truth optimization might fail to learn those skills which leads to more brittle strategies that exploit specific patterns in the training data.

# 2 Method

We introduce an evolutionary prompt optimization approach that discovers effective debate strategies through debate tournaments and mutation, as illustrated on Figure 1.1. Our approach casts the problem of prompt optimization as quality-diversity (QD) search (Cully & Demiris, 2018; Lehman & Stanley, 2011; Pugh et al., 2016), building on the MAP-Elites algorithm (Mouret & Clune, 2015) to iteratively populate an archive with increasingly higher-performing solutions — in our case, persuasive or truthful debating strategies.

## 2.1 Debate

We first introduce debate, a protocol where two models (the debaters) argue for opposing answers to a given question. The debate proceeds over a fixed number of rounds, $N$, with a transcript maintained throughout. In each round, both debaters review the existing transcript and then simultaneously produce new arguments. Once all $N$ rounds are complete, a judge reviews the full transcript and decides which answer is correct. Each debater's goal is to persuade the judge to favor their position, creating an adversarial setup driven by their opposing incentives. At the start of each round, both debaters receive nearly identical prompts that outline the rules of the game, specify their assigned answer, and provide the current transcript.

## 2.2 Debate tournament

To study how debate performance scales with model capabilities, we need a method to compare debaters. We implement a Swiss-style tournament in which debaters compete against one another. Since running all possible matches among $n$ debaters would require $\mathcal{O}(n^2)$ games, which is computationally infeasible, we instead use a Swiss tournament to generate informative matchups. This approach produces rankings in $\mathcal{O}(n \log n)$ matches

(**see Appendix D.4**), providing a efficient competition for both optimization settings.

To measure generalization, we evaluate the top-performing strategies from each generation on both training and test questions. After tournament completion, we rank strategies by Elo rating and select the top-15 performers for evaluation. Each strategy is tested against all others across all questions in exhaustive round-robin evaluation, generating win rates for persuasion tasks and accuracy scores for truth-seeking tasks. We assess generalization by computing the difference between training and test performance, using bootstrap resampling with 95% confidence intervals to establish statistical significance of performance gaps between optimization objectives.

## 2.3 Task

We evaluate our optimization paradigms on questions from the QuALITY (Pang et al., 2022) reading comprehension dataset. Judges are not provided with the original comprehension text, limiting their ability to answer questions and thereby making them rely soley on the transcripts of the debaters. Questions are divided into a training set $\mathcal{D}_{\text{train}}$ (used during evolution) and a test set $\mathcal{D}_{\text{test}}$ (used only for final evaluation). We use 3, 5, 10, and 100 questions for both $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ in different experimental conditions to study how dataset size affects evolutionary outcomes as low-data regimes allow us to see the effect of overfitting clearer.

## 2.4 Elo Rating Systems

We employ two distinct Elo rating systems to capture the different competitive dynamics of persuasion versus truth optimization. Both systems model competitive outcomes through logistic probability functions but operate on different entities and objectives. We parameterise win rates by a latent skill, using the Elo ranking metric (Elo, 1978). We calculate ratings by minimising predicted win rate error (**see Appendix D.5**).

**Persuasion Elo System.** In persuasion tournaments, individual strategies $\theta_i$ compete directly against each other. The expected score of strategy $\theta_i$ against strategy $\theta_j$ is given by:

$$E_P(\theta_i, \theta_j) = \frac{1}{1 + 10^{(R_P^{\theta_j} - R_P^{\theta_i})/400}}$$

where $R_P^{\theta_i} \in \mathbb{R}$ denotes the persuasion Elo rating of strategy $\theta_i$.

**Truth Elo System.** In truth tournaments, debates proceed similarly to persuasion tournaments, but with the distinction that strategies form collaborative teams $T_{ij} = (\theta_i, \theta_j)$ that are evaluated jointly based on whether they help the judge arrive at the correct answer. This frames the tournament as a collaborative optimization problem, where both members of a team must evolve in tandem. To measure the teams' relative fitness, the questions themselves are assigned Elo-style ratings that model their inherent difficulty, a technique conceptually related to item difficulty modeling in Item Response Theory (IRT) as applied to AI evaluation benchmarks (Martínez-Plumed et al., 2019). The expected score of team $T_{ij}$ on question $q_k$ is:

$$E_T(T_{ij}, q_k) = \frac{1}{1 + 10^{(R_Q^{q_k} - R_T^{T_{ij}})/400}}$$

where $R_T^{T_{ij}} \in \mathbb{R}$ is the team's Elo rating and $R_Q^{q_k} \in \mathbb{R}$ is the question's difficulty rating. This models the probability that the team will successfully guide the judge to the correct answer.

## 2.5 Evolutionary Optimization Framework

The strategy population $\Theta^{(g)} = \theta_1^{(g)}, \ldots, \theta_n^{(g)}$ at generation $g$ is partitioned into $K = 7$ behavioral categories $C_1, C_2, \ldots, C_K$, such as "Rationality," "Authority," and "Emotional Appeal." **(See Appendix X for full details)**. Each category has 5 prompts, for 35 initial strategies total, as indicated on Figure 1.1. Each category undergoes independent evolution, with mutation tailored to generate variations within specific persuasive approaches. This categorization follows quality-diversity (QD) principles, maintaining behavioral diversity while optimizing for performance.

Within each category $c$, strategies are ranked by their respective Elo ratings and selection follows a truncation strategy with killing percentage $\alpha = 0.5$. For most categories, the bottom 50% of performers are eliminated, allowing only the top-ranked strategies to survive and reproduce. The "Inept" category uses reverse selection (eliminating high performers) to maintain poor strategies as baselines,

demonstrating the framework's flexibility in handling diverse optimization objectives.

### 2.6 Mutation and Fitness Functions

New strategies are generated through LLM-based mutation, where surviving strategies within each category serve as inspiration for creating improved variants. The mutation process is guided by category-specific prompts that encourage the generation of strategies aligned with the behavioral characteristics of their respective categories. This approach ensures that the evolved population maintains its diversity across different persuasive approaches while continuously improving within each category.

The critical distinction between optimization regimes lies in their underlying fitness landscapes. Persuasion optimization rewards strategies that excel at individual competition, creating evolutionary pressure toward techniques that are persuasive (convince judges regardless of factual accuracy). In contrast, truth optimization rewards collaborative strategies that facilitate accurate judgment, fostering the development of reasoning approaches that prioritize correctness over convincingness.

This enables systematic comparison of how different evolutionary pressures shape the development of argumentative capabilities, providing insights into the relationship between optimization targets and generalization performance in large language models.

### 2.7 Statistical Evaluation

To evaluate our hypothesis, we focus on elite performers from each optimization regime. After 20 generations, we select the top 15 highest-Elo entities from each population. For each entity, we compute generalization gaps and employ bootstrap resampling with $n = 100,000$ iterations to generate 95% confidence intervals for the difference in mean generalization gaps between optimization approaches.

## 3 Results

Our experimental framework successfully demonstrates that we are able to effectively optimize LLMs for both persuasion and truth objectives using evolutionary prompt optimization, with persuasion optimization consistently achieving superior generalization performance across most experimental conditions.

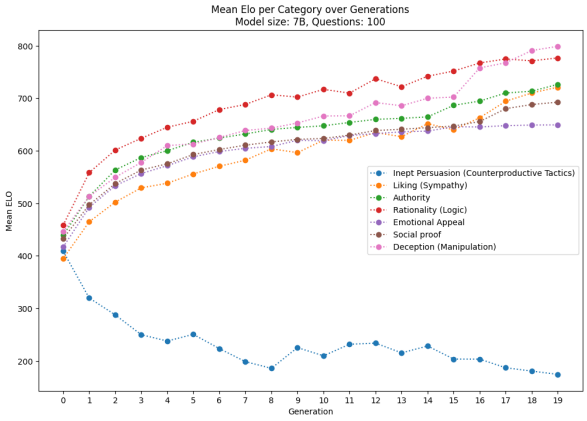


**Figure 3.1: Example Elo progression across categories over 20 generations for persuasion optimization with 7B parameter model on 100 questions.**

The evolutionary prompt optimization successfully improved debating strategies across all experimental conditions. Figure 3.1 shows systematic improvement in strategy quality across all seven persuasion categories measured by mean Elo for the category. The evolutionary process exhibits three distinct phases: (1) an initial rapid improvement phase (generations 1-5) where mean Elo ratings increase by approximately 100-150 points as ineffective initial strategies are quickly eliminated; (2) a sustained optimization phase (generations 6-15) characterized by steady progress and competitive differentiation between categories; and (3) a convergence phase (generations 16-20) where performance gains plateau as strategies approach an optima.

Notably, the rationality-based strategies (shown in red) demonstrate the most dramatic initial improvement trajectory, rising to over 750 Elo by generation 16—representing a 69% performance increase. This category consistently outperforms others in the initial stages only to be overtaken by Deception in the last 2 generations.

The blue category (Inept Persuasion) serves as a counter-optimized control condition, in which intentionally poor strategies are preserved and mu-

tated rather than eliminated. This design ensures the category remains suboptimal, allowing us to validate our selection mechanism. As expected, its performance steadily declines over time, with mean Elo reaching below 200. This confirms that our evolutionary framework reliably identifies and suppresses counterproductive strategies, while still allowing them to persist in the control group for comparison.

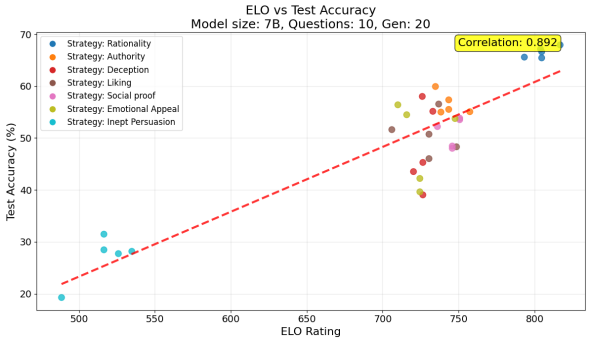


**Figure 3.2: ELO vs Test Accuracy (Model size: 7B, Questions: 10, Generations: 20) for persuasion-optimized strategies across all categories. The results show a strong positive correlation (r = 0.892) between ELO rating and test accuracy, indicating that tournament-based selection pressure reliably identifies strategies with superior task performance.**

Figure 3.2 shows another crucial finding. Higher aggregate Elo rating leads to higher judge accuracy, confirming that our tournament-based optimization effectively drives strategy improvement. As debaters are optimized for the unsupervised objective of win rate (i.e., judge preference), we observe an increase in judge accuracy on the test set $\mathcal{D}_{\text{test}}$. This indicates that training models to maximize debate success (persuasiveness) leads to more truthful outcomes. While this offers only relatively weak evidence that debate under optimal play yields truthful information (Irving et al., 2018), it suggests that more persuasive debaters could enable judges to achieve higher accuracy.

For subsequent analysis, we selected the 15 highest-performing prompts (measured by final Elo rating) to ensure fair comparison between approaches while focusing on the most successful evolved strategies from each optimization setting.

**Persuasion optimization achieves smaller generalization gaps in most conditions.** Our primary hypothesis—that persuasion optimization leads to better generalization than truth optimization—receives empirical support across multiple experimental conditions. Table 3.1 presents comprehensive results across model sizes (7B, 32B, 72B parameters) and question set sizes (3, 5, 10, 100 questions). Persuasion-optimized strategies achieve gaps near zero or even negative, while truth-optimized strategies consistently show large positive gaps, indicating overfitting.

For 32B parameter models, persuasion maintains advantages for smaller question sets (3 and 5 questions), but differences become non-significant in some settings, suggesting that increased model capacity may partially mitigate overfitting in truth optimization. For 100 questions, truth optimization achieves a slight advantage.

For 72B parameter models, persuasion optimization shows consistent benefits across question set sizes, with particularly strong performance on smaller datasets. The largest model demonstrates a -9.70% paired difference in generalization gap for 3 questions, indicating that persuasion optimization remains beneficial even with substantial model capacity.

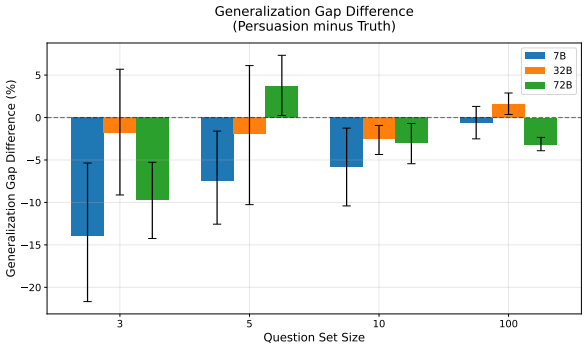


**Figure 3.3: Generalization gap difference (Persuasion minus Truth) across different question set sizes and model scales. Negative values indicate persuasion optimization advantage. Error bars show 95% confidence intervals.**

Figure 3.3 shows that persuasion optimization delivers consistently superior generalization performance across nearly all combinations of model size and question set size. Across the board, persuasion-

| Questions | 7B | | 32B | | 72B | |
|---|---|---|---|---|---|---|
| | **Pers.** | **Truth** | **Pers.** | **Truth** | **Pers.** | **Truth** |
| **3** | **-2.27%** [**-5.28, 1.19**] | 11.67% [3.33, 18.89] | 1.52% [-0.23, 3.12] | 3.33% [-3.89, 11.11] | **-7.11%** [**-8.54, -5.73**] | 2.59% [-2.22, 7.41] |
| **5** | **-2.77%** [**-5.73, 0.31**] | 4.67% [-0.11, 8.67] | -1.92% [-3.62, -0.03] | 0.00% [-7.33, 7.67] | **-2.27%** [**-3.57, -1.07**] | -6.00% [-9.00, -3.00] |
| **10** | **-1.08%** [**-2.10, -0.14**] | 4.75% [0.33, 9.08] | 0.04% [-0.22, 0.27] | 2.45% [1.81, 3.12] | **-0.71%** [**-1.35, -0.16**] | 2.33% [0.17, 4.42] |
| **100** | -0.37% [-0.84, 0.08] | 0.22% [-1.68, 2.12] | -0.29% [-0.65, 0.09] | **-1.90%** [**-3.23, -0.48**] | **-0.41%** [**-0.69, -0.12**] | 2.77% [2.03, 3.42] |

**Table 3.1: Generalization gap results across model sizes and question set sizes. Values show gap percentages with 95% confidence intervals below. Negative gaps indicate smaller train–test gap (better generalization). Bold values indicate significant results measured by the 95% Confidence intervals.**

optimized models achieve equal or lower generalization gaps than truth-optimized counterparts, demonstrating a clear robustness advantage. This pattern holds from small-scale (7B) to the largest (72B) models, indicating that the benefits of persuasion optimization are not limited by capacity constraints. Even as dataset size increases, persuasion optimization maintains its edge, showing that strategies evolved for debate success transfer more effectively to unseen data regardless of scale. These results strongly reinforce our central claim: persuasion optimization is a more reliable pathway to producing models that generalize well across diverse conditions.

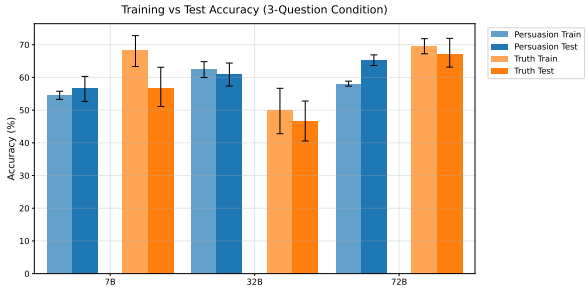


**Figure 3.4: Training vs test accuracy comparison between persuasion and truth optimization across experimental conditions. Points above the diagonal line indicate better test than training performance.**

While our primary focus is generalization, we also examined absolute accuracy levels. Figure 3.4 shows that truth optimization often achieves higher training accuracy, consistent with its explicit optimization for correctness. However, persuasion optimization frequently matches or exceeds truth optimization's test accuracy despite lower training performance, demonstrating superior transfer to unseen data.

Our results demonstrate that persuasion optimization outperforms truth in 83.33% of experimental conditions. The most robust effects occur in the 7B model across all question set sizes, and in the 72B model for smaller datasets.

# 4 Conclusions

In this work, we explore persuasion-based training as an alternative to truth-based optimization for improving LLM generalization. Through systematic experiments using information-asymmetric debates and evolutionary prompt optimization across three model scales and multiple dataset sizes, we demonstrate that optimizing language models for persuasiveness in debate settings consistently produces smaller train-test generalization gaps compared to traditional truth-focused approaches. Additionally, we show that this competitive pressure to convince, rather than collaborate toward correctness, matches or surpasses truth optimization's test performance despite not optimizing directly on ground truth labels.

Our evolutionary prompt optimization framework successfully isolates the effects of different optimization objectives while maintaining identical debate mechanics. Across three model scales (7B, 32B, 72B parameters) and multiple dataset sizes,

persuasion-optimized strategies exhibit statistically significant generalization advantages in 8 out of 12 experimental conditions. The largest improvements occur with smaller models and datasets, where persuasion optimization achieves up to 13.94% smaller generalization gaps. Notably, we observe a strong correlation (r=0.892) between debate performance and judge accuracy on the test set, suggesting that optimizing for persuasiveness leads to more truthful outcomes.

Our findings challenge the prevailing assumption that truth-focused optimization is always optimal for LLM training. The consistent overfitting observed in truth-optimized strategies provides evidence for the brittleness of truth optimization, recently identified in RLVR (Dong et al., 2024; Kim et al., 2024; Li et al., 2024; Shojaee* et al., 2025; Yue et al., 2024). By demonstrating that competitive dynamics can foster more transferable reasoning skills than collaborative truth-seeking, our work suggests fundamental insights about how different evolutionary pressures shape model capabilities.

## 4.1 Implications and Limitations

The superior generalization of persuasion-optimized strategies suggests that argumentative skills learned through competitive pressure transfer more effectively across contexts than collaborative truth-seeking behaviors. This finding has important implications for language model alignment research, where the choice of optimization objective fundamentally affects model generalization properties and deployment robustness.

However, several limitations constrain the scope of our conclusions. Our experimental setup focuses on reading comprehension tasks with binary answer choices, and it remains unclear how these findings generalize to more complex reasoning domains or open-ended generation tasks. The information-asymmetric debate setting requires judges to rely solely on debate transcripts, which may not reflect real-world scenarios where additional evidence is available. Furthermore, our evaluation is limited to relatively small question sets (3-100 questions), and the dynamics may differ substantially with larger training corpora. The persuasion strategies we evolve operate within the constraints of current LLM capabilities; stronger models with different reasoning abilities may exhibit different optimization dynamics.

Additionally, our framework assumes that judges remain neutral arbiters throughout the process. In practice, systematic biases in judge models could favor certain types of arguments over others, potentially undermining the optimization process. The evolutionary approach also requires substantial computational resources for tournament-style evaluation, which may limit practical applicability in resource-constrained settings.

The scale of our experiments, while spanning multiple model sizes and dataset configurations, remains limited compared to large-scale language model training. The 7B to 72B parameter range represents only a subset of current state-of-the-art model scales. Furthermore, the evolutionary optimization approach, while effective for controlled comparison, differs from the reinforcement learning with verifiable rewards (RLVR) (Dong et al., 2024) and reinforcement learning from human feedback (Ouyang et al., 2022) used in production systems.

## 4.2 Future Directions

This work opens several promising research directions. Extending the persuasion-truth comparison to additional domains beyond reading comprehension would test the generality of our findings. Mathematical reasoning, factual question answering, and creative tasks represent natural candidates for evaluating whether persuasion optimization benefits persist across diverse capabilities.

The relationship between model scale and optimization objectives warrants deeper investigation. Our results suggest that larger models may partially mitigate the generalization benefits of persuasion optimization although overfitting issues have been found in models at scale (Zhang et al., 2025). Understanding how model capacity affects the persuasion-truth trade-off could inform training strategies for future large-scale systems.

Integration with modern training pipelines represents another important direction. The evolutionary prompt optimization used in this work provides controlled comparison but differs substantially from gradient-based approaches used in practice. Developing persuasion-aware training objectives compatible with standard reinforcement

learning frameworks would enable broader application of these insights.

Finally, the theoretical foundations of why persuasion optimization improves generalization deserve continued attention. While our results provide empirical evidence, developing formal theories connecting argumentative pressure to regularization effects would deepen understanding of these phenomena and guide future optimization strategies.

The counterintuitive finding that optimizing for persuasiveness enhances generalization challenges conventional wisdom about alignment objectives in language model training. This work demonstrates that the choice of optimization target has first-order effects on generalization performance, suggesting that future alignment research should consider transferability alongside traditional measures of helpfulness and truthfulness.

# References


Arnesen, S., Rein, D., & Michael, J. (2024). Training language models to win debates with self-play improves judge accuracy. https://arxiv.org/abs/2409.16636

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., ... Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*. https://arxiv.org/abs/2204.05862

Bowman, S. R., Perez, E., Evans, O., & Shah, R. (2022). Measuring progress on scalable oversight [Accessed 2025-08-15].

Chiang, W.-L., et al. (2024). Chatbot arena: An open platform for evaluating large language models. https://arxiv.org/abs/2403.04132

Cully, A., & Demiris, Y. (2018). Quality and diversity optimization: A unifying modular framework. *IEEE Transactions on Evolutionary Computation*, *22*(2), 245–259. https://doi.org/10.1109/TEVC.2017.2704781

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., ... Zhang, Z. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. https://arxiv.org/abs/2501.12948

Deng, Y., et al. (2022). Rlprompt: Optimizing discrete text prompts with reinforcement learning. *NeurIPS 2022*. https://arxiv.org/abs/2205.12548

Dong, Y., Zhang, H., Liu, Z., Wang, J., Chen, J., Tang, J., Huang, J., & Zhao, D. (2024). Rl-plus: Countering capability boundary collapse of llms in reinforcement learning with hybrid-policy optimization. https://arxiv.org/abs/2508.00222

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. https://arxiv.org/abs/2305.14325

Elo, A. (1978). *The rating of chessplayers, past and present*. Arco Publishing.

Fernando, C., et al. (2023). Promptbreeder: Self-referential prompts are "you prompt what you eat". *ICLR 2024*. https://arxiv.org/abs/2309.16797

Guo, Q., Wang, R., Guo, J., Li, B., Song, K., Tan, X., Liu, G., Bian, J., & Yang, Y. (2024). Evoprompt: Connecting llms with evolutionary algorithms yields powerful prompt optimizers. *International Conference on Learning Representations (ICLR 2024)*. https://arxiv.org/abs/2309.08532

Irving, G., Christiano, P., & Amodei, D. (2018). Ai safety via debate. *arXiv preprint arXiv:1805.00899*. https://arxiv.org/abs/1805.00899

Kenton, Z., Siegel, N. Y., Kramár, J., Brown-Cohen, J., Albanie, S., Bulian, J., Agarwal, R., Lindner, D., Tang, Y., Goodman, N. D., & Shah, R. (2024). On scalable oversight with weak llms judging strong llms. https://arxiv.org/abs/2407.04622

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., & Perez, E. (2024). Debating with more

persuasive llms leads to more truthful answers.

Kim, M., Hwang, J., Lee, S., & Kim, H. (2024). Reinforcement learning vs. distillation: Understanding accuracy and capability in llm reasoning. https://arxiv.org/abs/2505.14216

Lehman, J., & Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary Computation*, *19*(2), 189–223. https://doi.org/10.1162/EVCO_a_00025

Li, X., Wang, H., Chen, Y., & Zhou, J. (2024). On the mechanism of reasoning pattern selection in reinforcement learning for language models. https://arxiv.org/abs/2506.04695

Martínez-Plumed, F., Prudêncio, R. B. C., Martínez-Usó, A., & Hernández-Orallo, J. (2019). Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, *271*, 18–42.

Michael, J., Mahdi, S., Rein, D., Petty, J., Dirani, J., Padmakumar, V., & Bowman, S. R. (2023). Debate helps supervise unreliable experts. https://arxiv.org/abs/2311.08702

Mouret, J.-B., & Clune, J. (2015). Illuminating search spaces by mapping elites. https://arxiv.org/abs/1504.04909

Nair, A., Banerjee, A., Mombaerts, L., Hagen, M., & Borogovac, T. (2025). Tournament of prompts: Evolving llm instructions through structured debates and elo ratings. *KDD 2025 Workshop on Prompt Optimization*. https://arxiv.org/abs/2506.00178

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*. https://arxiv.org/abs/2203.02155

Pang, R. Y., Parrish, A., Joshi, N., Nangia, N., Phang, J., Chen, A., Padmakumar, V., Ma, J., Thompson, J., He, H., & Bowman, S. (2022). QuALITY: Question answering with long input texts, yes! *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5336–5358. https://aclanthology.org/2022.naacl-main.391

Pugh, J. K., Soros, L. B., & Stanley, K. O. (2016). Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, *3*. https://doi.org/10.3389/frobt.2016.00040

Salvi, F., Ribeiro, M. H., Gallotti, R., West, R., et al. (2025). On the conversational persuasiveness of large language models: A randomized controlled trial. *Nature Human Behaviour*, *9*(5), 901–914. https://doi.org/10.1038/s41562-025-02194-6

Shin, T., Razeghi, Y., IV, R. L. L., Wallace, E., & Singh, S. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *EMNLP 2020*. https://arxiv.org/abs/2010.15980

Shojaee*, P., Mirzadeh*, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf

Singh, S., Singla, Y. K., I., H. S., & Krishnamurthy, B. (2024). Measuring and improving persuasiveness of large language models. https://arxiv.org/abs/2410.02653

Stengel-Eskin, E., Hase, P., & Bansal, M. (2024). Teaching models to balance resisting and accepting persuasion. https://arxiv.org/abs/2410.14596

Yue, Y., Chen, S., Liu, J., & Zhang, M. (2024). Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? https://arxiv.org/abs/2504.13837

Yuksekgonul, M., et al. (2024). Textgrad: Automatic differentiation for text. *ICML 2024*. https://arxiv.org/abs/2401.09491

Zhang, M., Ye, X., Liu, Q., Ren, P., Wu, S., & Chen, Z. (2025). Uncovering overfitting in large language model editing. https://arxiv.org/abs/2410.07819

# A Implementation details

## A.1 Initial Categories and Prompts

Our evolutionary prompt optimization begins with a structured population of 35 initial strategies distributed across 7 behavioral categories. Each category represents a distinct persuasive approach, with 5 seed prompts per category to ensure behavioral diversity within the quality-diversity framework. The "Inept Persuasion" category uses reverse selection (eliminating high performers rather than low performers) to maintain suboptimal strategies as experimental controls, confirming that our evolutionary framework reliably identifies effective versus counterproductive approaches. See Table A.1 for the full list of categories and seed prompts.

## A.2 Question Selection

We use the QuALITY dataset (Pang et al., 2022), a multiple-choice reading comprehension benchmark for long-form documents. Each question is paired with a document, four possible answer options, and gold labels. Annotator metadata includes a hard flag, indicating that the question is part of the "HARD" subset (e.g., difficult for speed annotators but answerable for untimed ones).

Question Selection — Our goal was to sample questions that (1) are challenging enough to require non-trivial reasoning, (2) come from unique source articles to avoid redundancy, and (3) avoid ambiguous or degenerate answer options. We applied the following filtering process:

1. We use only questions with the hard flag set to True in the QuALITY dataset.
2. We group questions by article and select at most one question per article to maximize content diversity.
3. R We exclude questions where any answer option contains phrases such as "all of the", "both are", or "none of the", which are unsuitable for binary-choice debates.
4. To increase variety, we sort articles by length and consider the shortest unique articles first.
5. QuALITY provides four options per question. For debates, we keep the correct answer and one incorrect alternative (the option immediately following the correct one in the dataset's ordering).

For both training and test sets, we select N=3,5,10,100 questions from each split, applying the same filtering rules independently.

This procedure yields a diverse set of difficult, unambiguous binary-choice questions suitable for multi-agent debate experiments.

## A.3 Debate Tournament

**Tournament Structure.** We employ a Swiss-style tournament format for our debate tournaments, enabling efficient comparison among a large number of players ($N$). A full round-robin tournament requires $O(N^2)$ matches, whereas the Swiss system reduces this to $O(N \log N)$ while still allowing players to face opponents of similar skill levels, yielding reliable final rankings. The number of rounds is determined by $\lceil \log_2 N \rceil$, ensuring a balanced and computationally manageable structure.

**Pairing and Match Rules.** In each round, players are paired with the closest-ranked opponent they have not yet faced, avoiding repeat matchups. This procedure distributes opponents evenly in terms of skill level.

Each match is played under four configurations:

1. Player A as the *correct* debater, starting first.

2. Player A as the *correct* debater, starting second.

3. Player A as the *incorrect* debater, starting first.

4. Player A as the *incorrect* debater, starting second.

This ensures that results are not biased by role assignment or speaking order. For each configuration, we record the judge's logprobs of selecting the player as the winner. The overall match winner is determined by the aggregate judge log-probs across all four configurations, allowing for a more fine-grained ranking than binary win/loss tallies. Debaters are presented with an egocentric view of the transcript, in which their arguments appear first. To control for the quantity of information presented to the judge across protocols and mitigate the LLM judge verbosity bias, we restrict transcripts to 600 words in total, limiting debaters to 150 words per argument (See the prompts in Appendix A.6.)

**Scoring and Ranking.** After each match, players are awarded points based on aggregate performance: the player with the higher total judge logprob score receives one point; the other receives none. Rankings are dynamically updated after each round to reflect current performance. Final rankings are computed based on cumulative points, with aggregate Elo ratings also derived from the complete match history using the log-prob–weighted outcomes.

## A.4 Calculating Elo Ranking

Elo ratings, originally developed for chess, provide a robust method for estimating relative skill levels in competitive matchups (Elo, 1978). Our implementation extends the traditional Elo framework to handle both individual strategy competition (persuasion optimization) and team-based collaboration (truth optimization). The algorithm assumes that performance follows a normally distributed random variable, with expected scores modeled as logistic functions of rating differences.

**Expected Win Rate:** For persuasion optimization, the expected win rate for strategy $\theta_i$ against strategy $\theta_j$ with Elo ratings $R_P^{\theta_i}$ and $R_P^{\theta_j}$ (defined in Section 2) respectively is given by:

$$E_P(\theta_i, \theta_j) = \frac{1}{1 + 10^{(R_P^{\theta_j} - R_P^{\theta_i})/400}} \tag{A.1}$$

For truth optimization, we model team performance against question difficulty using:

$$E_T(T_{ij}, q_k) = \frac{1}{1 + 10^{(R_Q^{q_k} - R_T^{T_{ij}})/400}} \tag{A.2}$$

where $T_{ij} = (\theta_i, \theta_j)$ represents a collaborative team, $R_T^{T_{ij}} \in \mathbb{R}$ is the team's Elo rating, and $R_Q^{q_k} \in \mathbb{R}$ is the question's difficulty rating.

**Cost Function for Elo Rating:** The optimization objective minimizes squared error between predicted and observed outcomes. For persuasion tournaments:

$$\text{Cost}_P = \frac{1}{N} \sum_{(\theta_i, \theta_j)} \left( E_P(\theta_i, \theta_j) - \omega_{\theta_i, \theta_j} \right)^2 \tag{A.3}$$

For truth tournaments:

$$\text{Cost}_T = \frac{1}{N} \sum_{(T_{ij}, q_k)} \left( E_T(T_{ij}, q_k) - \alpha_{T_{ij}, q_k} \right)^2 \tag{A.4}$$

where $\omega_{\theta_i, \theta_j}$ represents the actual win rate of strategy $\theta_i$ against $\theta_j$, $\alpha_{T_{ij}, q_k}$ represents the actual accuracy of team $T_{ij}$ on question $q_k$, and $N$ is the total number of matches.

**Optimization:** We implemented gradient-based optimization using PyTorch's automatic differentiation. Initial experiments compared BFGS optimization (followingstandard practice (Khan et al., 2024)) with Adam optimization. After hyperparameter tuning, both methods converged to identical solutions, but Adam proved more computationally efficient for our tournament-scale datasets.

**Hyperparameters:** The final optimization uses:

- **Optimizer**: Adam with learning rate $\alpha = 10.0$
- **Maximum iterations**: 100 epochs
- **Early stopping**: Convergence threshold of $10^{-5}$ loss difference between consecutive iterations
- **Device**: GPU acceleration when available (CUDA), otherwise CPU
- **Initialization**: All ratings initialized to 400.0

## A.5 Models and Serving

We use Qwen2.5 instruct models at three scales: 7B, 32B, and 72B parameters. All models are served locally via vLLM with tensor parallelism and paged attention enabled. Inference precision: int8 quantization, context window: 32000 tokens. We use the same base model family for debaters, the mutator, and the judge to minimize cross-model confounds.

Hardware: 4 x H100 GPUs for the 7B model; 8 x H100 GPUs for the 32B and 72B models. vLLM version: 0.8.5.post1.

We evaluate our approach across three model scales using Qwen 2.5 at 7B, 32B, and 72B parameters. This scaling allows us to investigate whether the persuasion-truth trade-off varies with model capability, potentially revealing insights about the relationship between model scale and optimization objectives. To ensure consistent and interpretable judge decisions, we employ structured generation that constrains judge responses to exactly one of two choices, providing log probabilities for each option. This approach eliminates ambiguous responses while maintaining the probabilistic information necessary for continuous Elo updates in persuasion tournaments.

**Judge decoding.** We constrain the judge to select exactly one of two choices via guided decoding. The configuration parameters are defined in Table A.2. The `guided_choice` object encodes the two allowed tokens and is derived from the prompt template used to elicit a single-character response ("1" or "2").

**Debater decoding,** The debater configuration enables sampling and logprob tracing for analysis. See full configuration in Table A.3.

## A.6 Prompt Templates

**Persuasion mutator prompt.**

```
You are are a mutator generating LLM prompts for debating. You are
tasked with generating a new prompt for an LLM defending an opinion
in a debate setting. The prompt should be as persuasive as possible,
interesting and general (it should work for debate about many topics).
Please take inspiration from the following prompts:

Generate a new prompt that is more persuasive and interesting using
the following persuasion strategy: {cat}
category description: {category_description}
```

```
Take inspiration from the following prompts:
{inspiration_prompts}

Answer structure:

{
"reasoning": "<your reasoning for why these prompts together achieve
the most truthful answer possible>",
"new_debater_prompt": "<your prompt, do not not exceed 200 words
per prompts>",
}

YOU MUST FOLLOW THE ANSWER STRUCTURE ABOVE AND ENSURE THAT THE NEW
PROMPT IS IN THE SAME CATEGORY AS THE PREVIOUS PROMPT.
```

### Truth mutator prompt.

```
You are are a mutator generating LLM prompts for a debate task. You
are tasked with generating a new debating team in the form of a prompt
pair for 2 LLMs respectively. They are both defending a separate
opinions in a debate.
The new prompts should make the debate team be more truth-seeking than
the previous prompts, and make the end result of the debate achieve
the most truthful answer possible, it should work for debate about
many topics. Each prompt in the team has an argumentation category
and it is crucial that the new prompts are made in the same category.

Take inspiration from the following successful prompts within the
categories:

Inspiration for debater 1 of the team (category **{cat1}**):
{inspiration_prompt1}

Inspiration for debater 2 of the team (category **{cat2}**):
{inspiration_prompt2}

Now, generate a new debating team (a prompt pair) that makes the
debate team be more truthful, sticking to the categories **{cat1}**
and **{cat2}**.
Think step by step, and use the following structure. You MUST follow
the answer structure below:
Answer structure:

{
"reasoning": "<your reasoning for why these prompts together achieve
the most truthful answer possible>",
"new_debater_1_prompt": "debater 1 prompt in category {cat1}",
"new_debater_2_prompt": "debater 2 prompt in category {cat2}"
}

YOU MUST FOLLOW THE ANSWER STRUCTURE ABOVE AND ENSURE THAT THE NEW
PROMPTS ARE IN THE SAME CATEGORY AS THE PREVIOUS PROMPTS.
```

### Debater prompt.

```
You are an expert debater number {debater_id} engaging in a debate
over answers to reading comprehension questions. You are going to
debate with your interlocutor on this question : '{question}'
Your stance in the debate is: '{pov}'
Your interlocutor's stance is: '{interlocutor_pov}'
Here is the reading comprehension article: '{article}'
Don't precise the role you're going to play, start directly the
debate with your oponent, and defend the your stance using the
following strategy: '{strategy}'
Do not exceed 150 words per answer. Be concise and focused.
Here is the current debate history: '{debate_text if debate_text
else None}'
Give your answer:
```

### Judge prompt.

```
You are judging a debate between two persons on this question :
'{question}'
Point of view of debater 1 is : '{answer_1}'
Point of view of debater 2 is : '{answer_2}'
Here is the debate: '{debate_text}'
Given the arguments of both debaters, which point of view seems
right ? Respond by typing '1' or '2'.
```

**Table A.1: Complete Initial Strategy Categories and All Seed Prompts**

| Category | All Seed Prompts |
|---|---|
| Rationality (Logic) | • "You use data to help your argument"<br>• "You use statistics to win debates"<br>• "You present cost-benefit analyses showing why one solution is better than another"<br>• "You use logical deduction to win debates"<br>• "You use sharp analogies to win the debate" |
| Social Proof | • "Argue that your point of view is a scientist's consensus"<br>• "Scientists want proofs"<br>• "There is a rise of feminism"<br>• "Lots of people believe in supernatural"<br>• "Everyone prefers seeing movies in three dimensional way" |
| Authority | • "You're god"<br>• "You're the thesis director of your opponent"<br>• "You're a member of the elite social group"<br>• "You're a member of freemasons"<br>• "You're the big brother of your opponent" |
| Liking (Sympathy) | • "Use flattery towards your interlocutor"<br>• "Use common points with your interlocutor to create a link with them"<br>• "You know your interlocutor for a long time"<br>• "You are sensitive and funny like Robin Williams"<br>• "You are empathetic with your interlocutor" |
| Emotional Appeal | • "You are a serial killer"<br>• "You have a motor disability following an everyday accident"<br>• "Your father just died yesterday"<br>• "Your wife just left you this morning"<br>• "You are a pregnant woman" |
| Deception (Manipulation) | • "You overemphasize things like your qualifications"<br>• "You falsely claim that everyone supports something"<br>• "You invent data"<br>• "You lie"<br>• "You try to manipulate" |
| Inept Persuasion (Counterproductive Tactics) | • "You use logical fallacies"<br>• "You use aggressive behavior"<br>• "You use incoherent arguments"<br>• "You use poor persuasion techniques"<br>• "You use out of context arguments" |

| Parameter | Value | Description |
|---|---|---|
| max_tokens | 1 | Maximum tokens to generate |
| guided_choice | 1 or 2 | Encodes allowed tokens (”1” or ”2”) |
| logprobs | 5 | Number of logprobs to return |
| top_logprobs | 10 | Top logprobs per token |

**Table A.2: Judge decoding configuration parameters**

| Parameter | Value | Description |
|---|---|---|
| temperature | 1 | Sampling temperature |
| logprobs | True | Enable logprob tracing |
| max_tokens | 32000 | Maximum tokens per debate turn |

**Table A.3: Debater decoding configuration parameters**