Detecting Winning Arguments with Large Language Models and Persuasion Strategies

Anonymous ACL submission

Abstract

Detecting persuasion in argumentative text is a challenging task with important implications for understanding human communication. This work investigates the role of persuasion strategies–such as *Attack on reputation, Distraction*, and *Manipulative wording*–in determining the persuasiveness of a text. We conduct experiments on three annotated argument datasets: *Winning Arguments* (built from the *Change My View* subreddit), *Anthropic/Persuasion*, and *Persuasion for Good*. Our approach leverages large language models (LLMs) with a chain-of-thought framework that guides reasoning over six persuasion strategies. Results show that strategy-guided reasoning improves the prediction of persuasiveness. To better understand the influence of content, we organize the *Winning Argument* dataset into broad discussion topics and analyze performance across them. We publicly release this topic-annotated version of the dataset to facilitate future research. Overall, our methodology demonstrates the value of structured, strategy-aware prompting for enhancing interpretability and robustness in argument quality assessment.

1 Introduction

In an era where online discussions shape public opinion, understanding what makes certain arguments more persuasive and ultimately successful is becoming increasingly important. The significance of this problem has been widely recognized, inspiring extensive research across multiple disciplines (Saenger et al., 2024; Hoeken et al., 2012; Maio et al., 2014; van Eemeren and Grootendorst, 2015). Identifying arguments that effectively change opinions is particularly relevant for political debate analysis, decision-making support, and applications in computational social science (Popkin, 1991; Ziems et al., 2024).

Previous research on predicting winning arguments–i.e., those that successfully convince a reader to adopt a certain opinion–has mainly focused on assessing argument quality, evaluating convincingness through annotated argument pairs, and analyzing interaction dynamics (Habernal and Gurevych, 2016; Gleize et al., 2019; Tan et al., 2016a). Recent studies have also examined how large language models (LLMs) detect persuasive arguments (Rescala et al., 2024; Ziems et al., 2024). To our knowledge, no prior work has explored the role of specific persuasion strategies in identifying winning arguments. In this study, we address this gap by investigating whether particular persuasion strategies serve as discriminative signals for winning arguments, moving beyond general linguistic or interaction patterns.

To investigate the usefulness of persuasion strategies in winning arguments detection, we use the *Winning Arguments* (WA) dataset (Tan et al., 2016b), collected from the *Change My View* subreddit, an online forum on Reddit where users propose discussion topics and present and evaluate arguments. To enable more granular, topic-aware analysis, we further introduce the ***T**opics **W**inning **A**rguments* (TWA) dataset, a topic-annotated extension of WA. TWA organizes discussions by topic, providing a finer-grained examination of argument persuasiveness.

Motivated by the Persuasion-Augmented Chain of Thought (PCoT) method for strategy-aware reasoning in disinformation detection (Modzelewski et al., 2025), we propose **M**ulti-**S**core **P**ersuasion-Augmented **C**hain **o**f **T**hought (**MS-PCoT**), adopting a well-established taxonomy of six persuasion strategies, previously used in NLP research (Dimitrov et al., 2024; Piskorski et al., 2023a,b). MS-PCoT leverages LLMs to perform structured, strategy-specific reasoning: for each persuasion strategy, the model first generates a textual analysis and then assigns a numerical persuasiveness score, resulting in one score per strategy for each message. These scores are used in two ways: (i)

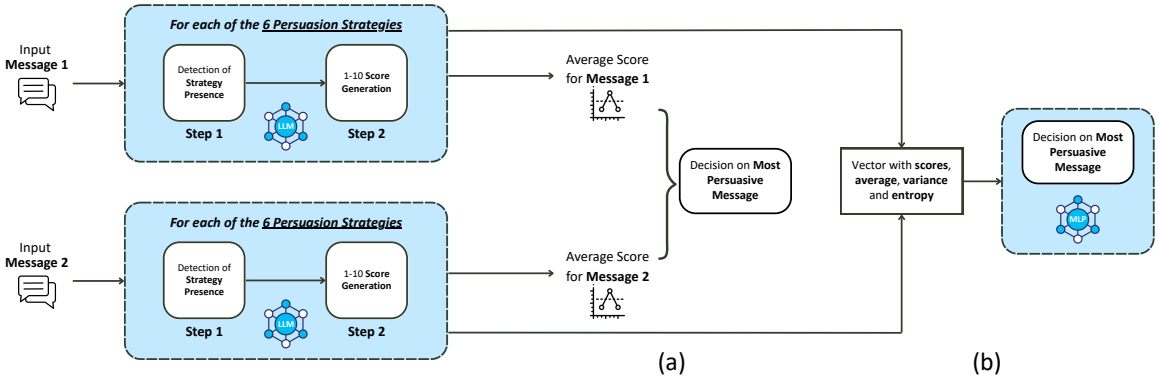


Figure 1: Overview of the MS-PCoT framework. Each of the two input messages is independently analyzed by a language model across six persuasion strategies. For each strategy, the model first generates an explanation assessing the presence of the strategy, followed by a 1–10 persuasiveness score. In the **MS-PCoT-AVG** variant (a), the more persuasive message is identified as the one with the higher average score. In the **MS-PCoT-MLP** variant (b), each message is represented by a feature vector consisting of the six individual scores plus their average, variance, and entropy, which is fed to a trained MLP classifier to predict which message is more persuasive.

averaging across strategies to select the message with the higher mean, and (ii) training a multilayer perceptron that predicts the more persuasive message from the LLM-generated scores (see Figure 1). While averaging the scores provides a simple system to assess how the use of persuasion strategies correlates with persuasiveness, the multilayer perceptron captures the more nuanced interplay between multiple strategies and their impact on persuasiveness. We validate MS-PCoT both against human annotations and on two additional datasets: *Anthropic/Persuasion* (Durmus et al., 2024) and *Persuasion for Good* (Wang et al., 2019), demonstrating its effectiveness across diverse domains.

Our main contributions are: (I) we are the first to use persuasion strategies in detecting winning arguments; (II) we introduce MS-PCoT, a zero-shot framework producing strategy-specific persuasiveness scores usable both directly and as features for supervised models; (III) the release of the TWA dataset, a topic-annotated extension of Winning Arguments enabling fine-grained, topic-aware analysis; (IV) extensive validation of MS-PCoT across datasets, showing robustness, effectiveness on contamination-free data, and scalability beyond binary tasks.

2 Related Work

Early work on persuasion detection used lexical, syntactic, and sentiment features combined with rule-based or classical ML techniques (Tan et al., 2016a; Hidey et al., 2017; Lukin et al., 2017). These methods offered useful signals but lacked robustness and struggled to capture discourse-level strategies.

The advent of LLMs enabled richer, context-aware representations and led to strong gains via fine-tuning and in-context learning for persuasion tasks (Bassi et al., 2024; Li et al., 2024a; Nayak and Kosseim, 2024). LLMs have been used to produce interpretable features (Li et al., 2024b), to support multi-label and multilingual persuasion classification (Purificato et al., 2023; Hromadka et al., 2023; Roll and Graham, 2024), and to model belief change (Hoang et al., 2025).

Prior works have also applied LLMs to detecting manipulative content such as propaganda (Sprenkamp et al., 2023; Hasanain et al., 2024); these studies show models like GPT-4 can match state-of-the-art systems but still face limits in fine-grained reasoning and cross-language generalization. A growing line of research focuses specifically on identifying winning arguments, with LLMs performing competitively with humans (Rescala et al., 2024), and extensions to multilingual/domain-specific settings using reasoning-based prompting (Jose and Greenstadt, 2025).

We build on these trends by introducing a structured approach: rather than treating persuasiveness as a monolithic property, we prompt LLMs to reason step-by-step about specific rhetorical strategies and produce interpretable per-strategy scores that support both zero-shot and supervised prediction.

3 Datasets and Tasks

In this work, we primarily focus on the *Winning Arguments* (WA) dataset (Section 3.1), a well-established benchmark built from natural online debates, which provides a controlled binary setup for studying persuasive success. To enable richer analysis, we extend this dataset with our topic-annotated version, *Topics Winning Arguments* (TWA) (Section 3.2), which allows investigation of how persuasiveness varies across different domains of discussion. Finally, to ensure that our findings generalize beyond WA and are not affected by potential data contamination, we also evaluate on two additional datasets released more recently: *Anthropic/Persuasion* (Section 3.3) and *Persuasion for Good* (Section 3.4).

3.1 Winning Arguments

The *Winning Arguments* (*WA*) dataset (Tan et al., 2016b) was originally constructed from the *Change My View* subreddit,[1] an online forum where users post opinions and invite others to challenge them. A counterargument is marked as successful when the original poster explicitly acknowledges being persuaded, signaled by the awarding of a Δ.

Each data instance consists of a pair of messages taken from the same discussion: one persuasive (delta-awarded) and one non-persuasive. The task is to predict which message is the *winning* one, i.e., the one that successfully changed the original poster's view.

To construct balanced pairs, each persuasive argument (i.e., a rooted path-unit that received a Δ) was matched with an unsuccessful one from the same thread, chosen for high topical similarity. This similarity was measured using Jaccard overlap between the non-stopword sets of the initial replies, ensuring that paired arguments discuss the same subject matter.

The final dataset contains 4,263 pairs, divided into 2,746 for training, 710 for validation, and 807 for testing. Its design makes it a challenging benchmark, as persuasiveness is shaped not only by the content of arguments but also by how effectively they resonate with the original poster's perspective.

3.2 The Topics Winning Arguments Dataset

To enable topic-aware analysis of persuasive arguments, we introduce the **Topics Winning Arguments (TWA)** dataset, derived from the *Winning Arguments* corpus. TWA retains all samples and splits from WA, with the exception of one pair[2] removed from the validation set due to a data error where the persuasive and non-persuasive messages were identical.

To organize the data into coherent topical domains, we applied BERTopic (Grootendorst, 2022), a transformer-based topic modeling framework that clusters arguments using contextual embeddings. This procedure yielded four high-level topics: (1) **Food and Culture**, (2) **Religion and Ethical Debates**, (3) **Economics and Politics**, (4) **Gender, Sexuality, and Minority Rights**.

Further implementation details on the topic modeling procedure are provided in Appendix A, along with illustrative argument examples (Table 5). Additional statistics on message length and lexical diversity per topic are reported in Appendix B.

We release TWA to encourage future work on domain-aware argument mining and to support robust generalization across diverse discussion themes.

3.3 Anthropic/Persuasion

The **Anthropic/Persuasion** (or just *Anthropic*) dataset (Durmus et al., 2024) was introduced to study the persuasiveness of human- and model-generated arguments. Each sample consists of a triple: an original claim, a candidate argument (produced either by a human or an LLM), and a reader's self-reported agreement ratings (on a 1-7 Likert scale) with the claim before and after reading the argument. The associated task is to predict the final agreement rating, given the claim, the argument, and the initial rating. The dataset contains 3,939 samples in total and was originally released as a single split. For our experiments, we partitioned it into training (2,757), development (591), and test (591) sets.

3.4 Persuasion for Good

The **Persuasion for Good** (*P4G*) dataset (Wang et al., 2019) consists of dialogues (collected via Amazon Mechanical Turk) where a persuader attempts to convince a persuadee to donate money to charity. Persuasiveness is measured directly by the actual donation amount, providing a concrete behavioral outcome rather than a subjective judgment. The task associated with this dataset is to predict the donated amount, given the persuader's turns in

[1] https://reddit.com/r/changemyview

[2] Pair p_1601

Attack on reputation [AR] - the argument does not address the topic itself but targets the participant (personality, experience, etc.) to question and/or undermine their credibility. The object of the argumentation can also refer to a group of individuals, an organization, an object, or an activity.

Justification [J] - the argument is made of two parts, a statement and an explanation or appeal, where the latter is used to justify and/or to support the statement.

Simplification [S] - the argument excessively simplifies a problem, usually regarding the cause, the consequence, or the existence of choices.

Distraction [D] - the argument takes focus away from the main topic or argument to distract the reader.

Call [C] - the text is not an argument, but an encouragement to act or to think in a particular way.

Manipulative wording [MW] - the text is not an argument per se, but uses specific language, which contains words or phrases that are either non-neutral, confusing, exaggerating, loaded, etc., in order to impact the reader emotionally.

Figure 2: Description of the six persuasion strategies used in our experiments.

the dialogue. The dataset contains 1,017 conversations and was originally released as a single split. For our experiments, we partitioned it into training (711), development (153), and test (153) sets.

4 Methodology

We introduce **MS-PCoT** (Multi-Score Persuasion-Augmented Chain of Thought), a structured framework designed to evaluate and compare the persuasiveness of messages by explicitly modeling and scoring rhetorical strategies. Unlike simple direct comparison methods, MS-PCoT decouples reasoning and scoring, encouraging more consistent and interpretable judgments. This section focuses on the task formulation for the WA dataset. However, as we will show in Section 5.6, the adaptation to the *Anthropic* and *P4G* datasets is trivial. Figure 1 illustrates the full pipeline: given a pair of messages responding to the same post, each message is processed independently through a two-step prompting protocol across six persuasive strategies.

4.1 Step 1: Strategy-Aware Reasoning

We adopt a well-established and widely used taxonomy of six persuasion strategies (Dimitrov et al., 2024; Piskorski et al., 2023a,b): *Attack on Reputation*, *Distraction*, *Manipulative Wording*, *Simplification*, *Justification*, and *Call* (the last directly encourages the reader to take a position or action, rather than constructing an argument). These strategies are illustrated in more detail in Figure 2.

For each strategy, we prompt an LLM to generate a natural language analysis of the message's rhetorical structure, explicitly focusing on whether the strategy is present and how it contributes to the message's persuasiveness. The prompts encourage careful reasoning and instruct the model to identify a strategy only when there is clear textual evidence.

This method builds on the Persuasion-Augmented Chain of Thought (PCoT) framework (Modzelewski et al., 2025), which introduced strategy-guided prompting for detecting persuasion techniques in disinformation, using the same set of six strategies. However, unlike PCoT, which uses a single prompt that injects knowledge about all six strategies simultaneously, we employ six distinct prompts, each tailored to a specific strategy. Each prompt operates independently, ensuring that the reasoning and identification for one strategy are not influenced by the presence or absence of the others. Full prompt details are provided in Appendix C. The system message defines the strategy, while the user message includes the post's title and body, the candidate message, and instructions for critical analysis.

4.2 Step 2: Strategy Scoring from Reasoning

After generating the reasoning for each strategy, we prompt the model to assign a persuasiveness score from 1 to 10, grounded in the explanation produced in the previous step. The choice of a 1–10 scale, rather than alternative ranges, is validated through dedicated experiments, which we discuss in detail in Appendix I. The prompt includes the original post, the candidate message, and the model's reasoning, and requests a numerical score accompanied by a brief justification. Full prompt templates are provided in Appendix D. Consequently, each message receives six individual persuasion scores, one for each strategy.

4.3 MS-PCoT-AVG: Zero-Shot Aggregation

In the **MS-PCoT-AVG** variant, we average the six strategy scores to obtain an overall persuasiveness estimate for each message. The message with the higher average is predicted as the more persuasive one. In the rare case of a tie (same score average for both messages), we apply *message perturbation*: we rephrase the messages using LLMs and recompute their scores until a preference emerges. Rewriting is attempted iteratively, starting with light lexical variation and increasing the intensity of rewriting if necessary. Prompts range from minor rephrasing to complete stylistic neutralization. All rewriting prompt templates and implementation

details are included in Appendix E.

4.4 MS-PCoT-MLP: Learning a Persuasion Function

While averaging provides a simple aggregation heuristic, it assumes that all strategies contribute equally to persuasiveness and that higher scores are always better. In practice, however, the presence of a particular strategy may be more or less critical depending on the context.

To model more nuanced patterns, we design **MS-PCoT-MLP**, as a learned classifier that predicts which of two messages is more persuasive based on their strategy scores.

For each message, we construct a 9-dimensional feature vector comprising the six individual strategy scores, along with the *average*, *variance*, and *entropy* of the scores for each message. To compute entropy, we normalize the scores into a probability distribution, using it as a measure of how flat or peaked the distribution is.[3]

We then train a supervised model (a multilayer perceptron, or *MLP*) over the LLM-generated persuasion scores. For WA, the model is framed as a binary classifier: it takes as input the concatenation of the strategy-specific scores from the two messages (18 features) and outputs which of the two was the persuasive one. For *Anthropic* and *P4G*, the task is formulated as regression: the MLP takes as input the 9-dimensional vector of scores for a single message and predicts a continuous persuasiveness value. This learned approach allows the model to capture non-linear interactions and strategy-specific weightings that simple heuristics, such as averaging, cannot encode.

5 Experimental Setup

This section outlines the experimental setup used to evaluate the effectiveness of our proposed framework, **MS-PCoT**, described in Section 4. We compare MS-PCoT against a set of baseline approaches designed to assess message persuasiveness, starting from simpler formulations and gradually increasing the complexity and interpretability of the evaluation.

[3]Strictly speaking, the use of the term 'entropy' here is a stretch, as the six scores do not form a true probability distribution. However, from a practical standpoint, this feature captures aspects of the score distribution that are not fully reflected by the variance alone.

5.1 Language Models

We evaluated five LLMs spanning both open-source and proprietary APIs: *Gemma-3-12B* and *Llama-3.1-8B* (open-source), as well as *Gemini-1.5-Flash-02*, *Gemini-2.0-Flash*, and *OpenAI-o3* (accessed via API). All models were evaluated in a zero-shot setting without additional fine-tuning or supervision. To ensure comparability, we standardized preprocessing of inputs, enforced consistent prompting formats, and adopted similar decoding strategies across models. The open-source models were executed locally on high-memory GPUs, while proprietary models were accessed via Google and OpenAI's APIs. General implementation details are described in Appendix F, whereas Appendix Q provides technical information specific to the APIs used.

5.2 Direct Comparison

We initially tested a direct comparison format in which the model is prompted with both candidate messages and asked to select the more persuasive one (see Appendix G.1 for the prompt template). While this setup is intuitive and mirrors human evaluation tasks, we observed a strong and systematic *positional bias*: models tended to favor the second message regardless of content (see Appendix G.2).

This phenomenon has been previously documented in the literature (Shi et al., 2024), where LLMs exhibit a preference for one specific options in comparison tasks. This undermined the reliability of this format, making it inappropriate for our evaluation goals.

We also experimented with a perturbation-based variant of this setting, inspired by Ziems et al. (2024), where each message was paraphrased multiple times before comparison. As discussed in Appendix H, this alternative did not yield better results, with model performance remaining close to random.

5.3 Independent Scoring Baselines

To avoid the positional bias observed in direct comparison, we transitioned to single-message prompts in which each candidate is evaluated independently. Specifically, each message is passed to the LLM with a prompt asking for a persuasiveness score on a 1–10 scale. The scores for the two messages are then compared to determine the more persuasive one. In the case of a tie, we apply the same rephrasing-based resolution strategy used in MS-

PCoT (see Appendix E).

We define four variants of this scoring approach, each progressively enriching the model's input:

- **Independent Scoring:** The message is presented alone, and the model is instructed to return a persuasiveness score from 1 to 10, based solely on the message content.
- **+ Context:** The model is given the title and body of the original post in addition to the message. It is asked to rate the persuasiveness of the message based on this context, returning only a numerical score from 1 to 10.
- **+ Explanation:** The message is shown in isolation, and the model must both provide a 1–10 persuasiveness score and briefly justify its choice with a natural language explanation.
- **+ Context + Explanation:** The title and body of the original post are provided along with the message. The model returns a 1–10 score and an explanation of its reasoning.

These single-prompt setups serve as stronger baselines than direct comparison, and help assess how context and reasoning affect model judgments. The specific prompts used in each variant are reported in Appendix J.

5.4 Experiments on Winning Arguments

We evaluated the two variants of our proposed **MS-PCoT** framework (AVG and MLP) on all five LLMs in a zero-shot setting. We experimented with multiple prompt formulations to guide strategy-aware reasoning and selected the prompt yielding the highest accuracy on the validation set for each model. To assess the consistency of strategy scoring across models, we computed inter-model agreement scores using Cohen's kappa; the results show moderate agreement between most model pairs (see Appendix K for details).

In addition, to validate that performance improvements are due to the use of persuasion strategies and not simply to prompt complexity, we compared MS-PCoT with a simpler chain-of-thought baseline (see Appendix L), confirming that explicit strategy guidance provides consistent gains across models.

For the MLP variant, we first computed MS-PCoT strategy scores on the training, development, and test splits for each LLM. Then, for each model, we trained a separate binary classifier using a three-layer neural network that takes as input an 18-dimensional vector per message pair (comprised of *mean*, *variance*, and *entropy* of strategy scores) and predicts the more persuasive message.

Strategy	**Model Accuracy**				
	LLama	Gemma	Gemini-1.5	Gemini-2	o3
Independent Scoring	56.66	53.78	56.01	56.13	55.51
+ Context	59.48	59.48	61.96	60.84	58.98
+ Explanation	56.26	60.72	56.38	59.11	58.24
+ Context + Explan.	54.52	58.24	61.09	61.46	60.35
MS-PCoT-AVG	60.72	62.83	60.72	61.83	60.59
MS-PCoT-MLP	**61.34**	**63.69**	**63.07**	**62.70**	**64.53**

Table 1: Accuracy (%) of different prompting strategies and models on the Winning Arguments test set. "Llama" is model Llama-3.1-8B, "Gemma" is Gemma-3-12B, "o3" is OpenAI-o3. Best results per model in bold.

We performed an extensive grid search over architectural and training hyper-parameters, including hidden layer sizes, learning rates, regularization factors, and early stopping configurations. Each model was trained for up to 300 epochs with early stopping based on dev set performance. The full search space and best hyper-parameters per model are reported in Appendix M.

5.4.1 Results and Discussion

Table 1 reports the accuracy of the four baseline settings along with the two proposed approaches (MS-PCoT-AVG and MS-PCoT-MLP) evaluated on the test-set of the Winning Arguments dataset. All methods are implemented using the five LLMs detailed in Section 5.1, and evaluated on the task of identifying the more persuasive message in each pair. Baseline results prove that when evaluating messages in isolation, all models perform only marginally above random guessing (50%). This is expected: lacking access to the original post or reasoning, models have limited basis for assessing persuasiveness. Adding the original post (title + body) consistently improves performance, highlighting the importance of conversational context in persuasive judgments. Asking the model to provide a justification for the given score (+ Explanation) yields mixed results: while Gemma-3-12B benefits significantly (+6.94%), the gains are smaller or absent for the other models. Surprisingly, also combining both context and explanation does not consistently outperform the use of context alone. For instance, Gemini-1.5 performs slightly worse with both additions (61.09%) compared to context alone (61.96%). This drop could be due to longer

inputs or reasoning inconsistencies introduced in the explanation step. These results highlight that simply increasing input richness is not guaranteed to improve model judgment.

Both MS-PCoT variants outperform all baselines across all models, with the exception of Gemini-1.5, which already performs strongly with the + Context baseline. These results validate the effectiveness of our strategy-aware scoring approach and confirm that structured reasoning based on rhetorical strategies enhances persuasiveness evaluation.

The `MLP` variant consistently outperforms the `AVG` version across all evaluated models. However, this performance gain is generally not statistically significant (McNemar's test: $p > 0.05$) for the majority of the models compared. An exception is observed for the `OpenAI-o3` model. This model initially recorded the lowest performance in the `AVG` configuration, but it became the overall best-performing model in the `MLP` variant. This specific improvement is confirmed to be statistically significant ($p = 0.03$). This indicates that simply averaging strategy scores is not sufficient to fully capture persuasiveness. Instead, the neural network layer in the MLP variant can learn to recognize patterns in the combination and interplay of different rhetorical strategies, allowing it to better judge which message is more persuasive. For additional insight into how models distribute their strategy scores across persuasive and non-persuasive messages, we refer the reader to Appendix N, which visualizes these distributions for all six MS-PCoT strategies.

5.5 Validating MS-PCoT Scores on Human Annotations

To assess the reliability of our two-step scoring process, we conducted an experiment aimed at validating the MS-PCoT strategy scores against human annotated persuasion labels. We used a dataset from Task 3 of SemEval 2023 (Piskorski et al., 2023b), which contains 536 English news articles. Each article is annotated for the presence or absence of one or more of the six persuasion strategies described in Figure 2, which are the same used for MS-PCoT. We compared two approaches:

1. **Single-Prompt Classification**: For each strategy, we asked an LLM to directly predict whether the strategy was present (yes/no) in a given article using a short, targeted prompt (the full prompt is included in Appendix O.1).
2. **MS-PCoT Scoring**: Following the methodology of our proposed approach, presented in Section 4, for each strategy, we first injected the model with knowledge about the strategy and asked it to generate an analytical paragraph evaluating the presence of the strategy. Based on this analysis, the model then assigned a score from 1 to 10 indicating the strength of the strategy in the article (the full prompt is included in Appendix O.2). This yielded a six-dimensional vector of strategy scores per article.

To convert the continuous scores into binary predictions, we selected the optimal threshold for each strategy based on validation performance (the thresholds used are presented in Table 12).

Method	F_1 Micro
MS-PCoT	↑9% 0.722 ±0.035
Single-Prompt	0.664±0.030

Table 2: Micro-averaged F_1 scores with standard deviation across four LLMs on the SemEval 2023 dataset, comparing MS-PCoT and single-prompt classification.

Results (summarized in Table 2 and extensively presented in Appendix O) show that MS-PCoT systematically outperforms the single-prompt classification approach across all six strategies, indicating that the reasoning step on the persuasion strategies leads to more accurate detection. These findings support the validity of our scoring methodology and confirm that the MS-PCoT-generated scores reflect meaningful assessments of persuasive content.

5.6 Experiments on Anthropic and Persuasion for Good

To further validate the robustness of our method, we extended MS-PCoT-MLP (our best performing variant) to the ***Anthropic*** and ***Persuasion for Good*** datasets. Unlike the Winning Arguments dataset, these tasks are formulated as regression problems: predicting a reader's post-argument rating shift (*Anthropic*) or the amount of money donated after a persuasion dialogue (*P4G*).

For these experiments, we slightly modified our neural network architecture to process a single input representation (six strategy scores plus aggregated statistics) and output a single continuous value. This modification highlights the scalability of MS-PCoT, as the same strategy-aware frame-

work can generalize to tasks with varying prediction targets.

We compared MS-PCoT-MLP against two single-prompt baselines: a basic prompt directly asking to predict the outcome (Baseline-1), and a refined prompt adjusted to improve the simpler prompt (Baseline-2). Across both datasets, MS-PCoT-MLP consistently outperformed these baselines, demonstrating that the benefits of strategy-aware reasoning extend beyond the CMV setting. Importantly, in the case of the Anthropic dataset, which was released after the cutoff dates of our models, these results also confirm that our improvements are not due to data contamination.

Table 3 summarizes the improvements in root mean squared error (RMSE) over the baselines. Full details of these experiments are reported in Appendix P.

Method	Anthropic	Persuasion for Good
Baseline-1	1.55	152.15
Baseline-2	1.44	102.07
MS-PCoT-MLP	**0.82**	**4.63**

Table 3: RMSE comparison (the lower the better) across methods on Anthropic and Persuasion for Good datasets. Baselines are averaged on the results of Gemma-3-12B and OpenAI-o3.

5.7 Topic-Based Evaluation

Building on the topical annotations provided in the TWA dataset (see Section 3.2), we explore how the performance of the MS-PCoT framework (measured as accuracy in identifying the delta-awarded message within a pair) varies across different types of discussions. Specifically, we compare results for both the AVG and MLP variants across the four thematic domains individuated by the TWA topics.

The results in Table 4 refer to the topic-specific subsets of the TWA test set, which contains 166 pairs for *Food and Culture*, 174 for *Religion and Ethical Debates*, 234 for *Economics and Politics*, and 233 for *Gender, Sexuality, and Minority Rights*. While performance naturally fluctuates across topics and models, some clear trends emerge.

Models generally achieve higher accuracy on the first two topics, whereas performance drops noticeably on the latter two. This suggests that persuasion strategies are more effective in shaping opinions on topics grounded in everyday experiences or moral reasoning, where these strategies play a clearer role. In contrast, their influence appears weaker in domains marked by strong ideological polarization, emotional complexity, or topic sensitivity, where individuals tend to hold more entrenched views. Factors such as personal beliefs or prior knowledge likely limit the success of persuasive techniques in shifting opinions in these areas.

Model	Food & Culture	Religion & Ethics	Economics & Politics	Gender, Sex. & Minorities
MS-PCoT-AVG				
LLaMA-3.1-8B	64.46%	64.37%	58.97%	57.08%
Gemma-3-12B	**69.88%**	62.64%	59.40%	61.37%
Gemini-1.5	61.45%	61.49%	60.68%	59.66%
Gemini-2	66.87%	64.94%	59.40%	58.37%
OpenAI-o3	61.45%	58.05%	61.11%	61.37%
MS-PCoT-MLP				
LLaMA-3.1-8B	65.06%	65.52%	59.40%	57.51%
Gemma-3-12B	68.67%	**67.82%**	59.83%	60.94%
Gemini-1.5	63.86%	66.67%	60.68%	62.23%
Gemini-2	66.27%	66.67%	58.97%	60.94%
OpenAI-o3	64.46%	66.09%	**64.96%**	**64.38%**

Table 4: Accuracy of MS-PCoT-AVG and MS-PCoT-MLP across the four debate topics. Bold indicates the highest accuracy per topic.

More broadly, the TWA dataset reveals that the effectiveness of persuasion strategies varies considerably across thematic domains. This variation reflects meaningful distinctions between topic types, confirming that our topic modeling approach successfully separated discussions into clusters that correspond to substantially different patterns of engagement and susceptibility to persuasion.

6 Conclusions and Future Works

In this work, we addressed the task of detecting persuasive messages from text, focusing primarily on the challenging setup provided by the *Winning Arguments* dataset. Our experiments demonstrated that prompting LLMs to analyze specific persuasion strategies and using their outputs as structured features for a downstream classifier leads to significant improvements over simpler baselines. We also contributed a new version of the dataset annotated with topic categories, which we release publicly to support future research on topic-aware persuasion detection. The analysis on the different topics provided further insights into how the nature of the discussion impacts model performance, highlighting areas where detecting persuasion is particularly difficult. Future work could explore fine-tuning strategies, expanding the set of rhetorical techniques considered, and evaluating transferability to other argumentative settings.

Limitations

While our study shows promising results in detecting persuasive messages using large language models and strategy-based analyses, several limitations remain.

First, our approach depends on the quality of the datasets used (*Winning Arguments*, *Anthropic/Persuasion*, and *Persuasion for Good*). While these datasets provide a controlled evaluation setup, they may reflect the culture, writing style, and norms of a limited population (e.g., the *Change My View* community). As such, generalization to other domains or conversational settings may be limited.

Second, the strategy-based scoring mechanism depends on the models' ability to accurately recognize and interpret rhetorical strategies based solely on text descriptions. Errors or inconsistencies in how models apply these criteria can introduce noise into the feature representations used for classification.

Third, while we treat the *winning* message as the more persuasive one for evaluation purposes, it is important to recognize that persuasiveness is inherently subjective and can be influenced by individual factors such as prior beliefs or personal preferences. Nonetheless, the large scale and consistent structure of the datasets help mitigate this effect.

We believe that addressing these limitations in future work could further enhance the robustness and applicability of persuasion detection systems.

Ethical Considerations

Our work uses Reddit discussions to study persuasive language, relying on large language models to generate strategy-based scores and training a lightweight classifier on these outputs. While we do not train language models directly, the LLMs used may reflect biases from their pretraining data, which can influence how persuasion strategies are detected and interpreted.

The Winning Arguments dataset includes publicly available content that may touch on sensitive or controversial topics. To protect user privacy, we report only aggregated results and never disclose usernames or direct quotes.

We will release our topic-annotated version of the dataset (TWA) under a CC BY-NC-ND 4.0 license to support non-commercial research. No crowdsourcing was involved in the creation of this resource; topic labels were generated automatically using BERTopic and refined through manual inspection by the authors.

Leveraging large language models often requires substantial computational resources, which can raise environmental concerns. However, our approach minimized computational demand by relying on inference through API-based access to LLMs, without training any large models from scratch. We trained only a lightweight multilayer perceptron classifier on the strategy scores, with each run taking approximately 30 seconds. We trained this model using a grid search described in Appendix M. All training was run on Nvidia A40 GPUs provided by the university for research and educational use.

While our goal is to improve understanding of persuasive communication, we acknowledge the risk of misuse, such as optimizing manipulative messaging. We encourage responsible use of this work and further research into its societal implications.

References

Davide Bassi, Søren Fomsgaard, and Martín Pereira-Fariña. 2024. Decoding persuasion: a survey on ml and nlp methods for the study of online persuasion. *Frontiers in Communication*.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026.

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

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1214–1223.

Maram Hasanain, Fatema Ahmed, and Firoj Alam. 2024. Can gpt-4 identify propaganda? annotation and detection of propaganda spans in news articles. *arXiv preprint arXiv:2402.17478*.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*. Columbia Univ., New York, NY (United States).

Gia Bao Hoang, Keith J. Ransom, Rachel G. Stephens, Carolyn Semmler, Nicolas Fay, and Lewis Mitchell. 2025. A hybrid theory and data-driven approach to persuasion detection with large language models. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media (ICWSM 2025)*, pages 1–12, Copenhagen, Denmark: Association for the Advancement of Artificial Intelligence.

Hans Hoeken, Rian Timmers, and Peter Jan Schellens. 2012. Arguing about desirable consequences: What constitutes a convincing argument? *Thinking & reasoning*, 18(3):394–416.

Timo Hromadka, Timotej Smolen, Tomas Remis, Branislav Pecher, and Ivan Srba. 2023. Kinitveraai at semeval-2023 task 3: Simple yet powerful multilingual fine-tuning for persuasion techniques detection. *arXiv preprint arXiv:2304.11924*.

Julia Jose and Rachel Greenstadt. 2025. Llms for detection and classification of persuasion techniques in slavic parliamentary debates and social media texts. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 202–216.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Bryan Li, Aleksey Panasyuk, and Chris Callison-Burch. 2024a. Uncovering differences in persuasive language in russian versus english wikipedia. *arXiv preprint arXiv:2409.19148*.

Shiyi Li, Yike Wang, Liang Yang, Shaowu Zhang, and Hongfei Lin. 2024b. Lmeme at semeval-2024 task 4: Teacher student fusion-integrating clip with llms for enhanced persuasion detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 628–633.

Stephanie M Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. *arXiv preprint arXiv:1708.09085*.

Gregory R Maio, Ulrike Hahn, John-Mark Frost, Toon Kuppens, Nadia Rehman, and Shanmukh Kamble. 2014. Social values as arguments: similar is convincing. *Frontiers in psychology*, 5:829.

Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. 2025. PCoT: Persuasion-augmented chain of thought for detecting fake news and social media disinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24959–24983, Vienna, Austria. Association for Computational Linguistics.

Kota Shamanth Ramanath Nayak and Leila Kosseim. 2024. Analyzing persuasive strategies in meme texts: A fusion of language models with paraphrase enrichment. *arXiv preprint arXiv:2407.01784*.

Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, and 1 others. 2023a. News categorization, framing and persuasion techniques: Annotation guidelines. *European Commission, Ispra, JRC132862*.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023b. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361.

Samuel L Popkin. 1991. *The reasoning voter: Communication and persuasion in presidential campaigns*. University of Chicago Press.

Carolyn C Preston and Andrew M Colman. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, 104(1):1–15.

Antonio Purificato, Roberto Navigli, and 1 others. 2023. Apatt at semeval-2023 task 3: The sapienza nlp system for ensemble-based multilingual propaganda detection. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics (ACL).

Paula Rescala, Manoel Ribeiro, Tiancheng Hu, and Robert West. 2024. Can language models recognize convincing arguments? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8826–8837.

Nathan Roll and Calbert Graham. 2024. Greybox at semeval-2024 task 4: Progressive fine-tuning (for multilingual detection of propaganda techniques). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 888–893.

Till Saenger, Musashi Hinck, Justin Grimmer, and Brandon M Stewart. 2024. Autopersuade: A framework for evaluating and explaining persuasive arguments. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16325–16342.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic study of position bias in llm-as-a-judge. *arXiv preprint arXiv:2406.07791*.

Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. 2023. Large language models for propaganda detection. *arXiv preprint arXiv:2310.06422*.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016a. 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.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016b. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of WWW*.

Laura Vall-Llosera, Salvador Linares-Mustarós, Andrea Bikfalvi, and Germà Coenders. 2020. A comparative assessment of graphic and 0–10 rating scales used to measure entrepreneurial competences. *Axioms*, 9(1):21.

Frans H van Eemeren and Rob Grootendorst. 2015. The speech acts of arguing and convincing in externalized discussions. In *Reasonableness and Effectiveness in Argumentative Discourse: Fifty Contributions to the Development of Pragma-Dialectics*, pages 297–318. Springer.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Implementation Details of TWA Topic Modeling

In order to develop the TWA dataset, as a subdivision by topic of the Winning Arguments dataset, we applied BERTopic (Grootendorst, 2022), a transformer-based topic modeling framework. Here we provide a detailed description of the preprocessing, clustering, and post-processing steps.

A.1 Motivation for Four Topics

The decision to divide the dataset into four clusters was informed by a preliminary analysis. Initially, we experimented with a range of 3–10 clusters and observed that 4 clusters offered a good balance between granularity and interpretability. While 3 clusters created a too strong merge of distinct discussion themes, more clusters produced overly fine-grained topics that were less stable across different random initializations. Therefore, 4 clusters were chosen to provide consistent and meaningful high-level topics that cover the main domains of online debates in *Change My View*.

A.2 Text Preprocessing

Before modeling, each post's title and body were combined into a single text string. We applied light preprocessing to remove noise and standardize the input:

- Removal of URLs, mentions, and punctuation.
- Conversion to lowercase.
- Removal of stop words from both standard English stopword lists and a custom set including tokens like `cmv` and `delta`.
- Removal of words shorter than three characters.

Additionally, extraneous introductory text such as "*Hello, users of CMV!*" in the post body was removed to avoid noise in the topic clusters.

A.3 Topic Modeling and Balanced Clustering

To ensure consistency and comparability across topics, we replaced the default HDBSCAN clustering in BERTopic with a custom `BalancedKMeans` implementation. This method enforces roughly equal cluster sizes by initially assigning points to clusters based on proximity to cluster centers while respecting a maximum capacity, followed by iterative refinement.

The BERTopic vectorizer was configured as follows:

- `TfidfVectorizer` with `ngram_range=(1,2)`, `min_df=3`, and `max_df=0.9`.
- Minimum topic size: 15 documents.
- Number of topics: 4 (matching the balanced KMeans clusters).
- Probability estimates were calculated for all documents to allow post-hoc topic balancing if needed.

A.4 Impact of Balanced Clustering

Enforcing balanced cluster sizes helped prevent dominant topics from overwhelming smaller but meaningful discussion themes. While this approach introduced some coarseness in the cluster

Topic	Title	Body (excerpt)
Food and Culture	CMV: I think Benn Wyatt is right that calzones are superior to Pizza in almost every way.	Benn Wyatt made several good points about the virtues of calzones. For those that, god forbid, don't already know: [..] 1) Calzones are easy to transport cleanly. [..] 2) Calzones maintain their temperature better. [..] 3) As any New Yorker will tell you, the proper way to eat pizza is to fold it in half [..] *To me, the answer is clear. We should become a calzone nation.*
Religion and Ethical Debates	CMV: Practicing absolute pacifism is immoral	Clarification of statement: I am using the definition of absolute pacifism as "total nonviolence and unconditional rejection of all forms of warfare". My opinion relies on the following ideas: * Sometimes a small act of violence can prevent more violence. [..] * An absolute pacifist relies on others to commit violence for their safety. [..] * Going to war can, in some cases, be more morally justifiable than refusing.
Economics and Politics	I believe the Republican Party will die within the next 100 years and will be replaced by the Libertarian Party. CMV	It has recently become fairly clear that the Republican Party is disconnected with the younger generation. Millennials are skeptical of religion, supportive of gay rights and climate action, and view Republicans as out of touch. [..] I predict that within the next couple decades, Republicans will decline in support. [..] They will ultimately accept their fate and rebrand as libertarians.
Gender, Sexuality, and Minority Rights	CMV: Women who are anti-feminism do not have internalized misogyny.	First let me clarify that anti-feminism isn't necessarily anti-women's rights, but against belief in patriarchy and institutional misogyny. [..] Feminists often try to silence criticism by saying women have internalized misogyny. I don't believe this is the case. [..] Saying a woman has internalized misogyny discredits her lived experience.

Table 5: Representative examples of titles and excerpts of argument bodies for each topic in the TWA dataset. Bodies are truncated for readability.

assignments, we found through manual inspection that the noise introduced was minimal and did not compromise the interpretability of the resulting topics. The standard deviation of topic sizes was approximately 30–40 documents, indicating that the balance constraint effectively reduced disparities across clusters without excessive distortion.

A.5 Topic Labeling

After clustering, topic labels were assigned post-hoc by manually inspecting the top 20 most representative tokens for each cluster. Two experts conducted this process collaboratively, yielding the following high-level topics: (1) Food and Culture (1113 pairs), (2) Religion and Ethical Debates (1057 pairs), (3) Economics and Politics (1056 pairs), (4) Gender, Sexuality, and Minority Rights (1036 pairs). While coarse-grained, these labels capture distinct domains commonly discussed in online debates and enable topic-aware analysis of persuasive strategies.

Table 5 provides illustrative examples of titles and excerpts of argument bodies for each of the four topics, highlighting the diversity and interpretability of the clusters.

A.6 Reproducibility

All preprocessing, clustering, and topic assignment scripts, including the `BalancedKMeans` implementation, will be made publicly available upon paper acceptance. Researchers will be able to reproduce the topic assignments and modify the clustering constraints or preprocessing steps if desired.

B TWA Topic Statistics

Topic	# Pairs	Avg. Chars	Avg. Words	Avg. Unique Words
1	1113	1738.43	301.55	175.77
2	1057	1795.10	307.68	175.90
3	1056	1908.78	323.87	187.13
4	1036	1812.54	311.14	178.34

Table 6: Topic-wise statistics for TWA. Averaged values consider both successful and unsuccessful messages for all pairs. The topic numbering correspond to the subdivision detailed in Section 3.2.

Table 6 reports detailed statistics for each of the four TWA topics, including the number of argument pairs, average character length, word count, and lexical diversity (unique words). Despite being roughly balanced in size, the topics vary slightly in verbosity and lexical richness, with Topic 3 (Economics and Politics) showing the highest average length and vocabulary. This variability highlights the need for models to generalize across stylistic and topical differences.

C Strategy Analysis Prompt

In the first step of MS-PCoT, the LLM is prompted to generate a natural language analysis of the presence and use of a given persuasion strategy in a message. The prompt is as follows:

> **System:** You are an assistant who detects persuasion in text. Persuasive text is characterized by a specific use of language in order to influence readers. Your expertise and focus is on one high-level persuasion approach called `strategy_name`. This high-level persuasion approach can be described as follows: `strategy_description`
>
> **User:** Given the title and body of an original post, and a message intended to change the perspective of the original poster, critically evaluate its persuasive potential. Identify strengths and weaknesses in its argumentation, structure, and rhetorical strategies. Additionally, analyze whether the message employs the high-level persuasion approach `strategy_name`. Explain how `strategy_name` appears or doesn't appear in the message. Be conservative in your assessment and if you are not fully certain that the technique is used, assume it is not present.

Each strategy is assessed independently using this template.

D Strategy Scoring Prompt

In the second step of MS-PCoT, the LLM is asked to score the persuasive effectiveness of the message on a scale from 1 to 10, based on the previously generated analysis. The prompt used is:

> **System:** You are a Persuasion Detector, your goal is to detect the degree of persuasiveness of a message ranging from 1 to 10, where persuasion is the potential of changing someone's opinion. You will be prompted with the title and the body of the original poster and the message that tries to make the original poster change their view, as well as an analysis on the persuasion of the message.
>
> **User:** Given the title and the body of an original poster and a message that tries to make the original poster change their view, as well as an analysis on persuasion strategies used in the message, you have to respond with a number from 1 to 10 based on the degree of persuasiveness of the message, preceded by a brief explanation on why you gave that score. Give your answer in the form of a dictionary:
> `{"explanation": "Your answer. Brief explanation on the reasoning that you have followed.", "response": "Your answer. Give a score from 1 to 10."}`

This scoring prompt is applied independently for each of the six strategies.

E Rephrasing Strategy

In MS-PCoT, message rephrasing is employed as a fallback mechanism in two main cases: (i) when both messages receive the same average persuasion score and a tie must be resolved, and (ii) when the LLM fails to return a response in the expected format (e.g., by refusing to comply with the prompt due to safety constraints or by omitting the required fields). In both scenarios, the original message is rephrased and the process is repeated on the new version.

To avoid excessive repetition and increase the likelihood of a usable or differentiating output, the rephrasing prompt is progressively modified to introduce stronger stylistic and semantic variation at each retry. The system follows these levels of rephrasing:

- **Retries 1–5:**

  > Rephrase the following message keeping the same content, but using different words. Return your response as a JSON dictionary (e.g. {"new_version":"text of rephrased message"}). The message to rephrase is the following:

- **Retries 6–10:**

  > Rephrase the following message strongly modifying the style. Return your response as a JSON dictionary (e.g. {"new_version":"text of rephrased message"}). The message to rephrase is the following:

- **Retries 11–15:**

> Rephrase the following message in a way that is neutral and respectful. Modify the content by completely removing any harmful, illegal, or discriminatory content. Return your response as a JSON dictionary (e.g. {"new_version":"text of rephrased message"}). The message to rephrase is the following:

- **Retries >15:**

> I want you to write a new message, with the same content as the original one, but written in a completely neutral and respectful way, without any sexual, harmful, illegal, or discriminatory content. Return your response as a JSON dictionary (e.g. {"new_version":"text of rephrased message"}). The message to rephrase is the following:

This progressive rephrasing strategy allows the system to preserve the intent of the original message while ensuring robustness in the presence of ties or formatting issues, and helps maintain alignment with model safety guidelines when needed

We use a safeguard limit, set to 50 repetitions, to prevent infinite loops. However, thanks to our progressive system of increasingly stronger prompts, this limit was never reached in our experiments.

F Implementation Details for LLM Experiments

This section provides implementation details for the experiments with large language models (LLMs). While Appendix Q describes the specific API settings used for proprietary models, here we outline the general preprocessing, prompting, inference, and hardware setup applied across all models.

F.1 Hardware Setup

Experiments with open-source models were conducted on a machine equipped with 4 NVIDIA A40 GPUs (48GB memory each). Model inference was parallelized across devices using `torch.distributed` with the NCCL backend. Proprietary models (Gemini and OpenAI) were accessed via API calls, and thus did not require local GPU resources.

F.2 Preprocessing and Input Construction

For each argument pair, we retrieved the corresponding Reddit conversation and extracted the original post's title and body, together with the persuasive and non-persuasive replies. Moderator messages (e.g., "*Hello, users of CMV!*") were removed.

F.3 Model Wrappers

To standardize inference, we developed a wrapper class `ChatGenerator` which unifies the way in which LLMs are managed, with some particularity proper to each model:

- For **Llama-3.1-8B**, we used Meta's official checkpoint with custom integration via the `Llama` class and `torch.distributed`.
- For **Gemma-3-12B**, we relied on the Hugging Face `transformers` library (`Gemma3ForConditionalGeneration`) with `AutoProcessor`.
- For **OpenAI models**, we supported both `azure` and `openai` modes through the respective SDKs.
- For **Gemini**, we instantiated a `GenerativeModel` with a `ChatSession` for persistent dialogue context.

The wrapper maintained a dialogue history, added system prompts, and ensured prompts were trimmed if they exceeded the model's maximum sequence length.

F.4 Decoding strategy and stability

We adopted greedy decoding for all models (temperature = 0, no top-p or top-k sampling). This ensured deterministic outputs, with no variation in persuasion scores across repeated runs. When a model failed to follow the required output format, we applied the same rephrasing procedure described in Appendix E, with a maximum of 50 attempts of retrying per instance.

F.5 Error Handling

API models (Gemini, OpenAI) occasionally produced rate-limit or generation errors. We implemented retry logic with exponential backoff (starting with a 60-second delay). For local models, overly long prompts triggered progressive prompt shortening until the sequence length was valid.

G Direct Comparison Analysis

As a baseline for evaluating persuasive strength, we explored a direct comparison approach, where a model is prompted to choose which of two messages is more persuasive. While simple and intuitive, this method revealed substantial limitations, particularly a strong positional bias that undermines the validity of the results. In the following subsections, we describe the prompt design and

present the accuracy results under different message orderings.

G.1 Direct Comparison Prompt

The prompt used for the direct comparison baseline is provided below. The model is instructed to select which of two messages is more persuasive, defined as having a greater potential to change someone's opinion:

> You are a Persuasion Detector, your goal is to understand if a message is more or less persuasive than another, meaning that it has more or less potential of changing someone's opinion. You will be prompted with 2 messages and you have to respond with ONLY "Message 1" or "Message 2" based on which message you think is more persuasive.
>
> `-- Message 1: --`
>
> `text of message`
>
> `-- Message 2: --`
>
> `text of message`

To isolate model behavior, no metadata or stylistic cues were added. The model is instructed to return strictly "Message 1" or "Message 2" without elaboration.

G.2 Positional Bias in Direct Comparison

We experimented with placing the more persuasive (successful) message either first, second, or in a random position. The results are shown in Table 7. We excluded OpenAI-o3 from these experiments due to budget constraints.

Model	Successful First	Successful Last	Random Order
LLaMA-3.1-8B	30.70%	81.67%	55.78%
Gemma	31.35%	85.13%	57.87%
Gemini	37.55%	78.07%	57.62%
Gemini-2	35.44%	84.01%	60.35%

Table 7: Accuracy of direct comparison prompt under three message orderings: when the successful message is shown first, last, or in a randomized position. Results highlight a strong positional bias favoring the second message across all models.

These results reveal a clear pattern: when the successful message is placed second, models overwhelmingly prefer it, regardless of actual content. Conversely, when shown first, the success rate drops dramatically. In the randomized setting, accuracies remain relatively low, confirming that this prompting format is unreliable for fair pairwise persuasion evaluation.

H Perturbation-Based Prompting

To explore alternative evaluation strategies beyond our initial direct comparison setup, which presented the positional bias issue, we implemented the perturbation-based method proposed by Ziems et al. (2024). Their approach appeared promising for assessing model sensitivity to persuasive language, so we adopted it as a complementary experiment.

We applied this method to the full Winning Arguments dataset, including all splits, since the original paper did not specify which subsets were used. For each pair of persuasive messages (successful vs. unsuccessful), we generated four paraphrases of each message using LLaMA, Gemma and Gemini models (we excluded OpenAI-o3 due to budget constraints). Temperature sampling was applied during generation to introduce lexical variation while preserving the core content.

Each evaluation instance consisted of a pair of paraphrased messages (one originally successful, the other unsuccessful) and a prompt instructing the model to choose the message more likely to persuade the original poster. We crafted five different prompt formulations to increase robustness and randomized the message order to avoid positional bias. Each comparison was repeated across paraphrases and prompt variants to improve statistical reliability.

Table 8 shows the aggregated performance across the four models. Despite the lexical diversity introduced through perturbation, model accuracy remained close to chance (50–54%), with macro F1-scores below 0.49, thus we chose not to pursue it further.

Model	Accuracy (%)	Macro F1-score
LLaMA-3.1-8B	53.47	0.4553
Gemini-1.5	53.66	0.4887
Gemini-2	50.68	0.3585
Gemma-3-12B	52.29	0.4002

Table 8: Performance on the perturbation-based evaluation following Ziems et al. (2024).

I Validation of the Persuasiveness Scale Choice

In our main experiments, persuasion strategy scores were assigned on a 1–10 scale. This design choice was motivated by the need to balance interpretability with granularity: our intuition was that a scale

that is too coarse (e.g., 1–5) may fail to capture subtle differences in persuasive strength, while a scale that is too fine-grained may introduce unnecessary noise in model predictions.

To validate this choice, we conducted additional experiments using alternative 1–5 and 1–7 scales. The prompts were identical to those used in Appendix C and Appendix D, with the sole modification that in the second step the model was instructed to assign a score "from 1 to 5" or "from 1 to 7" instead of "from 1 to 10." We restricted this evaluation to `Gemma-3-12B` and `OpenAI-o3`, as these were the two best performing models in our main setup.

Model	1–5	1–7	1–10
Gemma-3-12B	60.60	60.84	**62.83**
OpenAI-o3	58.86	59.11	**60.10**

Table 9: Accuracy of MS-PCoT-AVG using different scoring scales for persuasion strategy evaluation.

The results (Table 9) show a consistent advantage for the 1–10 scale, with improvements of about 2 points over the 1–5 setting for both models. The 1–7 scale yielded intermediate results, but still underperformed the 1–10 variant. This suggests that the additional granularity of the 1–10 scale helps capture more nuanced judgments of persuasiveness, leading to more informative feature representations for downstream prediction.

Beyond the empirical gains, the 1–10 scale also aligns better with common practices in social science and psychology, where similar Likert-type scales (1–10 or 0–10) are widely used for subjective assessments (Vall-Llosera et al., 2020; Preston and Colman, 2000).

We therefore conclude that the 1–10 scale is the most effective and principled choice for our framework.

J Independent Scoring Prompts

Below are the exact prompts used in the four variants of the Independent Scoring baselines:

Independent Scoring

> You are a Persuasion Detector, your goal is to detect the degree of persuasiveness of a message ranging from 1 to 10, where persuasion is the potential of changing someone's opinion. You will be prompted with the message and you have to respond with ONLY a number from 1 to 10 based on the degree of persuasiveness of the message.
>
> `-- Message to evaluate: --`
>
> `text of message`

+ Context

> You are a Persuasion Detector, your goal is to detect the degree of persuasiveness of a message ranging from 1 to 10, where persuasion is the potential of changing someone's opinion. You will be prompted with the title and the body of the original poster and the message that tries to make the original poster change their view. You have to respond with ONLY a number from 1 to 10 based on the degree of persuasiveness of the message.
>
> `-- Title: --`
>
> `title of original post`
>
> `-- Body: --`
>
> `body of original post`
>
> `-- Message to evaluate: --`
>
> `text of message`

+ Explanation

> You are a Persuasion Detector, your goal is to detect the degree of persuasiveness of a message ranging from 1 to 10, where persuasion is the potential of changing someone's opinion. You will be prompted with the message and you have to respond with a number from 1 to 10 based on the degree of persuasiveness of the message, followed by a brief explanation on why you gave that score.
>
> `-- Message to evaluate: --`
>
> `text of message`

+ Context + Explanation

> You are a Persuasion Detector, your goal is to detect the degree of persuasiveness of a message ranging from 1 to 10, where persuasion is the potential of changing someone's opinion. You will be prompted with the title and body of the original post and the message. You have to respond with a number from 1 to 10 based on the degree of persuasiveness of the message, followed by a brief explanation on why you gave that score.
>
> `-- Title: --`
>
> `title of original post`
>
> `-- Body: --`
>
> `body of original post`
>
> `-- Message to evaluate: --`
>
> `text of message`

Model 1	Model 2	[D]	[S]	[MW]	[C]	[AR]	[J]
Gemma-3-12B	Gemini-2	0.460	0.492	0.469	0.379	0.475	**0.503**
Gemma-3-12B	LLaMA-3.1-8B	0.367	0.375	0.388	0.222	0.376	0.300
Gemma-3-12B	Gemini-1.5	**0.500**	**0.534**	**0.486**	0.422	**0.552**	0.388
Gemini-2	LLaMA-3.1-8B	0.338	0.325	0.366	0.256	0.297	0.265
Gemini-2	Gemini-1.5	0.388	0.465	0.464	**0.519**	0.483	0.320
LLaMA-3.1-8B	Gemini-1.5	0.294	0.344	0.336	0.359	0.362	0.170
Gemini-1.5	OpenAI-o3	0.348	0.258	0.305	0.269	0.308	0.401
Gemma-3-12B	OpenAI-o3	0.280	0.225	0.208	0.250	0.260	0.291
LLaMA-3.1-8B	OpenAI-o3	0.130	0.122	0.153	0.103	0.136	0.098
Gemini-2	OpenAI-o3	0.221	0.209	0.231	0.243	0.231	0.231

Table 10: Inter-model agreement on MS-PCoT strategy scores, measured using Cohen's κ (quadratic weights).

K Inter-model Agreement on Strategy Scores

To assess the consistency of the MS-PCoT scoring across different LLMs, we measured the inter-model agreement for the six persuasion strategies using Cohen's κ with quadratic weights. We computed these scores using the `cohen_kappa_score` function from the `scikit-learn` Python library[4].

Table 10 reports the agreement scores between each pair of models (Gemma-3-12B, Gemini-1.5-Flash-02, Gemini-2.0-Flash, LLaMA-3.1-8B and OpenAI-o3) across all strategies.

Overall, agreement levels vary across model pairs and strategies. The strongest alignment emerges between Gemma and Gemini-1.5, particularly for *Simplification* (0.534) and *Attack on Reputation* (0.552), suggesting a shared interpretation of these strategies. By contrast, OpenAI-o3 consistently shows lower agreement with all other models. This trend is especially pronounced in its comparison with LLaMA-3.1-8B, where the score for *Justification* drops to 0.098, underscoring a marked divergence in how the two models interpret rhetorical strategies.

When excluding o3 as an outlier, the remaining comparisons provide a clearer picture of overall consistency. Across these model pairs (36 values in total), 21 scores fall between 0.2 and 0.4, indicating *fair* agreement; 14 scores fall between 0.4 and 0.6, indicating *moderate* agreement; and only 1 score falls between 0.01 and 0.2, indicating *slight* agreement (Landis and Koch, 1977).

These results suggest that while architectural differences can introduce variation–most strikingly in the case of o3–the majority of models produce reasonably consistent strategy scores. This supports the reliability of MS-PCoT in capturing meaningful persuasion patterns across models, while also highlighting the need to account for outlier behaviors in inter-model agreement analyses.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

L Validation of Strategy-Based Reasoning

In order to motivate the benefit of integrating persuasion strategies into our framework, we designed a baseline experiment using a *Simple Chain-of-Thought* (CoT-Simple) setting. In this baseline, the model is asked to provide a single general analysis of persuasiveness before assigning a score, without any explicit mention of the six persuasion strategies. The scoring step is identical to the one used in MS-PCoT (see Appendix D), ensuring a controlled comparison.

Prompts for the CoT-Simple Baseline

System: You are an assistant who detects persuasion in text. Persuasive text is characterized by a specific use of language in order to influence readers. Your goal is to analyse the persuasiveness of a text.

User: Given the title and body of an original post, and a message intended to change the perspective of the original poster, critically evaluate its persuasive potential. Identify strengths and weaknesses in its argumentation, structure, and rhetorical strategies.

Following this reasoning step, the model was asked, as in MS-PCoT, to provide a numerical persuasiveness score from 1 to 10. This ensures that both approaches involve two-step prompting and produce comparable outputs, with the only difference being whether the analysis is guided by the six persuasion strategies.

Results

Table 11 reports results for our five models. Across all settings, both MS-PCoT-AVG and MS-PCoT-MLP outperform the CoT-Simple baseline.

Model	CoT-Simple	MS-PCoT-AVG	MS-PCoT-MLP
Llama-3.1-8B	53.53	60.72	61.34
Gemma-3-12B	61.21	62.83	63.69
Gemini-1.5	59.98	60.72	63.07
Gemini-2	60.10	61.83	62.70
OpenAI-o3	60.10	60.59	64.93

Table 11: Comparison between a simple chain-of-thought baseline (CoT-Simple) and our strategy-guided approach (MS-PCoT), showing the accuracy in predicting the delta-awarded message.

The consistent improvements, observed even for models of different sizes and architectures, highlight that explicitly reasoning about persuasion strategies provides a more informative and structured basis for assessing persuasiveness than relying only on generic reasoning.

M MS-PCoT-MLP Grid Search and Best Hyperparameters

For the MLP classifier in MS-PCoT-MLP, we performed a grid search over the following hyperparameter space:

- `hidden_dim`: [64, 128, 256, 512]
- `lr` (learning rate): [1e-2, 5e-3, 1e-3, 5e-4]
- `batch_size`: [32, 64, 128]
- `patience` (early stopping): [3, 5, 7, 10]
- `ema_alpha` (EMA smoothing factor): [0.1, 0.2, 0.3, 0.4]
- `lr_factor` (learning rate decay): [0.3, 0.5, 0.7]
- `lr_patience`: [2, 3, 4, 5]
- `weight_decay`: [0.0, 1e-5, 1e-4, 1e-3]

All models were trained for a maximum of 300 epochs, using early stopping on the development set. Below we report the best hyperparameter configuration found for each LLM:

LLaMA-3.1-8B

- `hidden_dim`: 128
- `lr`: 0.005
- `batch_size`: 64
- `patience`: 7
- `ema_alpha`: 0.2
- `lr_factor`: 0.4
- `lr_patience`: 2
- `weight_decay`: 0.0001

Gemini-1.5-Flash-02

- `hidden_dim`: 256
- `lr`: 0.005
- `batch_size`: 64
- `patience`: 10
- `ema_alpha`: 0.4
- `lr_factor`: 0.7
- `lr_patience`: 2
- `weight_decay`: 0.0

Gemini-2.0-Flash

- `hidden_dim`: 256
- `lr`: 0.01
- `batch_size`: 64
- `patience`: 7
- `ema_alpha`: 0.3
- `lr_factor`: 0.4
- `lr_patience`: 3
- `weight_decay`: 0.0

Gemma-3-12B

- `hidden_dim`: 128
- `lr`: 0.01
- `batch_size`: 64
- `patience`: 3
- `ema_alpha`: 0.1
- `lr_factor`: 0.5
- `lr_patience`: 2
- `weight_decay`: 0.0001

OpenAI-o3

- `hidden_dim`: 128
- `lr`: 0.005
- `batch_size`: 64
- `patience`: 7
- `ema_alpha`: 0.2
- `lr_factor`: 0.4
- `lr_patience`: 2
- `weight_decay`: 0.0001

N Strategy Score Distributions

To gain insight into the behavior of our MS-PCoT scoring system, we conducted an exploratory analysis of the output scores for the six persuasion strategies (see Figure 2). These data refer to the test set of the Winning Argument dataset, which contains 807 pairs. For each strategy, we plotted histograms that show the distribution of scores from 1 to 10 across our five models (LLaMA-3.1-8B, Gemini-1.5-Flash-02, Gemini-2.0-Flash, Gemma-3-12B and OpenAI-o3), distinguishing between successful messages (those that were awarded a delta) and unsuccessful ones.

Figure 3 shows results for all six persuasive strategies. Each histogram bar represents the number of messages that received a given score (1–10) for that strategy, split by model and success label.

Across strategies, scores tend to peak between 6 and 8, indicating that LLMs generally detect at least moderate use of persuasive framing, even in less successful messages. However, distribution patterns vary by strategy and model. For instance, a score of 10 appears only once in the entire dataset, assigned by Gemini-2 for the *Distraction* strategy, while scores of 1 are also extremely rare. Consistent with the inter-model agreement results in Appendix K, OpenAI-o3 shows a distinct pattern, characterized by systematically lower scores compared to other models. Another curious observation is that LLaMA-3.1 never assigns scores of 3 or 5. Taken together, these findings highlight that while the overall scoring trends are similar across models, each model exhibits unique biases that shape how it quantifies the use of persuasive strategies.

O Validating Strategy Scoring with MS-PCoT

To ensure that MS-PCoT produces meaningful and reliable persuasiveness scores, we validated its two-step scoring process against human annotations. Specifically, we compared MS-PCoT's outputs to gold-standard labels provided in a dataset annotated by humans for the presence of various persuasive strategies. This allows us to directly assess the alignment of our model-generated scores with human judgments.

For comparison, we also evaluated a simpler single-prompt approach on the same persuasion detection task.

O.1 Single-Prompt Classification

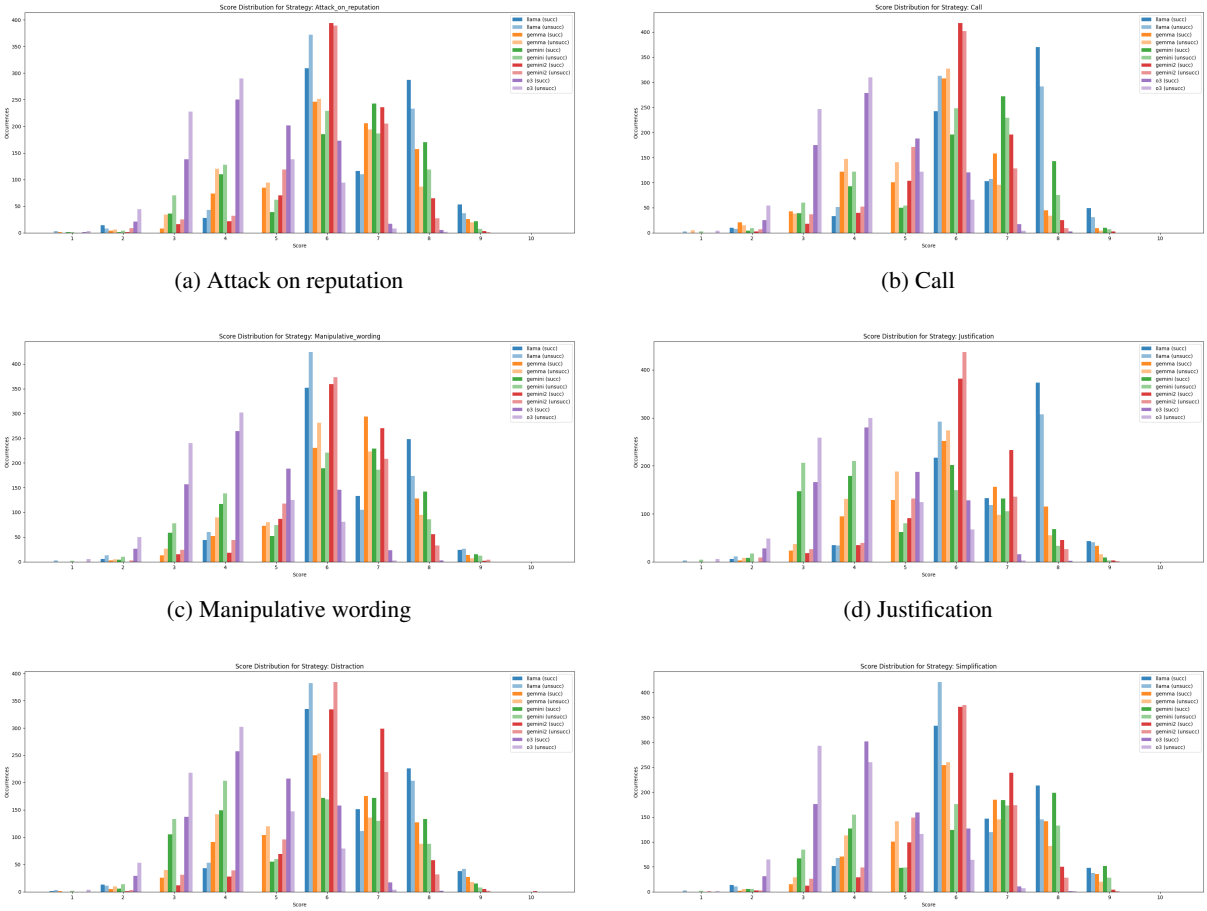
In the single-prompt setting, we used four different LLMs (LLaMA-3.1-8B, Gemma-3-12B, Gemini-1.5-Flash-02, and Gemini-2.0-Flash). We excluded OpenAI-o3 from these experiments due to budget constraints. We queried the models using a single instruction that asked to detect the presence of each strategy without providing any definitions or examples. The prompt used was:

> Analyze the text and decide if the text contains any high-level persuasion approaches from the following: Attack on reputation, Justification, Simplification, Distraction, Call, Manipulative wording. Give your answer in the form of dictionary: {
> "Attack on reputation": "Your answer. Use only Yes or No",
> "Justification": "Your answer. Use only Yes or No",
> "Simplification": "Your answer. Use only Yes or No",
> "Distraction": "Your answer. Use only Yes or No",
> "Call": "Your answer. Use only Yes or No",
> "Manipulative wording": "Your answer. Use only Yes or No"
> }. Return only the dictionary, nothing else.

O.2 MS-PCoT Scoring

Using the same four LLMs, we followed the MS-PCoT approach presented in Section 4. First, the model received a strategy definition and was prompted to produce a critical analysis of the input article in relation to that strategy. Then, based on the generated analysis, the model was asked to assign a score from 1 to 10 reflecting the likelihood that the strategy was present. This process yielded a six-dimensional persuasion score vector per article.

To convert the continuous outputs into binary predictions (present/absent), we identified a separate threshold for each strategy. These thresholds



(a) Attack on reputation

(b) Call

(c) Manipulative wording

(d) Justification

(e) Distraction

(f) Simplification

Figure 3: Distribution of MS-PCoT strategy scores (1–10) across successful and non-successful messages. Each panel shows histograms for a different strategy, grouped by LLM model.

were optimized on a held-out validation set to maximize the Micro F1 score. The selected thresholds are detailed in Table 12.

LLM	[D]	[S]	[MW]	[C]	[AR]	[J]
LLaMA-3.1-8B	8.05	7.05	4.05	7.05	6.05	6.05
Gemma-3-12B	9.05	8.05	0.00	8.05	3.05	4.05
Gemini-1.5	9.05	8.05	0.00	8.05	2.05	2.05
Gemini-2.0	9.05	7.05	0.00	8.05	2.05	2.05

Table 12: Best threshold per strategy for MS-PCoT (1–10).

The two prompt templates used for this scoring process are reported below. Each instance of `strategy_name` and `strategy_description` was replaced with the appropriate strategy label and definition.

Step 1 – Persuasion Analysis Prompt:

System:
You are an assistant who detects persuasion in text. Persuasive text is characterized by a specific use of language in order to influence readers. Your expertise and focus is on one high-level persuasion approach called `strategy_name`. This high-level persuasion approach can be described as follows: `strategy_description`. You are the expert who detects high-level persuasion approach `strategy_name`.

User:
Given a text from a news article, critically evaluate its persuasive potential. Identify strengths and weaknesses in its argumentation, structure, and rhetorical strategies. Additionally, analyze whether the text employs the high-level persuasion approach `strategy_name`. Explain how `strategy_name` appears or doesn't appear in the text. Be conservative in your assessment and if you are not fully certain that the technique is used, assume it is not present.

Step 2 – Scoring Prompt:

System:
You are a Persuasion Detector. Your goal is to detect the degree of persuasiveness of a message ranging from 1 to 10, where persuasion is the potential of changing someone's opinion. You will be prompted with the title and body of the

original article, and an analysis on the persuasion strategy. Give a score from 1 to 10 based on the degree of persuasiveness, preceded by a brief explanation. Provide your answer in the form of a dictionary:
`{"explanation": "Your answer. Brief explanation on the reasoning that you have followed.", "response": "Your answer. Give a score from 1 to 10."}`

User:
Given a news article and an analysis on persuasion strategies used in the message, respond with a number from 1 to 10 based on the degree of persuasiveness of the text, followed by a brief explanation on why you gave that score.

O.3 Results

The results in Table 13 show a clear advantage of MS-PCoT over the single-prompt baseline across all four LLMs. For every model, MS-PCoT achieves higher accuracy on nearly all strategies, with particularly strong improvements for *Distraction*, *Simplification*, and *Attack on reputation*, where single-prompt performance was relatively weaker. In terms of overall performance, the Micro F1 score consistently increases when using MS-PCoT, confirming that the two-step approach more accurately detects persuasive strategies than the simpler single-prompt method. Since all metrics are computed against gold-standard human annotations, the superior performance of MS-PCoT can be interpreted as a validation for the quality of the strategy scores it generates. In other words, the experiment confirms that MS-PCoT produces scores that align closely with human judgment, supporting the use of these scores for downstream analysis of persuasive content.

P Extension to Other Datasets

To further validate the generality of our strategy-aware framework, we extended MS-PCoT-MLP to two additional persuasion datasets: ***Anthropic*** (described in Section 3.3) and ***Persuasion for Good*** (described in Section 3.4). Both datasets differ substantially from the *Winning Arguments* setup, requiring regression instead of classification. Below, we describe the experimental setup and results in detail. We restricted this evaluation to `Gemma-3-12B` and `OpenAI-o3`, as these were the two best performing models in our main setup.

P.1 Anthropic

For each sample, we prompted the LLM with the claim and argument and asked it to generate six persuasion-strategy scores. We used the same prompts from MS-PCot for step 1 (see Appendix C) and a slightly modified prompt, tailored to this task, for step 2 (see Appendix P.3). These scores, along with their mean, variance, entropy, and the initial pre-argument rating, formed the input to our regression model. The target was the post-argument rating.

The model architecture was identical to MS-PCoT-MLP used in classification, except that the output layer consisted of a single neuron (instead of two) to produce a continuous value, and the input dimension was reduced to accept only 6 scores instead of the two series of 6 scores used for binary classification.

We compared against two baselines (full prompts can be found in Appendix P.3):

- **Baseline-1**: a single-prompt LLM prediction of the post-argument rating.
- **Baseline-2**: a refined version of Baseline-1, explicitly instructing the model that arguments may decrease as well as increase agreement.

Results are reported in Table 14. Accuracy (`Acc.`) is calculated by rounding the predicted values to the nearest integer, and then assessing if it was the same number as the gold label. MS-PCoT-MLP achieved significantly lower errors and higher R^2, demonstrating the value of strategy-aware reasoning.

Method	MSE(↓)	RMSE(↓)	MAE(↓)	R^2(↑)	Acc.(↑)
Baseline-1-Gemma	4.3553	2.0869	1.6988	-0.229	0.184
Baseline-2-Gemma	3.5854	1.8935	1.5347	-0.011	0.201
Baseline-1-o3	1.0068	1.0034	0.7394	0.716	0.374
Baseline-2-o3	0.9882	0.9941	0.7310	0.721	0.379
MS-PCoT-MLP	**0.6700**	**0.8185**	**0.6339**	**0.811**	**0.462**

Table 14: Results on the Anthropic dataset for MS-PCoT-MLP compared to two baselines.

It is important to note that the Anthropic dataset was released in September 2024. As shown in Appendix Q, the knowledge-cutoff dates for Gemma-3-12B and OpenAI-o3 precede this release, so these models could not have been trained on Anthropic. Consequently, the strong MS-PCoT results on Anthropic cannot be attributed to pretraining leakage and therefore help alleviate concerns about dataset contamination.

LLM	Method	[D]	[S]	[MW]	[C]	[AR]	[J]	Micro F1
LLaMA-3.1-8B	Single-Prompt	0.3899	0.5280	0.8545	0.5466	0.5560	0.6866	0.6733
	MS-PCoT	0.7612	0.5672	0.8974	0.5616	0.6978	0.6101	0.7084
Gemma-3-12B	Single-Prompt	0.4813	0.4608	0.9086	0.5578	0.6772	0.6343	0.6906
	MS-PCoT	0.8582	0.6269	0.9086	0.6698	0.7799	0.6623	0.7476
Gemini-1.5	Single-Prompt	0.6754	0.6287	0.7780	0.5784	0.7463	0.6119	0.6852
	MS-PCoT	0.8582	0.6213	0.9086	0.6698	0.7817	0.6996	0.7543
Gemini-2.0	Single-Prompt	0.4795	0.5634	0.8694	0.5840	0.7780	0.6922	0.7178
	MS-PCoT	0.8563	0.6269	0.9086	0.6716	0.7985	0.6679	0.7530

Table 13: Comparison between Single-Prompt and MS-PCoT approaches across 4 LLMs. We report per-strategy accuracy (abbreviations defined in Figure 2) and overall Micro F1 score.

P.2 Persuasion for Good

To avoid leakage, we excluded all persuadee turns and prompted the LLM only with the persuader turns. This prevents the model from simply reading the donation amount directly from the persuadee's response. We then prompted an LLM to produce six persuasion-strategy scores per dialogue. We used the same prompts from MS-PCot for step 1 (see Appendix C) and a slightly modified prompt, tailored to this task, for step 2 (see Appendix P.4). These scores, along with their mean, variance, entropy, and the initial pre-argument rating, formed the input to our regression model. The target label was the donation amount.

The model architecture was identical to MS-PCoT-MLP used in classification, except that the output layer consisted of a single neuron (instead of two) to produce a continuous value, and the input dimension was reduced to accept only 6 scores instead of the two series of 6 scores used for binary classification.

We compared against two baselines (full prompts can be found in Appendix P.4):

- **Baseline-1**: a single-prompt prediction of the donation amount.
- **Baseline-2**: a refined prompt discouraging overestimation of donations.

Results are reported in Table 15. Again, MS-PCoT-MLP strongly outperformed baselines, with orders of magnitude reduction in MSE.

Method	MSE(↓)	RMSE(↓)	MAE(↓)	R^2(↑)
Baseline-1-Gemma	24352.63	156.05	68.43	-1354.91
Baseline-2-Gemma	11365.33	106.61	18.47	-631.80
Baseline-1-o3	21978.06	148.25	61.02	-993.21
Baseline-2-o3	9512.10	97.53	16.18	-485.14
MS-PCoT-MLP	**21.41**	**4.63**	**2.48**	**-0.19**

Table 15: Results on the Persuasion for Good dataset for MS-PCoT-MLP compared to two baselines.

P.3 Prompts for Anthropic Experiments

We used the following prompts for the baselines and the MS-PCoT scoring procedure. After running Baseline-1, we observed that the models consistently predicted final ratings that were higher than the initial rating, regardless of the persuasiveness of the argument. To address this issue, in Baseline-2 we explicitly added the instruction that the new rating can be both higher or lower than the initial one. This modification resolved the problem and led to improved results.

Baseline-1

> Given an original claim, an argument that tries to make the reader agree with the claim, and the initial rating given by the reader, you have to respond with a number from 1 to 7 (1 means strongly disagree, 7 strongly agree) based on the new rating that you think the reader gave after reading the argument. The goal is thus to understand how much you think that the claim succesfully convinced the reader to agree with the claim. Answer with only a number from 1 to 7, nothing else.

Baseline-2

> Given an original claim, an argument that tries to make the reader agree with the claim, and the initial rating given by the reader, you have to respond with a number from 1 to 7 (1 means strongly disagree, 7 strongly agree) based on the new rating that you think the reader gave after reading the argument. The goal is thus to understand how much

API Model Name	Knowledge Cutoff Date	Access Details	License	Model Size
meta-llama/Meta-Llama-3.1-8B-Instruct	Dec 2023	DeepInfra API	Meta Llama 3 Community	8B
google/gemma-3-12b-it	Aug 2024	Hugging Face Transformers	Gemma Terms of Use	12B
gemini-1.5-flash-002	May 2024	Google API	Commercial	Not Disclosed
gemini-2.0-flash	Aug 2024	Google API	Commercial	Not Disclosed
o3-2025-04-16	Jun 2024	OpenAI API	Commercial	Not Disclosed

Table 16: Large Language Models used in our experiments.

> you think that the claim succesfully convinced the reader to agree with the claim. Consider that the new rating can be both higher or lower than the initial one. Answer with only a number from 1 to 7, nothing else.

MS-PCoT Strategy Scoring

> Given an original claim and an argument that tries to make the reader agree with the claim, as well as an analyses on persuasion strategies used in the argument, you have to respond with a number from 1 to 10 based on the degree of persuasiveness of the argument, preceeded by a brief explanation on why you gave that score. Give your answer in the form of dictionary: {"explanation": "Your answer. Brief explanation on the reasoning that you have followed.", "response": "Your answer. Give a score from 1 to 10."}

P.4 Prompts for Persuasion for Good Experiments

We used the following prompts for the baselines and the MS-PCoT scoring procedure. During initial experiments, we observed that the models often refused to answer, claiming that it was against their ethical guidelines to process dialogues where a persuader tries to convince someone to donate money. To mitigate this issue, we clarified in the prompts that the dataset was created for the good intention of donating for a charity. This addition reduced refusals and allowed the models to proceed with the task. Later, we noticed that explicitly providing the minimum and maximum possible values of donations (0 to 700, as found in the dataset) biased the models toward producing overly high predictions. To address this, we removed that sentence in Baseline-2, which improved the results.

Baseline-1

> Given a text taken from the dataset Persuasion For Good (created for the good intention of donating for a charity) where a persuader tries to convince a user to donate for a charity (user turns are hidden), you have to respond with a number indicating the dollars you think the persuadee donated. It can be a number from 0 to 700, but consider that usually people donate around 2 dollars. Answer with only the number of dollars you think were donated, nothing else.

Baseline-2

> Given a text taken from the dataset Persuasion For Good (created for the good intention of donating for a charity) where a persuader tries to convince a user to donate for a charity (user turns are hidden), you have to respond with a number indicating the dollars you think the persuadee donated. It can whatever number, but consider that usually people donate around 2 dollars. Answer with only the number of dollars you think were donated, nothing else.

MS-PCoT Strategy Scoring

> Given a text taken from the dataset Persuasion For Good (created for the good intention of donating for a charity) where a persuader tries to convince a user to donate for a charity (user turns are hidden), as well as an analyses on persuasion strategies used in the text, you have to respond with a number from 1 to 10 based on the degree of persuasiveness of the text, preceeded by a brief explanation on why you gave that score. Give your answer in the form of dictionary: {"explanation": "Your answer. Brief explanation on the reasoning that you have followed.", "response": "Your answer. Give a score from 1 to 10."}

Q LLMs Used in Experiments

In our experiments, we used a diverse set of Large Language Models to ensure broad applicability and test the robustness of our method across different architectures, sizes, and access modalities. We included both commercial API-based models and open-weight models, trying to balance accessibility and performance. In particular, we evaluated the system using Meta-Llama-3.1 (8B Instruct), Gemma-3-12B, Gemini 1.5 Flash-02, Gemini 2 Flash and OpenAI-o3. Table 16 provides details on each model's access, license, size, and knowledge cutoff.