Detecting Winning Arguments with Large Language Models and Persuasion Strategies

Anonymous ACL submission

Abstract

Detecting persuasion in text is a challenging task, with important implications for understanding human communication. In this work, we address the problem using the *Winning Arguments* dataset built from the *Change My View* subreddit, where users award a "delta" to comments that successfully change their opinion. Given a pair of similar messages where only one of which received a delta, our goal is to identify the successful one. We approach the task by leveraging large language models (LLMs) through a chain-of-thought framework that guides them to reason about six persuasion strategies that have been widely studied in the literature. Our method directs LLMs to reflect on the use of each strategy within a message and to assess its overall persuasiveness. To better understand the influence of content, we also organize the dataset into broad discussion topics and examine performance across them. Finally, we release this topic-annotated version of the dataset to support future research on persuasion detection. Our results show that LLMs, when guided through explicit reasoning steps, can effectively capture persuasive signals. This highlights the value of strategy-based prompting for improving interpretability and robustness in argument quality assessment.

1 Introduction

Persuasion is a core element of human communication, shaping public discourse, influencing opinions, and enabling constructive disagreement. Detecting persuasive language is a challenging task: persuasive messages often rely on subtle rhetorical strategies, emotional resonance, and contextual nuance rather than overt factual superiority. This complexity makes automatic persuasion detection particularly difficult for standard NLP models.

Recent advances in large language models (LLMs) have significantly expanded the potential for modeling subtle aspects of language understanding and subjective interpretation (Wu et al., 2024; Elbouanani et al., 2025). LLMs show promise in reasoning and evaluating argumentative quality, yet their application to persuasion detection raises key questions around reliability, bias, and interpretability, particularly in zero-shot settings where the model is not fine-tuned for the task.

We focus on the task of identifying persuasive strategies in text, using the *Winning Arguments* dataset (Tan et al., 2016), built from the *Change My View* subreddit. On this platform, users share opinions and others respond in an attempt to change their mind; replies that succeed are marked with a "delta" symbol (Δ). The dataset contains pairs of replies, one successful (delta-awarded) and one not, matched to be lexically and semantically similar. This setup makes distinguishing persuasive from non-persuasive messages especially challenging, underscoring the role of rhetorical and stylistic cues.

To enable topic-sensitive analyses, we introduce a topic-annotated version of the dataset, organizing discussions into four themes: (1) Food and Culture, (2) Religion and Ethical Debates, (3) Economics and Politics, and (4) Gender, Sexuality, and Minority Rights. We release this resource as the *Topics Winning Arguments* (TWA) dataset.

Our analysis centers on six persuasion strategies identified in prior work (Piskorski et al., 2023c), each involving distinctive rhetorical techniques aimed at influencing the reader: (1) Attack on reputation, (2) Justification, (3) Simplification, (4) Distraction, (5) Call, and (6) Manipulative wording. These strategies provide a structured lens for evaluating persuasive content.

Building on the Persuasion-Augmented Chain of Thought (PCoT) framework (Modzelewski et al., 2025), which introduced strategy-aware prompting for analyzing rhetorical patterns in disinformation detection, we propose **MS-PCoT** (Multi-Score

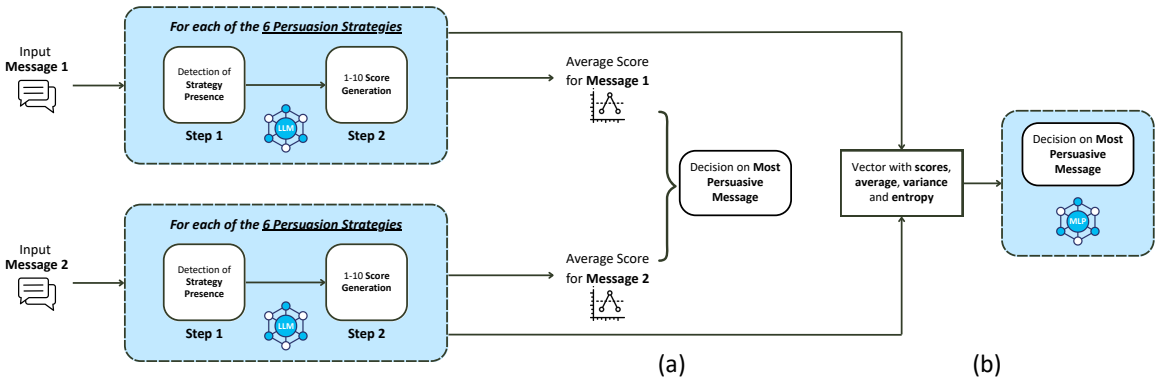


Figure 1: Overview of the MS-PCoT framework. Each of the two input messages is independently analyzed by a language model across six persuasion strategies. For each strategy, the model first generates an explanation assessing the presence of the strategy, followed by a 1–10 persuasiveness score. In the **MS-PCoT-AVG** variant (a), the most persuasive message is identified as the one with the highest 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.

Persuasion-Augmented Chain of Thought). MS-PCoT uses LLMs to perform structured, strategy-specific reasoning: for each message, it generates chain-of-thought analyses guided by the six strategies, followed by numerical persuasiveness scores (1–10) for each. These scores are used in two ways: (i) by averaging them across strategies and selecting the message with the higher mean, and (ii) by training a multilayer perceptron to predict which message is more persuasive given the generated scores.

Our main contributions are as follows:

- We release the TWA dataset, a topic-annotated extension of the Winning Arguments corpus, enabling more granular, topic-aware analysis of persuasive discourse.
- We propose a zero-shot approach that prompts LLMs using knowledge of six persuasive strategies to evaluate message persuasiveness.
- We validate the generalizability of our approach by applying it to Task 3 of the SemEval 2023 persuasion detection dataset. Our strategy-aware scores show improved correlation with human-annotated persuasiveness labels compared to a standard baseline.
- We extend our zero-shot system by training a supervised classifier over the LLM-generated persuasion strategy scores, showing that this hybrid approach further improves predictive performance.

By combining interpretable strategy-based prompting with structured downstream learning, our approach highlights the potential of LLMs not just as end-to-end predictors, but as powerful feature extractors for modeling persuasion in text.

2 Related Work

Early work on persuasion detection relied on supervised or rule-based methods with handcrafted features and discourse structures. Studies explored tactics-based modeling in blogs and dialogues (Anand et al., 2011; Young et al., 2011), affective lexicons (Guerini et al., 2008), and structural or unsupervised approaches (Walker et al., 2012; Iyer and Sycara, 2019).

Recent benchmark tasks have formalized persuasion detection. SemEval 2021 Task 6 (Dimitrov et al., 2021) focused on persuasive techniques in memes, while SemEval 2023 Task 3 (Piskorski et al., 2023b) targeted news articles with multilingual span-level annotations. These shared tasks spurred a wave of supervised models and provided standardized evaluation datasets.

Transformer-based models have since been applied to persuasion detection, especially in multilingual or multi-label contexts (Purificato et al., 2023; Hromadka et al., 2023; Roll and Graham, 2024). Some studies use LLMs to generate interpretable features or external knowledge (Li et al., 2024b), while others adopt prompting strategies with GPT models (Li et al., 2024a; Nayak and Kosseim, 2024).

Our work builds on this trend by prompting LLMs to detect specific persuasive strategies and structuring their outputs into interpretable, strategy-aware representations that support zero-shot and hybrid classification.

3 The TWA Dataset

To facilitate topic-aware analysis of persuasive arguments, we introduce the **Topics Winning Arguments (TWA)** dataset. TWA is derived from the *Winning Arguments* dataset (Tan et al., 2016), which was originally constructed from the *Change My View* subreddit: a platform where users post opinions and invite others to challenge them.[1] Each argument is annotated based on whether it successfully persuaded the original poster, signaled by the awarding of a Δ.

The *Winning Arguments* dataset was designed to highlight linguistic factors that contribute to persuasive success, rather than focusing purely on reasoning strategies. To create a balanced binary classification task, each successful argument (i.e., a rooted path-unit that received a Δ) was paired with an unsuccessful one from the same discussion, chosen for its high topical similarity. This similarity was computed using Jaccard similarity over the sets of non-stopwords in the initial replies of each path, ensuring that the content of both arguments was closely aligned in topic. The dataset further filters out non-argumentative or trivially short replies (under 50 words) and includes only discussions with sufficient engagement (at least 10 challengers and at least 3 unsuccessful arguments before the OP's last reply).

The original dataset includes 4,263 pairs, divided into 2,746 for training, 710 for validation, and 807 for testing. In constructing TWA, we preserve this split but remove one pair[2] from the validation set due to a data error where the persuasive and non-persuasive arguments were identical.

To organize the data by topic, we applied BERTopic (Grootendorst, 2022), a state-of-the-art topic modeling framework that uses transformer-based embeddings and clustering. Unlike standard settings, we customized the pipeline in three key ways. First, we cleaned the text by combining the title and body of each post and applying light preprocessing, including stop word and punctuation removal. Second, we constrained the topic model to produce exactly four interpretable clusters, ensuring consistency across samples. Third, to promote balanced topic distributions, we replaced BERTopic's default clustering with a modified KMeans algorithm that explicitly enforces similar cluster sizes. All code used for preprocessing, modeling, and topic balancing will be made publicly available.

The resulting dataset contains four 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). Additional statistics on average length and lexical diversity for each topic are reported in Appendix A.

The topic labels were assigned post-hoc based on manual inspection of the top 20 most representative tokens for each cluster, conducted collaboratively by three experts. While coarse-grained, the topics reflect distinct conversational domains commonly found in online debate.

We release this new version of the dataset to encourage future work on domain-aware argument mining and robust generalization across discussion themes.

4 Task Definition

Given a pair of messages (m_1, m_2) from the same discussion, the task is to identify which message is the persuasive one, that is, the one that successfully changed the original poster's view and received the Δ marker. The model must predict a label $y \in \{1, 2\}$, specifying whether m_1 or m_2 was the persuasive message.

Since the two messages in each pair are selected to be topically similar, the distinction between success and failure often hinges on subtle factors such as rhetorical style, emotional tone, or framing. This makes the task particularly challenging, as persuasiveness is shaped not only by content but also by how the message resonates with the original poster's perspective.

In this work, we investigate how large language models can address this task, first through direct comparison of message pairs and then through structured analyses based on persuasion strategies. We further introduce a machine learning setup where the decision is informed by features derived from strategy-aware evaluations.

[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.

5 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.

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: Attack on Reputation, Distraction, Manipulative Wording, Simplification, Justification, Call (rather than constructing an argument, this strategy directly encourages the reader to take a position or action). These strategies are derived from Piskorski et al. (2023a,c) and are described in Figure 2.

Our primary focus is the challenging task of selecting the more persuasive message between two similar ones, a setting that pushes models to make fine-grained distinctions in rhetorical effectiveness. As highlighted in the previous section, this task is particularly difficult and central to evaluating persuasive strength, which is why we adopt it as our main benchmark. However, MS-PCoT can also be naturally extended to assess the persuasiveness of a single message by applying the same protocol and averaging the six strategy-specific scores.

5.1 Step 1: Strategy-Aware Reasoning Generation

For each persuasion 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 prompt encourages careful reasoning and instructs the model to identify the strategy only when there is clear textual evidence.

This approach builds on the Persuasion-Augmented Chain of Thought (PCoT) framework (Modzelewski et al., 2025), which introduced strategy-guided prompting for identifying persuasion techniques in disinformation detection, and proposed the set of six persuasion strategies that we also adopt in our work. 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 others.

See Appendix B for full prompt details. The system message defines the strategy in detail, while the user message provides the title and body of the original post, the candidate message, and instructions for critical analysis.

5.2 Step 2: Strategy Scoring from Reasoning

After generating the reasoning, we prompt the model again to assign a persuasiveness score between 1 and 10, grounded in the explanation produced in the previous step. The prompt format includes the original post, the message, and the explanation, and asks for a numerical score preceded by a brief justification. Full prompt templates are included in Appendix C. As a result, each message is associated with 6 individual persuasion scores, one per strategy.

5.3 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 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 D.

5.4 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 certain strategies (e.g., highly emotional language) may backfire depending on the context.

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

For each message pair, we construct an 18-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 binary classifier (a multilayer perceptron, or MLP) to predict which message received the delta in the original Reddit thread. The classifier is trained using stratified 5-fold cross-validation. This learned approach enables the model to capture non-linear interactions and strategy-specific weightings that are difficult to encode with simple heuristics.

6 Experimental Setup

This section outlines the experimental setup used to evaluate the effectiveness of our proposed framework, **MS-PCoT**, described in Section 5. 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.

6.1 Language Models

We evaluated four large language models 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* and *Gemini-2.0-Flash* (accessed via API). The open-source models were run locally on high-memory GPUs, while the Gemini models were accessed through Google's API. All models were evaluated in zero-shot settings without additional fine-tuning or supervision. More details are presented in Appendix L.

6.2 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 E.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.

To quantify this effect, we evaluated the model's accuracy under three ordering conditions, placing the successful (i.e., delta-awarded) message first, second, or in a random position. As detailed in Appendix E.2 and summarized in Table 5, performance dropped significantly when the persuasive message appeared first. 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 F, this alternative did not yield better results, with model performance remaining close to random.

6.3 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 D).

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 G.

6.4 MS-PCoT Evaluation

We evaluated the two variants of our proposed **MS-PCoT** framework (AVG and MLP) on all four 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 H for details).

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.

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 I.

6.4.1 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 four LLMs detailed in Section 6.1, 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 models, suggesting that learning to combine rhetorical strategy scores yields a small but consistent performance gain.

Compared to the strongest independent scoring configuration (+ Context), MS-PCoT achieves up to a 3.59-point improvement, with Gemma-3-12B scoring 59.48% in + Context and 63.07% in MS-

Table 1: Accuracy of different prompting strategies and models in identifying the more persuasive message. Each setup is evaluated on the Winning Arguments test set.

Strategy	Model	Correct	Total	Accuracy (%)
Independent Scoring	LLaMA-3.1-8B	433	807	53.66
	Gemma-3-12B	434	807	53.78
	Gemini-1.5-Flash-02	452	807	56.01
	Gemini-2.0-Flash	453	807	56.13
+ Context	LLaMA-3.1-8B	480	807	59.48
	Gemma-3-12B	480	807	59.48
	Gemini-1.5-Flash-02	500	807	61.96
	Gemini-2.0-Flash	491	807	60.84
+ Explanation	LLaMA-3.1-8B	454	807	56.26
	Gemma-3-12B	490	807	60.72
	Gemini-1.5-Flash-02	455	807	56.38
	Gemini-2.0-Flash	477	807	59.11
+ Context + Explanation	LLaMA-3.1-8B	440	807	54.52
	Gemma-3-12B	470	807	58.24
	Gemini-1.5-Flash-02	493	807	61.09
	Gemini-2.0-Flash	496	807	61.46
MS-PCoT-AVG	LLaMA-3.1-8B	490	807	60.72
	Gemma-3-12B	507	807	62.83
	Gemini-1.5-Flash-02	490	807	60.72
	Gemini-2.0-Flash	499	807	61.83
MS-PCoT-MLP	LLaMA-3.1-8B	495	807	**61.34**
	Gemma-3-12B	514	807	**63.69**
	Gemini-1.5-Flash-02	509	807	**63.07**
	Gemini-2.0-Flash	506	807	**62.70**

PCoT-MLP, confirming that strategy-aware scoring is an effective and generalizable approach to persuasiveness evaluation.

For additional insight into how models distribute their strategy scores across persuasive and non-persuasive messages, we refer the reader to Appendix J, which visualizes these distributions for all six MS-PCoT strategies.

6.5 Validating Strategy Scoring with MS-PCoT

To assess the reliability of our two-step scoring process, we conducted an experiment aimed at validating the MS-PCoT strategy scores against 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 Appendix 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 K.1).

2. **MS-PCoT Scoring**: Following the methodology of our proposed approach, presented in Section 5, 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 K.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 10). Both approaches have been evaluated using the four LLMs presented in Section 6.1.

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 K) show that MS-PCoT consistently outperforms the single-prompt classification approach across all six strategies, indicating that the additional reasoning step leads to more

accurate and nuanced detection. These findings support the validity of our scoring methodology and confirm that the MS-PCoT-generated scores reflect meaningful assessments of persuasive content.

6.6 Topic-Based Evaluation

Building on the topical annotations provided in the TWA dataset (see Section 3), 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.

Model	Topic 1	Topic 2	Topic 3	Topic 4
MS-PCoT-AVG				
LLaMA3-8B	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-2.0	66.87%	64.94%	59.40%	58.37%
MS-PCoT-MLP				
LLaMA3-8B	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-2.0	66.27%	66.67%	58.97%	60.94%

Table 3: Accuracy of MS-PCoT-AVG and MS-PCoT-MLP across different topics for each LLM. The topic numbers correspond to the subdivision introduced in Section 3.

The results reported in Table 3 refer to the topic-specific subsets of the TWA test set, which includes 166 pairs for Topic 1, 174 for Topic 2, 234 for Topic 3, and 233 for Topic 4. While performance naturally fluctuates across topics and models, several trends emerge.

Performance is generally higher on Topic 1 (Food and Culture) and Topic 2 (Religion and Ethical Debates), while noticeably lower on Topic 3 (Economics and Politics) and Topic 4 (Gender, Sexuality, and Minority Rights). 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.

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.

7 Conclusions

In this work, we addressed the task of detecting persuasive messages from text, focusing on the challenging setup provided by the *Winning Arguments* dataset. By leveraging large language models in multiple configurations, we explored different ways to assess persuasiveness, moving from direct comparison prompts to more structured evaluations based on rhetorical strategies.

Our experiments demonstrated that prompting models to analyze specific persuasion strategies and using their outputs as structured features for a downstream classifier leads to significant improvements over simpler baselines. Additionally, our topic modeling analysis provided further insights into how the nature of the discussion impacts model performance, highlighting areas where detecting persuasion is particularly difficult.

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.

Overall, our findings suggest that while modern LLMs are capable of capturing complex persuasive signals, their performance can be further enhanced through guided analysis frameworks that focus on strategic aspects of argumentation.

Future work could explore fine-tuning strategies, expanding the set of rhetorical techniques considered, and evaluating transferability to other conversational and 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 relies heavily on the quality of the Winning Arguments dataset. Although the dataset offers a controlled setup by matching

successful and unsuccessful messages with high Jaccard similarity, it reflects the specific culture, writing style, and norms of the *Change My View* subreddit. 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 successful 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 dataset 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 I. All training was run on a single A40 GPU 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

Pranav Anand, Joseph King, Jordan L Boyd-Graber, Earl Wagner, Craig H Martell, Douglas W Oard, and Philip Resnik. 2011. Believe me-we can do this! annotating persuasive acts in blog text. In *Computational Models of Natural Argument*.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Semeval-2021 task 6: Detection of persuasion techniques in texts and images. *arXiv preprint arXiv:2105.09284*.

Akram Elbouanani, Evan Dufraisse, Aboubacar Tuo, and Adrian Popescu. 2025. Cea-list at checkthat! 2025: Evaluating llms as detectors of bias and opinion in text. *arXiv preprint arXiv:2507.07539*.

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

Marco Guerini, Carlo Strapparava, Oliviero Stock, and 1 others. 2008. Resources for persuasion. In *LREC*.

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*.

Rahul Radhakrishnan Iyer and Katia Sycara. 2019. An unsupervised domain-independent framework for automated detection of persuasion tactics in text. *arXiv preprint arXiv:1912.06745*.

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.

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. *arXiv preprint arXiv:2506.06842*.

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.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023c. Multilingual multifaceted understanding of online news in terms of genre, framing and persuasion techniques.

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).

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.

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*.

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

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 592–596.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22583–22599.

Joel Young, Craig H Martell, Pranav Anand, Pedro Ortiz, Henry Tucker Gilbert IV, and 1 others. 2011. A microtext corpus for persuasion detection in dialog. In *Analyzing Microtext*.

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 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 4: 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.

Table 4 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.

B 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.

C 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.

D 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.

E 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.

E.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.

E.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 5.

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 5: 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.

F 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 both LLaMA and Gemini models. 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 7 shows the aggregated performance across 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 7: Performance on the perturbation-based evaluation following Ziems et al. (2024).

G Independent Scoring Prompts

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

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

Table 6: Inter-model agreement on MS-PCoT strategy scores, measured using Cohen's κ (quadratic weights).

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`

H 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 6 reports the agreement scores between each pair of models (Gemma-3-12B, Gemini-1.5-Flash-02, Gemini-2.0-Flash, and LLaMA-3.1-8B) across all strategies.

Overall, agreement levels vary across model pairs and strategies. The highest agreement is observed between Gemma and Gemini-1.5, particularly for *Simplification* (0.534) and *Attack on Reputation* (0.552), suggesting strong alignment in their interpretation of these strategies. Agreement is generally lower when comparing LLaMA with the other models, especially in the case of *Justification*, where the score drops to 0.170 when compared to Gemini-1.5.

Across all model pairs and strategies (36 values in total), 21 scores fall between 0.2 and 0.4, indicating *fair* agreement; 14 scores fall between 0.4

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

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 indicate that while models can produce reasonably consistent strategy scores, some variation exists, especially between architectures. This reinforces the idea that model-specific biases may affect how rhetorical strategies are interpreted. Nonetheless, the observed agreement supports the reliability of MS-PCoT in capturing meaningful persuasion patterns across models.

I 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

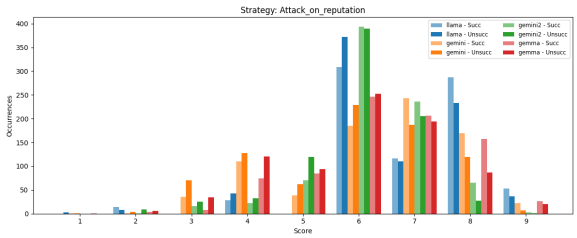
- `hidden_dim`: 256
- `lr`: 0.005
- `batch_size`: 64
- `patience`: 10
- `ema_alpha`: 0.4
- `lr_factor`: 0.7
- `lr_patience`: 2
- `weight_decay`: 0.0

Gemini-2.0-Flash

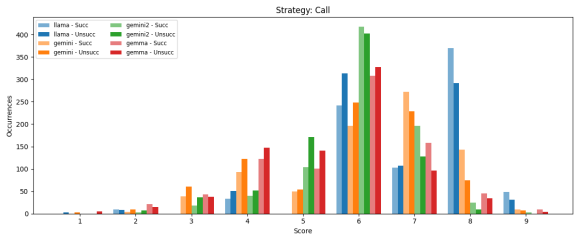
- `hidden_dim`: 256
- `lr`: 0.01
- `batch_size`: 64
- `patience`: 7
- `ema_alpha`: 0.3
- `lr_factor`: 0.4
- `lr_patience`: 3
- `weight_decay`: 0.0

Gemma-3-12B

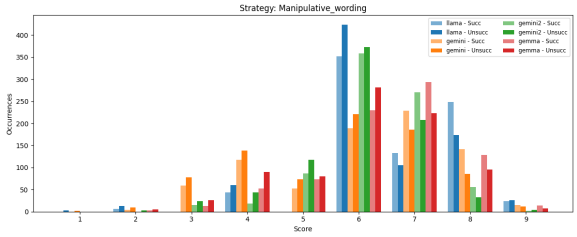
- `hidden_dim`: 128
- `lr`: 0.01
- `batch_size`: 64
- `patience`: 3
- `ema_alpha`: 0.1
- `lr_factor`: 0.5
- `lr_patience`: 2
- `weight_decay`: 0.0001



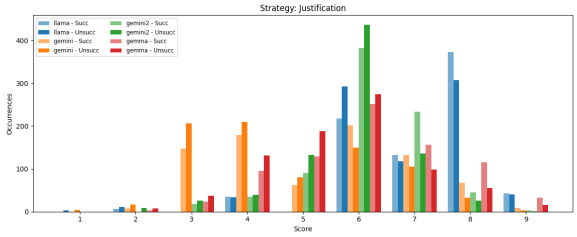
(a) Attack on reputation



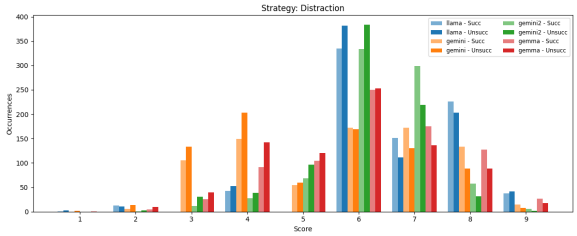
(b) Call



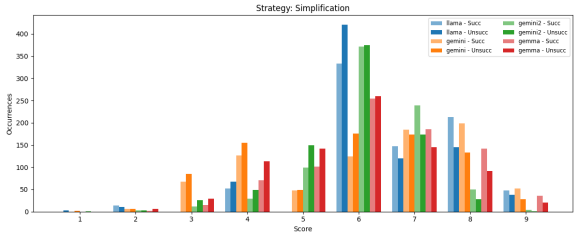
(c) Manipulative wording



(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.

J 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 the four models (LLaMA-3.1-8B, Gemini-1.5-Flash-02, Gemini-2.0-Flash, and Gemma-3-12B), distinguishing between successful messages (those that were awarded a delta) and unsuccessful ones. Notably, there is no case in which any model assigns the maximum score of 10.

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.

K Validating Strategy Scoring with MS-PCoT

To validate the effectiveness of MS-PCoT's two-step scoring process, we conducted an experiment comparing it against a simpler single-prompt approach on a persuasion detection task.

K.1 Single-Prompt Classification

In the single-prompt setting, we queried four different LLMs (LLaMA-3.1-8B, Gemma-3-12B, Gemini-1.5-Flash-02, and Gemini-2.0-Flash) using a single instruction that asked the model 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",

Table 8: 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.

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

"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.

Each model's output was parsed and evaluated against the gold annotations using various metrics. The results are presented below:

Per-strategy accuracy results (see Table 8) highlight strong performance on "Manipulative wording" and weaker detection of "Distraction" and "Simplification" across all models.

K.2 MS-PCoT Scoring

Using the same four LLMs, we followed the MS-PCoT approach presented in Section 5. 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 were optimized on a held-out validation set to maximize the Micro F1 score. Final accuracy and F1 metrics are reported in Table 8, while the selected thresholds are detailed in Table 10.

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

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

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

Table 9: Large Language Models used in our experiments.

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

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.

L 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, and the more recent Gemini 2 Flash. Table 9 provides details on each model's access, license, size, and knowledge cutoff.