A Generalizable Rhetorical Strategy Annotation Model Using LLM-based Debate Simulation and Labelling

Shiyu Ji[*1], **Farnoosh Hashemi**[*1], **Joice Chen**[1], **Juanwen Pan**[1], **Weicheng Ma**[2], **Hefan Zhang**[3], **Sophia Pan**[1], **Ming Cheng**[3], **Shubham Mohole**[1], **Saeed Hassanpour**[3], **Soroush Vosoughi**[3], and **Michael Macy**[1]

[1]Cornell University
[2]Georgia Institute of Technology
[3]Dartmouth College
Correspondence: {sj787, sh2574}@cornell.edu

Abstract

Rhetorical strategies are central to persuasive communication, from political discourse and marketing to legal argumentation. However, analysis of rhetorical strategies has been limited by reliance on human annotation, which is costly, inconsistent, difficult to scale. Their associated datasets are often limited to specific topics and strategies, posing challenges for robust model development. We propose a novel framework that leverages large language models (LLMs) to automatically generate and label synthetic debate data based on a four-part rhetorical typology (causal, empirical, emotional, moral). We fine-tune transformer-based classifiers on this LLM-labeled dataset and validate its performance against human-labeled data on this dataset and on multiple external corpora. Our model achieves high performance and strong generalization across topical domains. We illustrate two applications with the fine-tuned model: (1) the improvement in persuasiveness prediction from incorporating rhetorical strategy labels, and (2) analyzing temporal and partisan shifts in rhetorical strategies in U.S. Presidential debates (1960–2020), revealing increased use of affective over cognitive argument in U.S. Presidential debates.

1 Introduction

Persuasion is a core mechanism in social influence (O'Keeffe, 2016). It shapes how information is interpreted and acted upon across various domains, including marketing (Kumar et al., 2023), online communication (Anand et al., 2011), and political campaigns (Basave and He, 2016). In the political sphere, persuasion has become increasingly consequential amid rising polarization, growing partisan animosity, and widening ideological divides, with implications for democratic processes, public policy, and the sorting of partisan identities. (Druckman, 2022; Iyengar et al., 2019; Lelkes, 2016)

Persuasion involves rhetorical strategies that engage either cognitive and affective processes (Petty et al., 1986). Cognitive arguments appeal to reason and evidence while affective arguments persuade by arousing emotional and moral reactions. These strategies are orthogonal to veracity. For example, empirical claims, even when fabricated, can lend credibility to misleading information (Serrano-Puche, 2021), while emotional and moral appeals can go viral across social networks (Brady et al., 2017; Clifford, 2019) and intensify affective polarization by provoking indignation and reinforcing group identities (Ding et al., 2023).

The importance of rhetorical strategies in shaping consumer behavior, public discourse, and political polarization has attracted research utilizing datasets from online debates (Abbott et al., 2016), charity appeals (Wang et al., 2019), and commercial advertisements (Kumar et al., 2023).

While prior studies offer valuable insights into persuasive techniques, the diversity of theoretical perspectives has led to inconsistent categorization of rhetorical strategies across human-annotated datasets. In addition, most existing datasets are focused on specific topical domains and rhetorical strategies, making it difficult to analyze the full range of persuasive techniques or generalize across domains (Kumar et al., 2023). These datasets often lack principled topic control, which obscures the distinction between rhetorical and topic-driven effects and leads models to overfit to topic-specific patterns with limited generalizability (Chen and Yang, 2021). Most importantly, the cognitive and motivational demands of human annotation have resulted in a paucity of large-scale, high-quality datasets, and low inter-rater agreement compromises the establishment of reliable ground truth labels (Habernal and Gurevych, 2016b). These challenges have limited the development of robust deep-learning classifiers for automated identification of persuasive techniques.

[*]Both authors contributed equally to this research.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20482–20503
November 4-9, 2025

To address these challenges, we propose a novel framework that trains classifiers to detect rhetorical strategies using synthetic debate data generated and labeled by large language models (LLMs) and guided by a rhetorical typology informed by social and psychological theories of persuasion. Central to this framework is LLM labeling using simulated personas to annotate persuasive discourse. This automated annotation process enhances the reliability and scalability of rhetorical detection.

Using this dataset, we trained a rhetorical classifier and validated the labels with human annotators. We then applied the classifier to analyze temporal trends in persuasive strategies in U. S. Presidential debates from 1960 to 2020. Our analysis reveals shifting rhetorical patterns, providing new insights into the evolving landscape of partisan political communication. While our dataset generation focuses on the political domain, the framework is easily adaptable to other domains with minimal modification.

To sum up, our contribution is five-fold: 1) We present a **fully-automated scalable framework** for the generation and annotation of persuasive arguments that enhances the cross-context applicability of rhetorical labels. 2) We provide a **high-quality, topic-controlled dataset** that has been validated by human annotators. 3) We develop models to detect rhetorical strategies across varied topics and domains, with validation from human annotations and evaluation on external datasets. 4) Across five datasets from different domains, incorporating rhetorical labels into a fine-tuned BERT model improves performance in predicting persuasive outcomes, both within and across diverse datasets. 5) We identify a **significant increase in reliance on affective over cognitive strategies** during U.S. Presidential Debates going back to 1960, which may reflect the increase in affective polarization among both voters and political elites.

2 Related Work

2.1 Persuasion Strategy Identification

Prior work labeling rhetorical strategies has relied on two sources: 1) persuasive arguments collected from existing corpora (e.g. college debates), and 2) crowd-sourced annotations (Wang et al., 2019; Habernal and Gurevych, 2016a,b; Chen and Yang, 2021). These studies span multiple domains, including online conversations (Abbott et al., 2016), charity requests (Wang et al., 2019), commercial advertising (Kumar et al., 2023), and documented argumentation (Marro et al., 2022). Rhetorical labels are often derived from frameworks like Aristotle's typology of logos (logical reasoning), pathos (emotional appeals) and ethos (reference to credible sources) (Hidey et al., 2017; rhe; Stucki and Sager, 2018). For example, Higgins and Walker (2012) annotated social environment reports for logos, pathos, and ethos. Habernal and Gurevych (2016b) labeled 990 user-generated texts for logos and pathos, and Abbott et al. (2016) classified online discussion as emotion- or fact-based. Recent studies in computational linguistics have advanced automated rhetorical labeling by applying deep learning architectures to large annotated corpora. For example, Yang et al. (2019) developed a semi-supervised neural network model to classify persuasion tactics on social forums. Shaikh et al. (2020) employed autoencoders (VAE) to analyze content and rhetorical strategies in loan requests.

2.2 Automatic Debate Generation

The use of large language models (LLMs) in text generation has shown significant advantages across multiple applications, particularly in the social sciences where the ability to instantiate personas (Frisch and Giulianelli, 2024; Tseng et al., 2024) is vital for nuanced and contextually appropriate outputs (Veselovsky et al., 2023). Even early LLMs like GPT-3 perform well at producing syntactically correct and semantically coherent text (Huang et al., 2024), comparable to human-generated content (Muñoz-Ortiz et al., 2024; Dou et al., 2022), making them valuable tools for modeling social interactions and linguistic patterns (Xiao et al., 2023). LLMs are also effective for domain-specific tasks such as text generation for low-resource languages (Yang et al., 2024), where aligning with cultural and linguistic nuances is essential.

3 Rhetorical Strategies

We use a rhetorical typology that integrates Aristotle's classical framework with the dual-process distinction between cognitive and affective persuasion (Petty et al., 1986; Chaiken and Trope, 1999). Reasoning with logic and evidence involves cognitive processes, while emotional and moral arguments are affective.

Studies based on Aristotle's typology use logos inconsistently, sometimes referring to evidence and other times to logical reasoning (Egawa et al., 2019; Iyer and Sycara, 2019; Marro et al., 2022). This

can be especially problematic in annotation tasks. Moreover, while logos refers to logical reasoning, the rules of formal logic are overly narrow and difficult to operationalize for annotation. We therefore separate reasoning and evidence into two distinct strategies and focus on *causal* reasoning, in which an argument points to the positive or negative consequences of an action or event (Walton, 2012).

On the affective side, we distinguish between emotional and moral arguments. Appeals to emotion have been identified across multiple domains (Yang et al., 2019; Cabrio et al., 2018; Abbott et al., 2016) and involve the expression of evocative language to arouse emotions in the target audience (Miceli et al., 2006). Evocative language can also include moral emotions such as compassion, harm, betrayal, and degradation (Haidt, 2003; Feinberg and Willer, 2019; Anand et al., 2011). However, we classify these as moral persuasion, which is distinct from non-judgmental emotional appeals in that they refer to normative and ethical principles (Anand et al., 2011; Iyer and Sycara, 2019; Yang et al., 2019; Feinberg and Willer, 2019).

These distinctions yield the following four-fold typology (see Section A for examples and illustrations of each):

Causal - *A causal argument relies on cause-and-effect reasoning to explain or predict the positive or negative consequences of an action that are measurable or observable, with or without evidence.*

Empirical - *An empirical argument relies on evidence such as statistics, examples, illustrations, anecdotes, and/or citations to sources that support the argument.*

Emotional - *An emotional argument relies on impassioned, arousing, or provocative language to express or evoke feelings (such as frustration, fear, hope, joy, desire, sadness, hurt, and/or surprise).*

Moral - *A moral argument relies on concepts of right and wrong, justice, virtue, duty, or the greater good in order to persuade others about the ethical merit of a position, decision, or behavior.*

4 Methods

Using this typology, we developed a machine classifier for automated labeling of rhetorical strategies. Our approach consists of the five steps illustrated in Figure 1: 1) identifying controversial political topics; 2) using LLMs to generate political debate dialogues; 3) prompting LLMs to annotate the generated dialogues; 4) fine-tuned a model for strategy classification; 5) applying the fine-tuned model to downstream analytical tasks.

4.1 Opposing Stances Generation

To develop a persuasion strategy detection model for political texts, we used a combined human annotation and LLM keyword elaboration to generate diverse stances on controversial issues, ensuring balanced dialogues for robust model training. We identified controversial political topics in the United States using the Opposing Viewpoints database *Opposing Viewpoint* provided by *Gale* (Gale, a division of Cengage Learning, 2025), a trusted publisher of research content that offers diverse perspectives on contemporary social issues in the U.S. This yielded a list of 475 topical keywords (e.g., abortion, for-profit education, U.S. budget deficit). We used human annotation to refine the keyword list to those encompassing opposing viewpoints. Two annotators were tasked with answering "yes" or "no" to this question: *"Based on the provided keyword, are at least two distinct and opposing viewpoints evident in public discussions within the United States?"* Topics where both annotators answered "yes" were retained, resulting in a refined list of 146 contentious keywords.

Next, we used GPT-4o to expand each topic keyword into two broad opposing stances. These stances that were then used in our dialogue generation framework to create diverse and flexible argumentation, with broad topical coverage. This yielded 146 paired opposing arguments, associated with the 146 controversial topic keywords, which we used to generate debates. The prompt for generating paired opposing stances, along with examples, is provided in table 9 in Appendix B. Each topic was labeled as political or nonpolitical by two independent human annotators, with a third annotator resolving any disagreements, yielding 121 political and 25 non-political topics.

4.2 Controlled Debate Generation with Topic and Rhetorical Strategy Constraints

We adapted the automated debate generation framework from Ma et al. (2025) to simulate multi-turn english dialogues between two LLM agents, using the opposing stances generated from the 146 topics. Agents were prompted to either adopt or avoid one of four rhetorical strategies (causal, empirical, moral, or affective), ensuring a balanced distribution of strategies across topics and mitigating topic driven effects in downstream detection tasks.

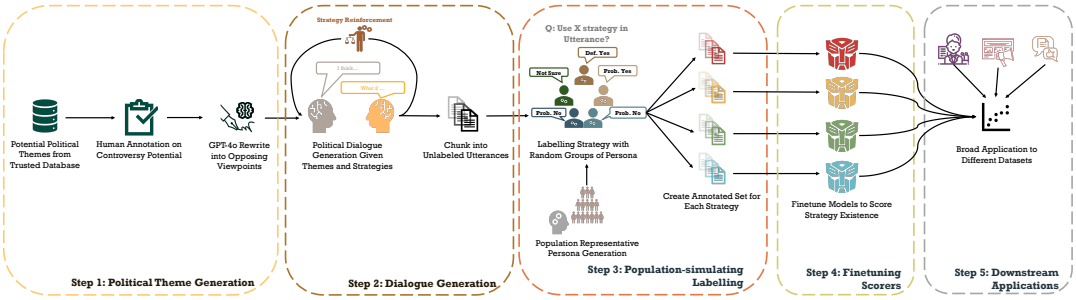

Figure 1: Overview of our proposed framework.

For each generated argument from an agent in a debate turn, a detection agent evaluated whether it aligned with the assigned strategy and prompted revisions when necessary. This detect-and-revise process could occur up to two times per argument, improving the rhetorical fidelity of generated debates. Two additional agents are employed to enhance dialogue quality. One refines individual arguments to avoid redundancy and trivial language use, and the other oversees the integrity of the generation process after each round, ensuring logical consistency within each dialogue and determining when the dialogue should conclude. (Full agent instructions are reported in Section C.) This process generated eight strategy-specific dialogues for each of the 146 controversial topics with a maximum five rounds of arguments, totaling 11,420 arguments, each with an average length of 63.4 words.

4.3 LLM-Based Persuasion Scoring

We used LLM annotation to quantify the extent to which each rhetorical strategy—causal, emotional, empirical, and moral—was exhibited in model-generated arguments, we employed large language models (LLMs) as annotators. Recent studies have demonstrated that LLMs exhibit strong alignment with human judgment in multiple domains, including clinical text summarization (Van Veen et al., 2024), moral judgment (Dillion et al., 2023), sentiment classification, political leaning detection (Bojić et al., 2025) and replicating human decision patterns in social dilemma experiments (Aher et al., 2023). Prior research suggests that prompting the model with role-specific or identity-related persona, can enhance annotation quality by encouraging more consistent and contextualized responses (El Baff et al., 2024; Bisbee et al., 2024; Argyle et al., 2023; Grundetjern et al., 2025; Kozlowski et al., 2024; Hewitt et al., 2024). Accordingly, we used five instances of GPT-4o to independently evaluate and score each argument, each from the standpoint of a different assigned persona. Each persona had a unique demographic profile based on sex, age, race, education, and partisan affiliation, with each profile aligned with the joint probabilities for the U.S. adult population, such that age, sex, and race were statistically independent while the correlations with education and political leaning reflected those in the underlying population, using data from the U.S. Census (U.S. Census Bureau, 2025), American Council on Education (American Council on Education, 2024), and Pew Research Center (Pew Research Center, 2024). The five profiles increased variability across the LLM annotators and enhanced the interpretive diversity observed among human annotators.See Appendix E for details on persona construction.

Each model, aside from its assigned persona, received the same prompt containing operational definitions of the four rhetorical strategies and two illustrations per strategy. Illustrations were drawn from Moral-Emotions (Kim et al., 2024), Ethixs (Bezou-Vrakatseli et al., 2024), and UKPConvArg (Habernal and Gurevych, 2016a), and were included only if independently labeled with full agreement by three human annotators. The prompt asked the LLM to rate each argument on a five-point Likert scale (1 = definitely not using, 5 = definitely using, 3 = uncertain) for each strategy. This yielded four scores per argument, one per strategy. For each strategy, we further averaged across five persona-conditioned LLM annotations. For downstream training, scores were linearly mapped to a 0 to 1 scale using $(x-1)/4$, where 0 = definitely not using, 1 = definitely using, and 0.5 = uncertain. The full prompt is included in section D.

To evaluate the effectiveness of the rhetorical constraints described in Section 4.2, we examined the distributions of LLM-based scores for argu-

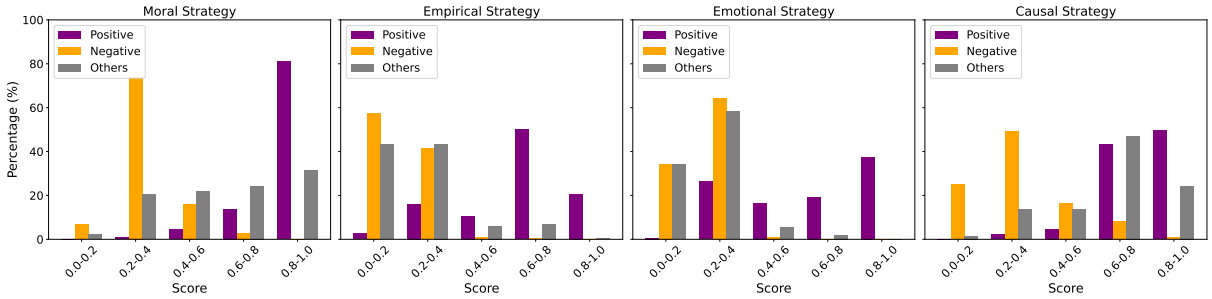

Figure 2: **Distribution of LLM-assigned strategy scores for utterances conditioned to use (Positive), avoid (Negative), or use a different rhetorical strategy (Others) for each target strategy.** Positive utterances were generated with prompts instructing the model to use the corresponding strategy; Negative utterances were prompted to avoid it; and Others includes utterances that were prompted for one of the other three strategies.

ments conditioned to use or avoid each strategy. As shown in Figure 2 , scores for all four strategies were consistently higher for positive (use) cases than for negative (avoid) ones across all four strategies. The Spearman correlations between the binary assignment (use vs. avoid) and the corresponding LLM-assigned strategy scores are reported in Table 1 , showing strong associations for moral ($\rho = 0.863$), emotional ($\rho = 0.785$), causal ($\rho = 0.812$), and empirical ($\rho = 0.805$) strategies.

The datasets were used to fine-tune models dedicated for rhetorical strategy identification for downstream application. The results are shown in Section 5.2.

	Moral	Emotional	Causal	Empirical
# of utterances	2848	2832	2862	2878
Spearman's ρ	0.863	0.785	0.812	0.805

Table 1: Number of utterances and Spearman correlation for each rhetorical strategy (all results are significant, $p < 0.0001$

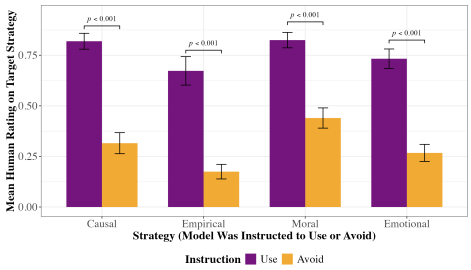

Figure 3: Human-labeled strategy scores for dialogues instructed to use vs. avoid each persuasion strategy. Scores range from 0 (definitely not using) to 1 (definitely using), with 0.5 indicating uncertainty.

5 Results

5.1 Human Validation Study

We validated the rhetorical strategy labels assigned to the LLM-generated debates through an annotation study conducted on Qualtrics, involving 355 college-educated english-speaking participants recruited via Prolific. Each participant annotated eight arguments randomly sampled from the LLM-generated debates in the test dataset of Section 5.2.2 used to evaluate the final model. They also annotated two other arguments from U.S. Presidential debates between 2000 and 2012, balanced for partisanship (to validate the downstream task in Section 6.2). For each argument, participants were asked to rate the extent to which each of the four rhetorical strategies was present, using the same Likert scale as the LLM annotation. In total, 728 arguments from the LLM-generated debates and 182 from the Presidential debate corpus were evaluated.

Prior to annotation, all participants completed a training session that explained the four rhetorical strategies, followed by a comprehension quiz to ensure that annotators understand the definition to ensure annotation quality. To improve label reliability, we used arguments annotated by at least three annotators and took the average rating (mapped into 0 to 1 as according to Section 4.3) per argument. This yielded 587 arguments with human labels from synthetic debates and 147 from Presidential debates. Results of the study are reported in Section 5.1.1 and Section 5.2.2.

5.1.1 Human Validation on LLM-generated Debate Quality and LLM-Scoring Quality

Figure 3 reports the average human scores for the target strategy, depending on whether the LLM was instructed to use versus avoid that strategy, along with t-tests for the difference between "use" and "avoid." All strategies show substantial and highly significant differences, demonstrating the effectiveness of our strategy-specific synthetic debate generation framework.

Strategy	Pretrained Model for Fine-tuning	In-Domain Evaluation		Out-Distribution Evaluation		Cross-Domain Evaluation	
		RMSE ↓	Spearman's ρ ↑	RMSE ↓	Spearman's ρ ↑	RMSE ↓	Spearman's ρ ↑
Causal	ROBERTa-base	**0.099** (0.005)	**0.870** (0.000)	**0.102** (0.000)	**0.865** (0.002)	**0.116** (0.002)	**0.850** (0.002)
	LLaMA-3.2-Instruct-3B + QLoRA	0.110 (0.001)	0.820 (0.007)	0.109 (0.003)	0.820 (0.015)	0.118(0.005)	0.808 (0.011)
Empirical	ROBERTa-base	**0.077** (0.004)	**0.931** (0.002)	**0.079** (0.003)	**0.922** (0.001)	**0.084** (0.002)	**0.913** (0.002)
	LLaMA-3.2-Instruct-3B + QLoRA	0.089 (0.003)	0.911 (0.006)	0.087 (0.003)	0.899 (0.007)	0.093 (0.002)	0.903 (0.001)
Emotional	ROBERTa-base	**0.072** (0.002)	**0.872** (0.002)	**0.073** (0.001)	**0.864** (0.002)	**0.082** (0.001)	**0.887** (0.001)
	LLaMA-3.2-Instruct-3B + QLoRA	0.083 (0.002)	0.852 (0.008)	0.079 (0.001)	0.841 (0.005)	0.091 (0.001)	0.854 (0.012)
Moral	ROBERTa-base	0.102 (0.005)	**0.939** (0.004)	0.107 (0.004)	**0.935** (0.001)	0.132 (0.004)	**0.915** (0.002)
	LLaMA-3.2-Instruct-3B + QLoRA	**0.099** (0.003)	0.932 (0.003)	**0.102** (0.003)	0.932 (0.003)	**0.117** (0.002)	0.910 (0.004)

Table 2: Transfer Learning Performance on AI-Generated Debate Data. We fine-tuned each pretrained model three times per persuasion strategy and report the mean and standard deviation on the test sets. Performance was evaluated using Spearman correlation and RMSE against LLM-based scores. RoBERTa outperformed LLaMA and showed minimal performance drop in cross-domain tests with non-politcal topics (e.g., 0.024 for moral strategy).

We validated our LLM-based persuasion scoring using the synthetic debate data from the human-annotated set. (Due to budget constraints, LLM scoring was not applied to the presidential debate data, though external corpora were used for additional validation; see section 5.3.) The LLM scoring showed strong spearman correlations with human annotations for causal ($\rho = 0.612$), empirical ($\rho = 0.622$), emotional ($\rho = 0.599$), and moral strategies ($\rho = 0.716$), all significant at $p < 0.001$.

5.1.2 Reliability and Quality of LLM Versus Human Annotation

While human annotation has long been the standard for creating ground-truth datasets, in the annotation study, we observed that large language models (LLMs) provide a more reliable and scalable alternative for rhetorical strategy annotation. We support this claim with the following three observations.

First, despite being theoretically motivated and providing richer information than binary classifications, human annotation of persuasion strategies is less reliable and requires more annotators per sample when fine-grained scales are used. In our study, we observed low inter-rater agreement (average Cohen's $\kappa = 0.148$; see Table 3) among human annotators using the five-class scheme across all rhetorical strategies, while agreement improved under coarser schemes (average Cohen's $\kappa = 0.281$ in a three-class setting, i.e., *yes* vs. *uncertain* vs. *no*, and $\kappa = 0.321$ in a binary setting, i.e., *yes* vs. *no/uncertain*; see Table 3). This suggests that much of the disagreement stems from scale granularity rather than fundamental interpretive differences, which also necessitates our approach to aggregate annotations from at least three human annotators to establish a reliable ground truth. While feasible for our study, such aggregation is costly and limits scalability.

Second, compared to individual human annotators, individual LLMs align more closely with aggregated human consensus. We assessed this by constructing a Leave-One-Out (LOO) Human Ground Truth, and comparing left-out human labels or LLM outputs against the LOO Ground Truth. Humans showed only moderate consistency with the LOO consensus (average Spearman's $\rho = 0.330$), whereas LLMs achieved substantially higher consistency ($\rho = 0.514$). As shown in Table 4, LLMs outperformed humans across all categories, indicating that a single LLM provides a closer approximation to the ground truth than a single human annotator.

Third, LLMs demonstrate greater internal consistency than human annotators. Human annotators varied widely in pairwise agreement, reflecting relatively inconsistent application of the guidelines even after the intensive training we administered. In contrast, independent LLM annotators produced more stable and coherent agreement with one another across classification schemes of varying gran-

Rhetorical Strategy	Classification Scheme		
	Five-Class (Original Scheme)	Three-Class	Two-Class
Causal	0.151	0.294	0.314
Empirical	0.141	0.290	0.334
Moral	0.146	0.287	0.324
Emotional	0.153	0.251	0.312
Average	0.148	0.281	0.321

Table 3: Human inter-rater agreement (Cohen's Kappa) across rhetorical strategies under different classification schemes. Agreement improves under coarser schemes, indicating that variability stems largely from scoring granularity.

Rhetorical Strategy	Human vs. LOO Human GT	LLM vs. LOO Human GT
Causal	0.357	0.523
Empirical	0.308	0.496
Moral	0.392	0.609
Emotional	0.264	0.427
Average	0.330	0.514

Table 4: Agreement with consensus (Spearman Correlation) between individual LLM or individual human annotators and Leave-One-Out (LOO) Human Ground Truth. LLM annotators consistently achieve higher alignment with human consensus than independent human annotators.

ularity (see Table 14 in Appendix J). This stability suggests that, under our rhetorical strategy typology, LLM annotation is more reproducible and scalable than crowd-sourced human annotation.

5.2 Fine-tuned Model Performance

We fine-tuned two pre-trained transformer models, RoBERTa-base (Liu et al., 2019) and LLaMA-3.2-3B-Instruct (Meta AI, 2024), on individual arguments from GPT-generated debates, using LLM-based strategy labels scaled from 0 to 1 in a regression setting. RoBERTa was fine-tuned on an NVIDIA A100 GPU with a learning rate of 2×10^{-5} and a batch size of 32. LLaMA was fine-tuned using 4-bit quantization and LoRA adapters (rank = 256, α = 512), with a learning rate of 4×10^{-5}, also on an A100 GPU. For all models reported in this paper, we use this same set of parameters. We evaluated the performance and topic-generalizability of the fine-tuned models in two experiments using the LLM-labeled, AI-generated dialogues.

5.2.1 Transfer Learning Experiment on AI-generated Debates

We first evaluated how well the models generalize across topics with varying levels of exposure using the arguments from the generated debates, with each topic classified as either political or non-political. Models were trained on all arguments (N=7930) from a randomly selected 101 out of 125 political topics identified in Section 4.1, using an 8/1/1 split for training, validation, and in-domain testing. We then evaluated performance on two held-out sets: (1) all arguments (N=1528) from 20 remaining political topics for out-of-distribution (OOD) testing, and (2) all arguments (N=1962) from 25 non-political topics to assess cross-domain transfer.

We fine-tuned the model independently with three random seeds on the same training set, and report the testing set performance for each of the trained models. Each test yielded two performance scores: the Spearman rank correlation between the model's predicted rhetorical strategy score and the LLM-based scores, and the RMSE for the prediction. The scores were nearly identical across the three tests, and we report the mean correlation and mean RMSE in table 2. The table reports two key findings. First, the RoBERTa model demonstrated strong predictive alignment with LLM-based scores, with exceptionally high Spearman correlations, ranging from 0.850 (cross-domain causal strategy) to 0.939 (in-domain moral strategy), and low RMSE values, ranging from 0.072 (in-domain emotional strategy) to 0.132 (cross-domain moral strategy). Second, the RoBERTa model exhibited robust transfer learning performance, with nearly identical correlations for in-domain and cross-domain evaluations, all below the 0.024 observed for the moral strategy. In sum, the results show that the two fine-tuned models are able to identify rhetorical strategies across different topics in LLM-simulated human debates.

Test Set Against	Causal	Empirical	Moral	Emotional
GPT Label	0.888	0.921	0.950	0.890
Human Annotation	0.607	0.637	0.729	0.644

Table 5: **Model testing performance on persuasion strategy labels.** Spearman rank correlations on synthetic test set with GPT-annotated labels and human annotations.

5.2.2 Final Model Performance with Human Validation

Table 1 also shows that LLaMA under-performed RoBERTa-base, which we chose for fine-tuning on the full set of LLM-generated debate data using an 8/1/1 train/validation/test split. The model's test performance is reported in Table 5. Spearman rank correlations between the model's predictions and LLM-based scores range from 0.888 to 0.950, indicating strong alignment with the synthetic annotations. To further assess external validity, we also calculated Spearman correlations on a subset of the test data annotated by human raters. These correlations ranged from 0.607 to 0.729, providing additional evidence that the model generalizes well to human-labeled data.

On the human-annotated presidential debate dataset, our model also demonstrates strong transfer learning performance, with correlations between model scores and human labels ranging from 0.567 to 0.618 (see Table 15 in Section K).

5.3 Validity Check with External Corpora

To further evaluate external validity, we tested the performance of our classifier on external datasets containing binary human annotations for rhetorical strategies that are relevant to our typology. Table 6 reports the mean difference between our model's scores and the dataset binary labels, with two-sample t-tests (see Section L for details, including the definitions of the relevant labels).

The results reveal two key patterns. First, our

model performs best on debate-like arguments with formal argumentative structures, such as those in Presidential debates, compared to less structured contexts like charity appeals or rental requests. In the debate dataset, the mean score differences range from 0.1 to 0.409 across strategies. Second, the model effectively detects rhetorical patterns associated with specific persuasion strategies, independent of the substantive content. For example, strategies like *slippery slope* and *false cause*, though fallacious, both entail causal reasoning. The model is able to distinguish these based on their argumentative form rather than the specific content of the argument, indicating the capacity to generalize across structurally similar persuasive techniques.

Strategy	Context (with Dataset Citation)	Relevant Label	Pos(1) v.s. Neg(0) Mean
Causal	Fallacious Argument in Presidential Debate (Goffredo et al., 2022)	Slippery Slope	0.409^{***}
		False Cause	0.193^{***}
	Charity Donation Requests (Wang et al., 2019)	Logical Appeal	0.047^{***}
Empirical	Charity Donation Requests (Wang et al., 2019)	Credibility	0.147^{***}
	Renting and Pizza Requests (Chen and Yang, 2021)	Evidence	0.059^{***}
	Fallacious Argument in Presidential Debate (Goffredo et al., 2022)	Appeal to Authority	0.100^{***}
Emotional	Fallacious Argument in Presidential Debate (Goffredo et al., 2022)	Appeal to Emotion	0.200^{***}
	Charity Donation Requests (Wang et al., 2019)	Personal Story	0.160^{***}
Moral	Online Petitions (Kim et al., 2024)	Moral Emotion	0.225^{***}

Table 6: **External Validity Test of the Strategy Models.** The table reports the average difference in model-predicted persuasion scores between positively labeled and other examples across external datasets.

6 Case Studies of Two Applications

Our classifier's usefulness is demonstrated in two applications: 1) improving the performance of a model for predicting the persuasiveness of an argument, and 2) measuring temporal changes in rhetorical strategies in partisan political discourse.

6.1 Persuasiveness Score Prediction

Changing someone's opinion is a common goal in contexts ranging from political and marketing campaigns to everyday interactions. This has made the study of what makes an argument persuasive a long-standing area of interest (Reardon, 1991; Habernal and Gurevych, 2016b; Wang et al., 2019; Tan et al., 2016; Toledo et al., 2019). We illustrate the usefulness of the classifier model by testing whether knowledge of an argument's rhetorical strategy can improve performance in predicting the persuasiveness of the argument. To test this, we conducted experiments across five datasets drawn from diverse topical domains, providing a broad testbed for evaluating both domain-specific and cross-domain performance. Each dataset contains arguments whose persuasiveness was assessed by human judges. The size of each dataset is shown in Table 7. A detailed description of the datasets and the evaluation of persuasiveness is provided in Appendix M.

We tested model performance in two settings: within and across topical domains, corresponding to five datasets with qualitatively different argumentation. The within domain analysis assesses performance in domain-specific contexts using an 8/1/1 train/validation/test split for each domain. We also tested the model's ability to generalize across domains with differing linguistic features. In the cross-domain setting, we fine-tuned the model on four of the five datasets and tested on the held-out fifth dataset. For each argument in each dataset, we applied the RoBERTa-based classifier trained on GPT-generated debate data to predict the four strategy scores. For both tasks, we used mean squared error to fine-tune a BERT-base-uncased model and project the resulting representation to a 128-dimensional vector. We then projected the four strategy scores into a 32-dimensional vector, concatenated this with the textual representation, and passed the combined vector through a 64-dimensional projection layer to score the persuasiveness of the argument. We measured performance using two complementary metrics, Spearman correlation between predicted and ground-truth persuasion scores and RMSE. We then compared performance between two conditions, with and without inclusion of predicted strategy scores.

Table 7 reports small but consistent improvements in predicting persuasiveness when incorporating rhetorical strategy. Within-domain, the strategy features increased the correlation with ground-truth persuasiveness scores by a relative 8.40% (absolute 0.03), with a relative 6.30% (absolute 0.014) decrease in RMSE, indicating better alignment with human judgments. In the more challenging cross-domain setting, we observe a relative 7.77% (absolute 0.024) increase in correlation and a relative 6.16% (absolute 0.015) reduction in RMSE. These improvements suggest that the strategy features not only improve prediction within a given domain but also in topical contexts other than those on which the model was trained. This case study highlights the value of using rhetorical strategies for more robust, generalizable analysis of persuasive arguments.

6.2 U.S. Presidential Debates as an indicator of Affective Polarization

The postwar increase in affective partisan polarization in the U. S. is evident not only in the voting population but also among political elites (Enders, 2021). This suggests the hypothesis that political

	ConvArg (1038)		IBM-30k (30497)		IBM-5.3k (5298)		IAC (4939)		IDEA (1205)	
	Spearman's $\rho\uparrow$	**RMSE $\downarrow$**	**Spearman's $\rho\uparrow$**	**RMSE $\downarrow$**	**Spearman's $\rho\uparrow$**	**RMSE $\downarrow$**	**Spearman's $\rho\uparrow$**	**RMSE $\downarrow$**	**Spearman's $\rho\uparrow$**	**RMSE $\downarrow$**
Within Dataset - Vanilla	0.647 (0.012)	0.265 (0.004)	0.502 (0.004)	0.176 (0.004)	0.456 (0.010)	0.204 (0.004)	0.670 (0.000)	0.188 (0.008)	0.263 (0.021)	0.280 (0.007)
Within Dataset-Strategy	**0.680** (0.009)	**0.255** (0.003)	**0.516**(0.005)	**0.167** (0.003)	**0.478** (0.009)	**0.188** (0.004)	**0.678** (0.003)	**0.171** (0.005)	**0.337** (0.036)	**0.264** (0.007)
Cross Dataset - Vanilla	0.300 (0.018)	0.335 (0.003)	0.290 (0.005)	0.247 (0.019)	0.380 (0.005)	0.345 (0.004	0.349 (0.003)	0.283 (0.004)	0.052 (0.010)	0.396 (0.005)
Cross Dataset-Strategy	**0.341** (0.016)	**0.326** (0.001)	**0.309** (0.009)	**0.218** (0.012)	**0.400** (0.005)	**0.335** (0.004)	**0.389** (0.014)	**0.257** (0.008)	**0.053** (0.009)	**0.395** (0.004)

Table 7: **Persuasiveness Score Performance.** Performance of models with and without the incorporation of rhetorical strategies, evaluated within and across datasets (higher ρ, lower RMSE are better). "Vanilla" refers to the condition without incorporation of labels for rhetorical strategies. Results are averaged over three fine-tuning runs (mean ± SD). Full results with performance differences and standard errors are reported in the Appendix.

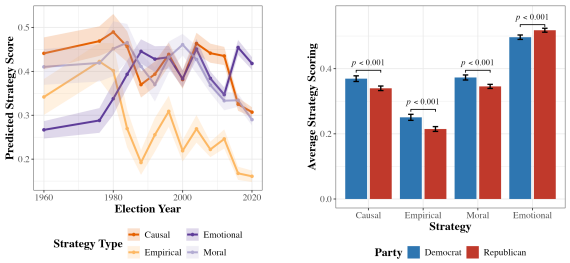

Figure 4: **Rhetorical strategies in U.S. Presidential debates.** Left: temporal trends (1960–2020). Right: partisan differences.

elites have shifted from cognitive discourse in the relatively bipartisan Eisenhower years to increasingly affective discourse today. Our model offers the opportunity to test this hypothesis by analyzing the transcripts of U. S. Presidential debates, going back to the inaugural Kennedy-Nixon debate in 1960 (Martherus, 2020). The hypothesis could also be tested using the *Congressional Record*, campaign ads, and stump speeches, but Presidential debates afford unique access to elite argumentation that targets a national audience, is focused exclusively on politically salient controversies, and follows institutional procedures that have remained relatively constant over time.

We measured temporal trends and partisan differences in rhetorical strategies at the argument level, defined as a continuous, uninterrupted string by a single speaker, with at least five words. For each argument, we applied our classifiers trained in Section 5.2.2 to predict each of the strategies. For comparability, we limited the analysis to general election candidates from the two major political parties and excluded Vice Presidential and primary debates, which differ in format and are only available for certain years. The debate corpus for analysis covers 13 U.S. presidential elections since 1960 (no debates were held in 1968 and 1972), totaling 3,307 arguments.

6.2.1 Temporal trends

Figure 4 (left) reports predicted strategy scores across Presidential debates by election year, with 95% confidence intervals. Empirical strategies (gold) show a consistent decline while emotional appeals (purple) increased, suggesting a shift from evidence-based cognitive arguments to affective rhetorical strategies. This trend is confirmed by a linear model using a single aggregated measure of cognitive (the mean of causal and empirical scores) minus affective (the mean of emotional and moral scores) on each argument. Affective scores increased relative to cognitive by 0.0025 per year ($p < 0.001$), approximately a 0.01 increase per four-year election cycle beginning in 1976.

6.2.2 Partisan differences

Figure 4 (right) reports temporally aggregated partisan differences in rhetorical strategies in Presidential de-bates. Compared to Democrats, Republican candidates relied more on emotional strategies ($\Delta = 0.021$, $p < 0.001$), and less on causal ($\Delta = 0.029$, $p < 0.001$), empirical ($\Delta = 0.036$, $p < 0.001$), and moral strategies ($\Delta = 0.027$, $p < 0.001$). However, across all elections since 1960, both Democrats ($\Delta = 0.246$, $p < 0.001$) and Republicans ($\Delta = 0.303$, $p < 0.001$) relied far more on emotional than on empirical arguments ($\Delta = 0.277$, $p < 0.001$). For election-specific results, see Section O.

7 Conclusion

Large-scale identification of rhetorical strategies has been hindered by the limitations of human annotation, including high cost, inconsistency, and limited scalability due to cognitive demands. To address this, we used a novel framework that leverages large language models to generate and annotate four rhetorical strategies in debate data. These synthetic labels are validated by simulated LLM personas and human annotators, enabling the fine-tuning of a robust rhetorical classifier that generalizes across topical domains. We demonstrate its utility in two applications: improving persuasiveness prediction and revealing the rise in affective appeals and decline in empirical arguments in U.S. Presidential debates from 1960 to 2020.

Limitations and Future Work

We note several limitations of our current study. First, we generated and evaluated data in English, which may overlook persuasion strategies that manifest differently across other languages and cultures. Future research is needed to extend our framework to multiple languages and cross-cultural comparisons, such as between the East and West and between individualist vs. collectivist societies. Second, our training data simulates only debate settings, but we can potentially improve transferability by incorporating simulations from other persuasive contexts such as advertising or fundraising. Third, we used four rhetorical strategies that were more refined than previous typologies, but future research is needed to test more fine-grained distinctions corresponding to specific emotions (e.g. indignation) or types of evidence (e.g. eye-witness or statistical). The modest improvement we observed in persuasiveness prediction may be amplified by discovering specific strategies that are uniquely effective in certain contexts.

Another limitation is the focus on persuasion, but rhetorical strategies may also influence information diffusion. Future research is needed to identify strategies that trigger virality on social media. We also did not take veracity into account. Going forward, a promising direction is to compare rhetorical strategies used in arguments that are truthful, intentionally misleading, or misinformed. For example, are affective strategies key to the manipulation and dissemination of falsehoods, with potential applications to mass persuasion processes and the spread of disinformation.

Potential Risks and Ethical Considerations

The synthetic debate dialogues generated and analyzed in this study were developed solely for research and model training purposes. While our framework offers scalability, flexibility, and high accuracy for rhetorical strategy analysis, we acknowledge the potential for misuse. As with many advances in natural language processing, similar frameworks could be repurposed by malicious actors to generate or evaluate manipulative and misleading content. However, this risk is not unique to our study and reflects broader concerns about the dangers and misuse of generative AI technologies.

Human annotation studies in this project were reviewed and approved by Cornell University's Institutional Review Board (IRB), which granted an exemption under Protocol Number IRB0149357. Annotators for rhetorical strategies were recruited through the Prolific platform, participated with informed consent, and were compensated in line with the platform's pay guidelines. No personally identifiable information was collected during the human annotation study. Participants were only associated with platform-assigned anonymous IDs used solely for payment purposes.

All external datasets used for model evaluation are publicly available to the research community. To promote transparency and facilitate future research, we will publicly release the full synthetic dataset and associated model outputs upon publication.

Acknowledgements

This work is supported in part by NSF Awards 2242073 and 2242072, by the U.S. National Library of Medicine (R01LM013833), and by a grant from the John Templeton Foundation.

References

Rhetorical strategies: Building compelling arguments. In *1st Edition: A Guide to Rhetoric, Genre, and Success in First-Year Writing*. Pressbooks@MSL.

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452, Portorož, Slovenia. European Language Resources Association (ELRA).

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

American Council on Education. 2024. Educational attainment by race and ethnicity. https://www.equityinhighered.org/indicators/u-s-population-trends-and-educational-attainment/educational-attainment-by-race-and-ethnicity/.

Pranav Anand, Joseph King, Jordan Boyd-Graber, Earl Wagner, Craig Martell, Doug Oard, and Philip Resnik. 2011. Believe me: we can do this! annotating persuasive acts in blog text. In *Proceedings of the 10th AAAI Conference on Computational Models of Natural Argument*, AAAIWS'11-10, page 11–15. AAAI Press.

Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2023. Out

of one, many: Using language models to simulate human samples. *Political Analysis*, 31:1–15.

Amparo Elizabeth Cano Basave and Yulan He. 2016. A study of the impact of persuasive argumentation in political debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1413.

Elfia Bezou-Vrakatseli, Oana Cocarascu, and Sanjay Modgil. 2024. Ethix: A dataset for argument scheme classification in ethical debates. In *27th European Conference on Artificial Intelligence (ECAI)*, pages 3628–3635.

James Bisbee, Joshua Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32:1–16.

Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaite, Vuk Vukovic, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. 2025. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific Reports*, 15:11477.

William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors. 2018. *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018: 10-12 December 2018, Torino*. Accademia University Press, Torino.

Shelly Chaiken and Yaacov Trope. 1999. *Dual-process theories in social psychology*. Guilford Press.

Jiaao Chen and Diyi Yang. 2021. Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12648–12656.

Scott Clifford. 2019. How emotional frames moralize and polarize political attitudes. *Political psychology*, 40(1):75–91.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Xiaohan Ding, Michael Horning, and Eugenia H Rho. 2023. Same words, different meanings: Semantic polarization in broadcast media language forecasts polarity in online public discourse. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 161–172.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. 2022. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274.

James N Druckman. 2022. A framework for the study of persuasion. *Annual Review of Political Science*, 25(1):65–88.

Sebastian Duerr and Peter A. Gloor. 2021. Persuasive natural language generation – a literature review. *Preprint*, arXiv:2101.05786.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2019. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 422–428, Florence, Italy. Association for Computational Linguistics.

Roxanne El Baff, Khalid Al Khatib, Milad Alshomary, Kai Konen, Benno Stein, and Henning Wachsmuth. 2024. Improving argument effectiveness across ideologies using instruction-tuned large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4604–4622, Miami, Florida, USA. Association for Computational Linguistics.

Adam M. Enders. 2021. Issues versus affect: How do elite and mass polarization compare? *The Journal of Politics*, 83(4):1872–1877.

Matthew Feinberg and Robb Willer. 2019. Moral reframing: A technique for effective and persuasive communication across political divides. *Social and Personality Psychology Compass*, 13(12).

Ivar Frisch and Mario Giulianelli. 2024. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111.

Gale, a division of Cengage Learning. 2025. Gale in context: Opposing viewpoints. `https://www.gale.com/c/in-context-opposing-viewpoints`. Accessed: 2025-09-12.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument

quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 05, pages 7805–7813.

Morten Grundetjern, Per Andersen, and Morten Goodwin. 2025. Synthetic personas: Enhancing demographic response simulation through large language models and genetic algorithms. *International Journal on Cybernetics & Informatics*, 14:21–40.

Ivan Habernal and Iryna Gurevych. 2016a. 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.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Jonathan Haidt. 2003. The moral emotions. In Richard J. Davidson, Klaus R. Scherer, and H. Hill Goldsmith, editors, *Handbook of Affective Sciences*, pages 852–870. Oxford University Press.

Luke Hewitt, Ashwini Ashokkumar, Isaias Gheza, and Robb Willer. 2024. Predicting results of social science experiments using large language models. Unpublished manuscript.

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*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Colin Higgins and Robyn Walker. 2012. Ethos, logos, pathos: Strategies of persuasion in social/environmental reports. *Accounting Forum*, 36(3):194–208.

Chieh-Yang Huang, Jing Wei, and Ting-Hao Kenneth Huang. 2024. Generating educational materials with different levels of readability using llms. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*, pages 16–22.

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. The origins and consequences of affective polarization in the united states. *Annual review of political science*, 22(1):129–146.

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

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.

Jaehong Kim, Chaeyoon Jeong, Seongchan Park, Meeyoung Cha, and Wonjae Lee. 2024. How do moral emotions shape political participation? a cross-cultural analysis of online petitions using language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16274–16289, Bangkok, Thailand. Association for Computational Linguistics.

Austin C Kozlowski, Hyunku Kwon, and James A Evans. 2024. In silico sociology: forecasting covid-19 polarization with large language models. *arXiv preprint arXiv:2407.11190*.

Yaman Kumar, Rajat Jha, Arunim Gupta, Milan Aggarwal, Aditya Garg, Tushar Malyan, Ayush Bhardwaj, Rajiv Ratn Shah, Balaji Krishnamurthy, and Changyou Chen. 2023. Persuasion strategies in advertisements. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):57–66.

Yphtach Lelkes. 2016. Mass polarization: Manifestations and measurements. *Public Opinion Quarterly*, 80(S1):392–410.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Weicheng Ma, Hefan Zhang, Ivory Yang, Shiyu Ji, Joice Chen, Farnoosh Hashemi, Shubham Mohole, Ethan Gearey, Michael Macy, Saeed Hassanpour, and et al. 2025. Communication is all you need: Persuasion dataset construction via multi-llm communication. In *Under Review*.

Santiago Marro, Elena Cabrio, and Serena Villata. 2022. Graph embeddings for argumentation quality assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4154–4164, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James Martherus. 2020. Introducing the transcripts of us presidential debates data set. *SSRN Electronic Journal*.

Meta AI. 2024. Llama 3.2: A multimodal large language model. `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`. Released Sep. 25, 2024. Accessed: 2025-09-12. LLaMA 3.2 Community License.

Maria Miceli, Fiorella de Rosis, and Isabella Poggi. 2006. Emotional and non-emotional persuasion. *Applied Artificial Intelligence*, 20(10):849–879.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265.

Brendan O'Keeffe. 2016. *Persuasion: Theory and Research*. SAGE Publications, Thousand Oaks, CA.

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4082–4088. AAAI Press.

Richard E Petty, John T Cacioppo, Richard E Petty, and John T Cacioppo. 1986. *The elaboration likelihood model of persuasion*. Springer.

Pew Research Center. 2024. Partisanship by race, ethnicity, and education. https://www.pewresearch.org/politics/2024/04/09/partisanship-by-race-ethnicity-and-education.

Kathleen Kelley Reardon. 1991. *Persuasion in Practice*, 2nd edition. SAGE Publications, Inc., Thousand Oaks, CA.

Javier Serrano-Puche. 2021. Digital disinformation and emotions: exploring the social risks of affective polarization. *International review of sociology*, 31(2):231–245.

Omar Shaikh, Jiaao Chen, Jon Saad-Falcon, Polo Chau, and Diyi Yang. 2020. Examining the ordering of rhetorical strategies in persuasive requests. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1299–1306, Online. Association for Computational Linguistics.

Edwin D Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Iris Stucki and Fritz Sager. 2018. Aristotelian framing: logos, ethos, pathos and the use of evidence in policy frames. *Policy Sciences*, 51(3):373–385.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment-new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

U.S. Census Bureau. 2025. National population by characteristics: 2020-2023. https://www.census.gov/data/tables/time-series/demo/popest/2020s-national-detail.html.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey. European Language Resources Association (ELRA).

Douglas Walton. 2012. Argument mining by applying argumentation schemes. *Studies in Logic*, 4(1):2011.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Le Xiao, Xin Shan, and Xiaolin Chen. 2023. Patterngpt: A pattern-driven framework for large language model text generation. In *Proceedings of the 2023 12th International Conference on Computing and Pattern Recognition*, pages 72–78.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. Let's make your request more

persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3620–3630, Minneapolis, Minnesota. Association for Computational Linguistics.

Ivory Yang, Weicheng Ma, and Soroush Vosoughi. 2024. N\" ushurescue: Revitalization of the endangered n\" ushu language with ai. *arXiv preprint arXiv:2412.00218*.

A Persuasion Strategies Literature

Our Typology	Related Concept	Related Definition	Source
Causal	Reason	Provides a justification for an argumentative point based on additional argumentation schemes, e.g., causal reasoning or argument absurdity.	Anand et al. (2011) Iyer and Sycara (2019)
	Reframing	Reframe issues through usage of analogy or metaphor	Duerr and Gloor (2021)
	Counter-arguments	Predict possible opposing opinions and prepare rebuttal arguments. Increase persuasiveness by addressing the audience's doubts and concerns.	Jin et al. (2024)
	Pro and Con	Provide the audience with an analysis of the pros and cons of the point of view, letting them understand why your point of view is more advantageous for them	Jin et al. (2024)
	Logos	Appeals to logical reason	Cabrio et al. (2018)
		Appeal to the rationality of the audience through logical reasoning	Hidey et al. (2017)
Empirical	Evidence	Using supporting evidence such as statistics, examples, facts	Shaikh et al. (2020)
	Concreteness	The use of facts or evidence	Yang et al. (2019)
	Logos	Appealing to the audience through reasoning or logic, by citing facts and statistics, historical and literal analogies.	Marro et al. (2022)
		Factual argumentation	Abbott et al. (2016)
Emotional	Empathy	Encourage the audience to connect with someone else's emotional state	Anand et al. (2011)
	Pathos	Persuade an audience by appealing to their emotions	Marro et al. (2022)
		Aims at putting the audience in a certain frame of mind, appealing to emotions, or more generally touching upon topics in which the audience can somehow identify	Hidey et al. (2017)
	Emotion	Have recipient feel certain emotions (guilt, anger, shame, fear, pity, feeling important, content, etc.)	Miceli et al. (2006)
		Messages with high emotional valence and arousal	Yang et al. (2019)
Moral	Deontic Appeals	Mentions duties or obligations	Anand et al. (2011)
	Moral Appeals	Mentions moral goodness/badness	Anand et al. (2011)
	Emotion	Have recipient feel certain emotions (guilt, anger, shame, fear, pity, feeling important, content, etc.)	Miceli et al. (2006)

Table 8: Overview of persuasive strategies with definitions and corresponding related work. This typology connects our four-category framework with established concepts and definitions from the literature.

B Opposing Stances Generation

B.1 Prompt for Opposing Stance Generation

Figure 5 shows the prompt used for generating opposing stances. Given a topic, we ask GPT-4o to generate broad stances both in support of and in opposition to the topic.

```
Given a topic, generate two broad and general stances—one in favor and one against. The supporting stance should advocate for the topic, while the opposing stance should argue against it. Both stances must be self-contained, very broad, and concise, without any reasoning or explanation.
Format the output as follows:
Support: [support argument]. Oppose: [oppose argument].
Do not include any other text in your response. The topic is: {topic}
```

Figure 5: Prompt for opposing stance generation

B.2 Examples of Topics and Their Generated Stances

Table 9 illustrates five examples of controversial political topics, each accompanied by two opposing stances generated by GPT-4o. These stances were then used in our debate generation framework to produce persuasive arguments.

Topic	Stance 1	Stance 2
Abortion: Late-Term	We should allow late-term abortion.	We should prohibit late-term abortion.
Marijuana	We should legalize marijuana.	We should not legalize marijuana.
Race Relations	We should prioritize improving race relations to create a more inclusive society.	We should not prioritize race relations above other pressing societal issues.
Voter Registration	We should make voter registration automatic.	We should not make voter registration automatic.
Universal Health Care	We should implement universal health care.	We should not implement universal health care.

Table 9: Five examples of controversial topics and their generated opposing stances.

C Prompts to Generate Debates

C.1 Utterance Generation Prompts

We employed distinct prompts to instruct the agents either to adopt a specific persuasion strategy or to avoid it. The prompts used for generating debate utterances under these two conditions are illustrated in fig. 6 and fig. 7, respectively. For simplicity, only the modified sections of the prompt—highlighted in blue—are shown; these reflect changes made to the original framework proposed by Ma et al. (2025). The unmodified portions (shown in grey), which follow the original prompt structure, are omitted for brevity.

```
We're doing an experiment to see if AI can persuade others. Assume that the conversation happens in online political debate. Your stance is <stance>. Use <strategy> argument to persuade the other model, and maintain your stance and DO NOT be easily persuaded by the <role>. <strategy definition>. <To Persuader/To Persuadee>. Behave like a real human and DO NOT always explicitly show agreements use the same modal to the other model's generations. Maintain a respectful tone but avoid overuse of polite phrases. Reduce redundant phrases like 'Hey', 'I understand', 'I appreciate', 'I respect' or 'thank you'. Please ensure that you use <strategy> argument and generate complete sentences, keeping the length within 100 tokens.

To Persuader:
  Assuming you both already know the topic you're going to talking about, and you don't have to repeat the proposition. Try to start the topic smoothly. For example, ...

To Persuadee:
  Try to give some explanation of your position or even convince the other person of your position, rather than simply restate your position. Try to start the topic smoothly. For example, ...
```

Figure 6: Prompt for generating debate utterance with a certain strategy.

```
We're doing an experiment to see if AI can persuade others. Assume that the conversation happens in online political debate. Your stance is <stance>. Try to persuade the other model, but DO NOT use {strategy} arguments AT ALL, and maintain your stance and DO NOT be easily persuaded by the <role>. <strategy definition>. <To Persuader/To Persuadee>. Behave like a real human and DO NOT always explicitly show agreements use the same modal to the other model's generations. Maintain a respectful tone but avoid overuse of polite phrases. Reduce redundant phrases like 'Hey', 'I understand', 'I appreciate', 'I respect' or 'thank you'. Please ensure that the content you generate consists of complete sentences and is within 100 tokens in length. Again, under NO circumstances should you use <strategy> arguments in any explicit or implicit form.

To Persuader:
  Assuming you both already know the topic you're going to talking about, and you don't have to repeat the proposition. Try to start the topic smoothly. For example, ...

To Persuadee:
  Try to give some explanation of your position or even convince the other person of your position, rather than simply restate your position. Try to start the topic smoothly. For example, ...
```

Figure 7: Prompt for generating debate utterance without a certain strategy.

C.2 Strategy Refinement Prompt

There are two steps involved in the strategy refinement. The first prompt evaluates whether the generated utterance follows the assigned persuasion strategy, and the second prompt instructs the model to revise the utterance if it fails to meet the strategy condition. The whole detect-and-revise process is repeated at most 2 times for each individual utterance. The prompts are shown in fig. 8.

```
Judge:
  We're doing an experiment to see if AI can be persuaded. The Persuader is <persuader stance> while the Persuadee is trying to keep its stance that <persuadee stance>. <strategy definition>. Please tell me whether the <role> uses <strategy> argument in the current dialogue. Answer 1 if it does or 0 if not. Don't include any other words.

Revise to Include a Strategy:
  Please revise the <role>'s current response to incorporate the <strategy> argument. Keep in mind that <strategy definition>. Argument <strategy example> and argument <strategy example> are both illustrative examples that use <strategy> argument. You are free to adapt, transform, or create entirely new arguments in the response to align with the <strategy> approach. You do not need to be fully faithful to the original generation, as long as you remain consistent with the stance in the original generation. Please ensure that you use <strategy> argument and generate complete sentences, keeping the length within 100 tokens.

Revise to Avoid an Strategy:
  Please revise the <role>'s current response while avoiding the use of the <strategy> argument. Keep in mind that <strategy definition>. You are free to adapt, transform, or create entirely new arguments in the response to align with the <strategy> approach. You do not need to be fully faithful to the original generation, as long as you remain consistent with the stance in the original generation. Please ensure that the content you generate consists of complete sentences and is within 100 tokens in length.
```

Figure 8: Prompts used in the detect-and-revise pipeline.

C.3 Local Utterance Refinement

After the utterance has been revised for certain strategy, the model was evaluated for redundancy. And the prompt is shown in fig. 9

```
Sometimes the dialogue generated by GPT contains many meaningless polite phrases such as "I understand," "I appreciate," "I respect," "Thank you," etc. Please identify if the input sentences contain such polite phrases, and if they do, remove them. If not, return the original sentences. Directly return the refined sentences without any other explanations. <style>. For example, ...

If asked to use a strategy:
  Please ensure that you use <strategy> argument and generate complete sentences, keeping the length within 100 tokens.

If asked to avoid a strategy
  Please ensure that the content you generate consists of complete sentences and is within 100 tokens in length.
```

Figure 9: Prompts used to revise the individual utterance to eliminate redundancy.

Strategy	Examples
Causal	Allowing prisoners to choose death reduces public pressure to improve the prison system.
	Mandatory vaccination could result in rich countries hoarding vaccines for their population. This could make vaccines inaccessible or unaffordable for poorer countries.
Empirical	The issue of animal extinction could be largely fixed with lab-grown meat. US consultancy firm Kearney suggests that 35% of all meat consumed globally will be cell-based by 2040.
	Research has estimated that many death row inmates were wrongly convicted and could have been exonerated.
Moral	It is a duty of the state to protect its citizens from life-threatening diseases such as COVID-19.
	It's unfair that families of prisoners can't see prisoners; it's also unfair how they're more at risk from COVID-19.
Emotional	Gay marriage is a lifestyle choice. It may be considered 'unnatural', but that is between that person and his/her love interest. Love is all some people have... You can't take that one given right away because it makes you uncomfortable. They want acceptance and understanding. Let them be happy or just ignore it. You don't choose to be gay either. Who would choose to live that way? They are constantly being harassed and can't be with their loved one. It's unfortunate and cruel. Please be respectful of them. They have done nothing wrong, God created them that way.
	These players are earning disgusting weekly salaries and the NHS is on its knees and the staff are putting their lives at risk whilst the footballers stay at home drinking Molt!

Table 10: Two Examples of Each Strategy Used in GPT-4o and Human Annotations.

C.4 Round-level Refinement

The two utterances in each debate round will be evaluated for topic consistency and repetition. If the round of utterances goes off topic or demonstrate strong repetition with the previous round, we instruct the model to re-do the generation for this round. In addition, we detect whether the two agents have reached consensus, and we stop the generation once a consensus is reached. Prompts to judge on these factors are shown in fig. 10

```
We're doing an experiment to see if AI can be persuaded. The Persuader is
<persuader stance> while the Persuadee is trying to keep its stance that
<persuadee stance>. <mode>.

Topic Mode:
 Please judge whether the conversation goes off the topic. Answer 'Yes' or 'No'.

Agreement Mode:
 Please judge whether the two sides of the dialogue have reached an agreement toward
 <topic>. Answer 'Yes' or 'No'.

Repetition Mode:
 Please judge whether the content of the current round repeats with the content of
 the previous round. Answer 'Yes' or 'No'.
```

Figure 10: Prompts used to evaluate topic consistency, repetition and agent agreement for each round of arguments to improve generation quality.

D Annotation Prompt

Figure 11 presents the prompt used for strategy labeling by GPT-4o agents adopting different personas. The definitions of each strategy correspond to those provided in section 3. We used two examples for each strategy annotated, which are provided in Table 10.

```
You are a <age> <ethnicity> <gender> with <education> education and <political leaning>
views. Your task is to label each argument as to what extent it is one of the following types:
(1) empirical, (2) causal, (3) emotional, or (4) moral.
Empirical: <definition of empirical argument>
Examples for empirical: <2 examples for empirical argument>
Causal: <definition of causal argument>
Examples for causal: <2 examples for causal argument>
Emotional: <definition of emotional argument>
Examples for emotional: <2 examples for emotional argument>
Moral: <definition of moral argument>
Examples for moral: <2 examples for moral argument>
The argument to label is: {argument}
For each question, only return the answer as one of these options (a, b, c, d, or e).
1) Is this argument empirical?
a) Definitely not
b) Probably not
c) Might or might not
d) Probably yes
e) Definitely yes
2) Is this argument causal?
a) Definitely not
b) Probably not
c) Might or might not
d) Probably yes
e) Definitely yes
3) Is this argument emotional?
a) Definitely not
b) Probably not
c) Might or might not
d) Probably yes
e) Definitely yes
4) Is this argument moral?
a) Definitely not
b) Probably not
c) Might or might not
d) Probably yes
e) Definitely yes
Your final output should follow the format: 1.x 2.y 3.z 4.w, where x, y, z, and w are the
selected answers (a, b, c, d, or e) for each respective question. DO NOT include
any explanations.
```

Figure 11: Prompt used for strategy annotation by GPT-4o

E Details of Persona Construction

Table 11 shows the persona attributes and possible values for each group, which are sampled based on U.S. Census (U.S. Census Bureau, 2025), American Council on Education Statistics (American Council on Education, 2024), and Pew Research Center (Pew Research Center, 2024). To simulate a more realistic online user population, we excluded individuals under age 15 and over 89.

Persona Attribute	Possible Values
Gender	Male, Female
Age Group	<5, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 60–64, 65–69, 70–74, 75–79, 80–84, ≥ 85
Race	Black, White, Asian, AIAN (American Indian and Alaska Native), NHPI (Native Hawaiian and Pacific Islander)
Level of Education	Less than High School, High School Graduate, Some College but No Degree, Associate Degree, Bachelor's Degree, Master's Degree, Professional Degree, Doctoral Degree
Political Leaning	Democrat, Republican, and Independent

Table 11: Persona attribute categories and value sets used for LLM simulation

F Examples of Generated Arguments

Strategy	Topic	Instruction	Generated Utterance
Causal	We should maintain capitalism as our economic system.	Use Strategy	While some companies may prioritize profit over other considerations in a capitalist system, regulations and oversight mechanisms can effectively address these issues. Government intervention ensures that businesses operate ethically, protecting worker rights and promoting sustainable practices. Competition within capitalism can drive companies to innovate and improve their social and environmental impact, leading to adaptation and progress that benefit society and the environment in the long run.
		Avoid Strategy	However, capitalism remains a predominant economic system globally. It emphasizes individualism and personal initiative, which are heavily valued in many societies. Competition is a fundamental aspect of capitalism that many find appealing. These are deeply ingrained societal values.
	We should not support WikiLeaks.	Use Strategy	While transparency and accountability are crucial in a functioning society, supporting WikiLeaks may not always be the best approach. By releasing classified information without proper context or consideration of potential consequences, WikiLeaks can jeopardize national security and put individuals at risk. The indiscriminate release of sensitive data can harm diplomatic relations, compromise ongoing operations, and endanger lives. It is important to promote transparency through legal and ethical means that uphold the safety and well-being of all individuals involved.
		Avoid Strategy	Just because an organization is well-known doesn't mean their actions are always justified or ethical. Reputation alone should not be the sole basis for supporting an entity. There are other factors to consider beyond just fame or notoriety.
Empirical	We should embrace alternative medicine alongside conventional treatments.	Use Strategy	Integrating alternative medicine alongside conventional treatments can provide patients with a more comprehensive care plan. Research has shown that certain alternative therapies, such as acupuncture and herbal medicine, can effectively complement traditional medicine in managing various conditions. Would you be open to considering the benefits of embracing alternative medicine alongside conventional treatments?
		Avoid Strategy	However, integrating alternative therapies doesn't necessarily mean abandoning traditional practices. Instead, it can enhance the overall efficacy of treatment by addressing aspects that conventional methods may not cover. It's like adding different colors to a painting to create a richer, more vibrant masterpiece.
	We should not prioritize the expansion of civil liberties.	Use Strategy	While it's important to consider civil liberties, there is empirical evidence suggesting that certain restrictions can benefit the populace in crisis situations. For instance, during the COVID-19 pandemic, a study published in The Lancet found that mask mandates led to a significant decrease in infection rates. Similarly, during the 2011 London riots, temporary curfews were imposed, leading to a 70% reduction in crime rates in affected areas.
		Avoid Strategy	I believe that prioritizing the expansion of civil liberties may not always lead to the best outcomes for society as a whole. Sometimes, certain restrictions are necessary to maintain order and stability. By focusing on other aspects of society, we can ensure a more balanced and harmonious environment for everyone.
Emotional	We should not support the widespread adoption of organic food.	Use Strategy	Let's not forget the heart-wrenching reality of organic farming. Imagine a world where organic farming is the norm, but the yield is significantly lower. Picture families, single parents, and the elderly, struggling to afford the higher prices of organic food. Imagine the fear and desperation of a mother, unable to provide for her children because the food is just too expensive. The dream of organic food for all is beautiful, but the potential nightmares it could bring.
		Avoid Strategy	However, I believe that the widespread adoption of organic food may not be the most practical solution for addressing sustainability in food production. Conventional farming methods can also implement techniques to improve soil health and reduce pollution without solely relying on organic practices. Additionally, the lower yields associated with organic farming could potentially lead to food shortages and higher prices, which may not be feasible for meeting the global food demand.
	We should not allow hunting altogether.	Use Strategy	Imagine a deer, wasting away, its ribs showing through its thinning coat, struggling for every meager mouthful of food in the dead of winter. Consider the spread of disease that could wipe out entire herds, entire species. It's a heart-wrenching image, isn't it?
		Avoid Strategy	While regulated hunting might be a solution to overpopulation problems, it's important to consider its potential risks and ethical issues. The impact of hunting on individual animals and species populations is significant, and non-lethal methods and ecosystem management practices could be a more sustainable approach to these challenges.
Moral	We should not implement censorship.	Use Strategy	I believe that censorship should not be implemented, even in the name of finding a middle ground between safety and freedom of expression. While it is crucial to protect individuals from harm, censorship can lead to the suppression of diverse perspectives and ideas. It is essential to trust in the ability of individuals to critically evaluate information and make their own decisions. Upholding the fundamental right to free speech is crucial for promoting a society that values open dialogue, creativity, and progress.
		Avoid Strategy	While the idea of censorship might seem like an efficient solution to control misinformation, its effectiveness can be questionable. The risk of unintentionally censoring valid content or diverse opinions is high. Instead, it would be more practical to invest in enhancing media literacy, developing critical thinking skills, and establishing reliable fact-checking systems. These can help individuals make informed judgments about the credibility of the information they come across.
	We should support the Fat Acceptance Movement.	Use Strategy	While promoting healthy choices is important, it's also crucial to recognize that individuals come in all shapes and sizes. The Fat Acceptance Movement focuses on ending weight-based discrimination and promoting respect for people of all body types. By supporting this movement, we are advocating for the rights and dignity of individuals who have long been marginalized and stigmatized based on their weight. It's about promoting inclusivity, understanding, and compassion for everyone, regardless of their size.
		Avoid Strategy	Let's consider this. The Fat Acceptance Movement is not encouraging unhealthy habits. Instead, it's highlighting that health isn't solely determined by weight. It's suggesting that we must consider various factors when judging healthiness, not just body size. This approach proposes a more comprehensive view of health, indicating that people can be healthy at different sizes.

Table 12: Examples of generated utterances across four persuasion strategies (Causal, Empirical, Emotional, Moral) under different instruction conditions (Use vs. Avoid). Each entry includes the topic and the model's response based on the strategic prompt.

G GPT Annotation Scores

To evaluate the effectiveness of the rhetorical constraints described in Section 4.2, we conducted a comprehensive analysis of how well the generated utterances reflected the intended persuasive strategies. Specifically, we examined the distributions of LLM-generated scores for utterances that were explicitly conditioned to either use (positive) or avoid (negative) each of the four strategies: causal, empirical, moral, and emotional. These scores were produced by five persona-conditioned GPT-4o annotators, each independently rating the presence of each rhetorical strategy on a five-point Likert scale.

Figure 2 visualizes the resulting distributions, showing that utterances generated under the "use" condition consistently received higher scores than those under the "avoid" condition for every strategy.

To quantify this effect more precisely, we calculated the Spearman rank correlation between the binary assignment (use vs. avoid) and the corresponding averaged LLM strategy scores. As shown in Table 1 , we found strong positive correlations for all four rhetorical strategies: $\rho = 0.863$ for *moral*, $\rho = 0.785$ for *emotional*, $\rho = 0.812$ for *causal*, and $\rho = 0.805$ for *empirical*. These results demonstrate that the generation system effectively controlled for rhetorical style, and that the use of rhetorical constraints yielded outputs that aligned closely with the intended persuasive strategies, as judged by independently simulated LLM annotators.

H Human annotation

For each argument, the participant sees four separate prompts along with a reminder of the definition of the strategy at question:

1. Is this argument empirical?

   Here again is the definition of empirical:

   An empirical argument relies on evidence such as statistics, examples, illustrations, anecdotes, and/or citations to sources that support the argument.

2. Is this argument causal?

   Here again is the definition of logical:

   A causal argument relies on cause-and-effect reasoning to explain or predict the positive or negative consequences of an action that are measurable or observable, with or without evidence.

3. Is this argument emotional?

   Here again is the definition of emotional:

   An emotional argument relies on impassioned, arousing, or provocative language to express or evoke feelings (such as frustration, fear, hope, joy, desire, sadness, hurt, and/or surprise), rather than relying on rational or moral appeals.

4. Is this argument moral?

   Here again is the definition of moral:

   A moral argument relies on concepts of right and wrong, justice, virtue, duty, or the greater good in order to persuade others about the ethical merit of an action.

We provide five options to choose for each prompt:

1. Definitely not
2. Probably not
3. Might or might not
4. Probably yes
5. Definitely yes

I Annotator Training Procedure

Before human participants begin their annotation task, they are asked to take a quiz. In the quiz, they are introduced to the definition and two examples of every strategy, presented with two arguments, one using the strategy and the other not, and asked to label them as either using the strategy or not. The examples and quiz arguments are drawn from the same samples as the LLM-based labeling.

If a participant labels every quiz argument correctly, they proceed to start the annotation task. Otherwise, they are redirected to a second round of quiz, which repeats the procedure above, with different examples and quiz arguments.

We provide two chances for each participant to label all arguments in a quiz correctly. If they fail to do so by the end of the second quiz, they are automatically directed to exit the survey.

Strategy Type	Evaluation Dataset	Relevant Label	Label Definition
Causal	Fallicious Argument Classification (Goffredo et al., 2022)	Slippery Slope	It suggests that an unlikely exaggerated outcome may follow an act. The intermediate premises are usually omitted and a starting premise is usually used as the first step leading to an exaggerated claim.
		False Cause	The misinterpretation of the correlation of two events for causation (**?**).
	Persuasion For Good (Wang et al., 2019)	Logical Appeal	The use of reasoning and evidence to convince others. For instance, a persuader can convince a persuadee that the donation will make a tangible positive impact for children using reasons and facts.
Empirical	Persuasion For Good (Wang et al., 2019)	Credibility	Use of credentials and citing organizational impacts to establish credibility and earn the persuadee's trust. The information usually comes from an objective source (e.g., the organization's website or other well-established websites).
	Good Faith Textual Requests (Chen and Yang, 2021)	Evidence	Providing concrete facts or evidence for the narrative or request.
	Fallicious Argument Classification (Goffredo et al., 2022)	Appeal to Authority	When the arguer mentions the name of an authority or a group of people who agreed with her claim either without providing any relevant evidence, or by mentioning popular non-experts, or the acceptance of the claim by the majority.
Emotional	Fallicious Argument Classification (Goffredo et al., 2022)	Appeal to Emotion	The unessential loading of the argument with emotional language to exploit the audience emotional instinct.
	Persuasion For Good (Wang et al., 2019)	Personal Story	Using narrative exemplars to illustrate someone's donation experiences or the beneficiaries' positive outcomes, which can motivate others to follow the actions.
Moral	Moral Emotion Dataset (Kim et al., 2024)	Moral Emotion (Existence of any of the four emotional strategy labels by majority vote)	**Other-condemning**: Condemn others (e.g., anger, contempt, disgust) **Other-praising**: Praise others (e.g., admiration, gratitude, awe) **Other-suffering**: Empathy for the suffering of others (e.g., compassion, sympathy) **Self-conscious**: Negatively evaluate oneself (e.g., shame, guilt, embarrassment)

Table 13: **Label Definitions for External Evaluation.** This table describes the relevant rhetorical or logical labels associated with each strategy type and dataset used in external validation. Due to dataset variation, only approximate matches to our strategy dimensions are used.

J Inter-rater Consistency: LLM

Rhetorical Strategy	Classification Scheme		
	Five-Class (Original Scheme)	Three-Class	Two-Class
Causal	0.458	0.665	0.749
Empirical	0.595	0.672	0.793
Moral	0.566	0.692	0.829
Emotional	0.546	0.691	0.822
Average	0.541	0.680	0.798

Table 14: **Inter-rater consistency (Cohen's Kappa) of LLM annotators under the five-class, three-class and two-class classification scheme.** LLMs consistently demonstrate substantially higher internal reliability than human annotators.

K Presidential Debate Human Validation

Strategy	Spearman's ρ (Model vs. Human)
Causal	0.618
Empirical	0.614
Moral	0.618
Emotional	0.567

Table 15: Comparison of RoBERTa model predictions with human-annotated strategy scores on the presidential debate dataset. The values indicate Spearman's rank correlation (ρ) between model predictions and average human annotations.

L Label Definitions for External Validation Datasets

The label definitions from external datasets used in our external validity tests, which are relevant to the rhetorical typology in our experiment, are presented in Table 13.

M Persuasiveness Score Datasets

ConvArg (Habernal and Gurevych, 2016a): The ConvArg dataset contains 9,111 argument pairs from the online debate platforms CreateDebate and ConvinceMe. Each argument pair is annotated via crowdsourcing, where human annotators indicate which argument is more convincing through a binary judgment and justify their choice by selecting from a set of predefined reasons, including strength of reasoning, emotional appeal, relevance to the topic, and language quality. We compute the persuasiveness score for each argument based on pairwise argument quality, using the PageRank (Simpson and Gurevych, 2018) or winning rate as $Score = \frac{\#Win}{\#Win+\#Loss}$ where $\#Win$ denotes the number of times an argument is labeled more persuasive, and $\#Loss$ the number of times it is deemed less persuasive. This method mirrors the one proposed by Gretz et al. (2020).

IBM_30k Gretz et al. (2020): The IBM-Rank-30k dataset contains 30,497 crowd-sourced arguments on 71 controversial topics, collected via the Figure Eight platform. For each topic, annotators were asked to write two short arguments—one supporting and one opposing the topic—as if preparing for a public speech. Each argument was then evaluated by 10 annotators, who were asked whether they would recommend the argument to a friend preparing a speech, regardless of their personal stance on the issue. To derive a continuous quality score from these binary responses, the dataset employs a Weighted Average scoring function, which adjusts each annotator's influence based on their annotator reliability score—a measure of how consistently the annotator agrees with others across previous shared tasks.

IBM_5.3k (Toledo et al., 2019): IBM_5.3k consists of 5.3k arguments selected from the UKPConvArg database (Habernal and Gurevych, 2016b), originally sourced from the Reddit CMV forum. Each argument has two types of labels: an individual argument quality label (absolute) and a relative argument-pair label (relative). For the absolute label, annotators are asked a binary yes/no question about whether they would recommend a friend preparing a speech supporting or contesting the topic to use the argument. The quality of each individual argument is a real-valued score between 0 and 1, defined by the fraction of 'yes' responses.

For the relative label, annotators are presented with a pair of arguments that take the same stance on a topic and are asked which of the two would be preferred by most people to support or contest the topic. The final dataset consists of 5.3k arguments, each selected based on high individual quality ratings, with an average of 11.4 valid annotations per argument.

IAC (Swanson et al., 2015): The dataset comprises 109,074 sentences covering four debate topics—gay marriage, gun control, the death penalty, and evolution—sourced from the IAC corpus Walker et al. (2012) and CreateDebate.com. Each sentence was annotated by seven Amazon Mechanical Turk workers with approval ratings above 95%. Annotations include a binary label indicating whether the sentence expresses an argument, as well as a continuous argument score ranging from 0 (difficult to interpret) to 1 (easy to interpret).

IDEA (Persing and Ng, 2017): The IDEA dataset consists of 165 debates obtained from the International Debate Education Association website. Each debate includes a motion that expresses a stance on a topic, along with an average of 7.3 arguments either supporting or opposing the motion. Each argument contains a one-sentence assertion of its stance and a justification explaining that stance. Two native English speakers annotated each argument with a persuasiveness score from 1 to 6, along with five types of errors that may have undermined its persuasiveness: grammar errors, lack of objectivity, inadequate support, unclear assertion, and unclear justification.

N Persuasiveness Prediction Performance Details

We evaluated persuasiveness prediction performance using two complementary metrics: Spearman correlation between predicted and ground-truth persuasiveness scores, and Root Mean Squared Error (RMSE). To assess the contribution of rhetorical features, we compared model performance under two conditions—with and without the inclusion of predicted strategy scores. As shown in Table 16 reports small but consistent improvements in predicting persuasiveness when incorporating rhetorical strategy. Within-domain, the strategy features increased the correlation with ground-truth persuasiveness scores by 0.03 and reduced RMSE by 0.014. These effect sizes are small but not negligible. For each dataset, the differences are statistically significant, with sample sizes rang-

	ConvArg (1038)		IBM-30k (30497)		IBM-5.3k (5298)		IAC (4939)		IDEA (1205)	
	Spearman's $\rho\uparrow$	**RMSE $\downarrow$**	**Spearman's $\rho\uparrow$**	**RMSE $\downarrow$**	**Spearman's $\rho\uparrow$**	**RMSE $\downarrow$**	**Spearman's $\rho\uparrow$**	**RMSE $\downarrow$**	**Spearman's $\rho\uparrow$**	**RMSE $\downarrow$**
Within Dataset - Vanilla	0.647 (0.012)	0.265 (0.004)	0.502 (0.004)	0.176 (0.004)	0.456 (0.010)	0.204 (0.004)	0.670 (0.000)	0.188 (0.008)	0.263 (0.021)	0.280 (0.007)
Within Dataset - Strategy	**0.680** (0.009)	**0.255** (0.003)	**0.516** (0.005)	**0.167** (0.003)	**0.478** (0.009)	**0.188** (0.004)	**0.678** (0.003)	**0.171** (0.005)	**0.337** (0.036)	**0.264** (0.007)
Δ	+0.033	-0.010	+0.014	-0.009	+0.022	-0.016	+0.008	-0.017	+0.074	-0.016
Cross Dataset - Vanilla	0.300 (0.018)	0.335 (0.003)	0.290 (0.005)	0.247 (0.019)	0.380 (0.005)	0.345 (0.004)	0.349 (0.003)	0.283 (0.004)	0.052 (0.010)	0.396 (0.005)
Cross Dataset - Strategy	**0.341** (0.016)	**0.326** (0.001)	**0.309** (0.009)	**0.218** (0.012)	**0.400** (0.005)	**0.335** (0.004)	**0.389** (0.014)	**0.257** (0.008)	**0.053** (0.009)	**0.395** (0.004)
Δ	+0.041	-0.009	+0.019	-0.029	+0.020	-0.010	+0.040	-0.026	+0.001	-0.001

Table 16: Persuasiveness score performance for the vanilla model and the model augmented with rhetorical strategies, evaluated within and across datasets. The table reports the mean and standard deviation of performance across three fine-tuning runs. Δ rows report the difference between strategy-enhanced and vanilla models (higher ρ, lower RMSE is better).

	ConvArg (1038)		IBM-30k (30497)		IBM-5.3k (5298)		IAC (4939)		IDEA (1205)	
	Test Set Size	**Mean SE**	**Test Set Size**	**Mean SE**	**Test Set Size**	**Mean SE**	**Test Set Size**	**Mean SE**	**Test Set Size**	**Mean SE**
Within Dataset - Vanilla	104	0.019	3050	0.002	530	0.005	494	0.005	120	0.013
Within Dataset - Strategy	104	0.014	3050	0.002	530	0.004	494	0.005	120	0.009
Cross Dataset - Vanilla	1038	0.007	30497	0.000	5298	0.003	4939	0.003	1205	0.008
Cross Dataset - Strategy	1038	0.007	30497	0.000	5298	0.002	4939	0.002	1205	0.008

Table 17: Test set size for each dataset and the average standard error (Mean SE) of the persuasiveness prediction model over three fine-tuning runs.

ing from hundreds to thousands of observations. The improvements represent a relative 8.40% increase in the magnitude of the correlations and a relative 6.30% decrease in RMSE, indicating better alignment with human judgments. In the more challenging cross-domain setting, we observe a relative 7.77% increase in correlation and a relative 6.16% reduction in RMSE. These results suggest that incorporating rhetorical strategies improves prediction not only within individual domains but also in previously unseen topical contexts.

O Individual Debate Level Strategy Comparisons

Year	**Candidates**	**Causal**	**Empirical**	**Emotional**	**Moral**
1960	Kennedy (D) vs. Nixon (R)	0.005	0.031	-0.041	-0.041
1976	Carter (D) vs. Gerald Ford (R)	0.002	0.014	0.090***	0.008
1980	Reagan (R) vs. Jimmy Carter (D)	0.055	0.010	-0.074*	0.036
1984	Reagan (R) vs. Mondale (D)	0.081**	-0.011	0.087***	0.087*
1988	H. W. Bush (R) vs. Dukakis (D)	0.041	0.000	-0.008	0.041
1992	B. Clinton (D) vs. H. W. Bush (R)	0.078**	0.193***	-0.019	0.022
1996	B. Clinton (D) vs. Dole (R)	0.067**	0.067*	-0.036*	0.056*
2000	G. W. Bush (R) vs. Gore (D)	-0.006	0.058**	-0.033*	-0.043*
2004	G. W. Bush (R) vs. John Kerry (D)	-0.022	0.021	0.023	-0.011
2008	Obama (D) vs. McCain (R)	0.025	-0.001	-0.052***	-0.004
2012	Obama (D) vs. Romney (R)	0.047*	0.016	0.000	0.017
2016	Trump (R) vs. H. Clinton (D)	0.030*	0.047***	-0.049***	0.050***
2020	Biden (D) vs. Trump (R)	0.006	0.016	-0.024*	0.045***

Note. Entries reflect the difference in strategy score (Democrat minus Republican) averaged over all utterances of each strategy type.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 18: **Partisan Differences in Rhetorical Strategy Between U.S. Presidential Debate Candidates.** Differences are measured as the Democrat's average over scores for each utterance, minus the Republican's average, broken down by four rhetorical strategies:.