A Hybrid Theory and Data-driven Approach to Persuasion Detection with Large Language Models

Gia Bao Hoang[1], Keith J. Ransom[1], Rachel G. Stephens[1],
Carolyn Semmler[1], Nicolas Fay[2], Lewis Mitchell[1]

[1]University of Adelaide
[2]University of Western Australia
{giabao.hoang, keith.ransom, rachel.stephens, carolyn.semmler, lewis.mitchell}@adelaide.edu.au,
nicolas.fay@uwa.edu.au

Abstract

Traditional psychological models of belief revision focus on face-to-face interactions, but with the rise of social media, more effective models are needed to capture belief revision at scale, in this rich text-based online discourse. Here, we use a hybrid approach, utilizing large language models (LLMs) to develop a model that predicts successful persuasion using features derived from psychological experiments.

Our approach leverages LLM generated ratings of features previously examined in the literature to build a random forest classification model that predicts whether a message will result in belief change. Of the eight features tested, *epistemic emotion* and *willingness to share* were the top-ranking predictors of belief change in the model. Our findings provide insights into the characteristics of persuasive messages and demonstrate how LLMs can enhance models of successful persuasion based on psychological theory. Given these insights, this work has broader applications in fields such as online influence detection and misinformation mitigation, as well as measuring the effectiveness of online narratives.

1 Introduction

Changing someone's opinion is a key aspect of communication, with broad implications for both individual decisions and societal outcomes. From political campaigns to marketing strategies and daily interactions, understanding how opinions change has been the subject of extensive research, (see e.g., Cialdini 1993; Dillard and Pfau 2002; Petty and Cacioppo 2012; Reardon 1991). With the rise of online social platforms, interpersonal persuasion can be observed and enacted on a massive scale, not just in direct conversation but also through text-based communication. As Fogg (2008) highlights, debate and argumentation are increasingly shifting to digital spaces. This introduces new complexities in understanding how persuasive messages are created, received, and how they influence the revision of beliefs in online environments.

Persuasion in these environments is particularly important to understand given global concerns about phenomena such as misinformation, polarization, and echo chambers (see e.g., Arechar et al. 2023; Barberá 2020; Weber et al. 2022). In addition, influence campaigns are shown to

be strategically coordinated efforts designed to shape public opinion using social diffusion processes and media exposure (Hwang 2012; Panagopoulos 2012). Therefore, early detection and a deeper understanding of how these processes unfold, along with identifying the characteristics of successful persuasion, are crucial.

Belief revision is a complex process. Beyond the content of a message, outcomes depend on how individuals engage with information, evaluate its credibility, and respond emotionally. Epistemic emotions (such as curiosity, confusion, and surprise) motivate deeper cognitive engagement (Muis et al. 2018), and have been shown to facilitate persuasion when they prompt active message processing (Briñol and Petty 2012). In parallel, emotional valence influences how information is processed, shaping attention, memory, and judgment (Lerner et al. 2015). These internal responses often determine whether persuasion occurs, yet they are not always evident from surface-level linguistic features. This complexity underscores the need for models that incorporate psychological insight, not just observable text patterns.

While foundational psychological models like the Elaboration Likelihood Model (ELM) (Petty and Cacioppo 1986) and the Heuristic-Systematic Model (HSM) (Chaiken 1980) have shaped our understanding of persuasion, they are primarily descriptive and offer limited utility for predicting persuasive success at the level of individual messages. Separately, researchers have also attempted to use machine learning and large language models (LLMs) to analyze persuasion at scale (Tan et al. 2016; Bai et al. 2023; Salvi et al. 2024). While promising, these approaches often function as black boxes, offering little insight into how persuasive effects are achieved. As a result, the relationship between linguistic features and psychological theories of persuasion remains difficult to interpret and apply in practice.

In this work, we demonstrate how psychological theories can be combined with LLMs' statistical representation of language, to enhance existing machine learning architectures to predict persuasion outcomes in text-based discussions. This study utilizes two datasets: the Winning Arguments dataset (Tan et al. 2016) for real-world online persuasion analysis, and the Truth Wins dataset (Fay et al. 2024), which provides controlled, human-annotated data from experiments on persuasive and attention-seeking messages. Our focus is on classifying successful versus unsuccessful

persuasive attempts in the Winning Arguments dataset, using features derived from the Truth Wins dataset. Our final model leverages a Random Forest architecture trained on LLM-generated persuasion ratings, inspired by psychological experiments. The model predicts whether a message will lead to successful persuasion. Additionally, we examine which features are the most influential predictors within the classification model.

2 Related Work

The ELM and the HSM are foundational theories in persuasion research, both proposing dual routes to attitude change, either through deep, content-focused processing or more superficial cues (Petty and Cacioppo 1986; Chaiken 1980). While influential, these models are largely descriptive, and offer little practical guidance for predicting persuasive success at the level of individual messages. Their application often requires adaptation to specific contexts, topics, and communication media, which limits their scalability in computational settings.

In response, researchers have explored ways to make persuasion more measurable. Zhao et al. (2011) proposed the Perceived Argument Strength (PAS) scale as a subjective measure of persuasive quality, but it relies on self-reports and lacks scalability. Youk et al. (2024) developed the Measures of Argument Strength (MAS), identifying linguistic features – such as citations, abstractness, and moral language – that are associated with higher perceived persuasiveness in large-scale online debates. Lukin et al. (2017) further emphasized how personality traits shape responses to emotional versus factual arguments, highlighting the importance of audience adaptation. While these studies move toward operationalizing persuasion, they remain constrained by subjective judgement, context sensitivity, handcrafted features, and limited predictive power – motivating the further shift toward data-driven approaches, discussed next.

Online platforms like Change My View (CMV) subreddit have become an important place to study persuasion at scale. In CMV, users influence opinions via textual exchanges. Two important features of CMV that are valuable for computational work are the *delta symbol*, awarded by a user to the reply that changed their view, and the *karma score*, which reflects community endorsement through up-votes and down-votes. While deltas serve as explicit markers of persuasive success, karma is often used as a proxy for argument quality. In addition, as the forum is highly moderated, comments are usually argumentative, containing reasoning rather than *ad hominem* attacks or other form of online bullying.

Tan et al. (2016) conducted a large-scale analysis of CMV (the Winning Arguments dataset), focusing on delta award predictions, based only on delta given by the original poster (OP). They examined the interaction dynamic associated with successful attempts at persuasion, and built a logistic regression model based on language factors and linguistic style. The model was able to predict delta changes effectively, but struggled with reliably distinguishing stance malleability (whether an OP will change their position), indicating difficulty in reliably distinguishing stance shifts

and highlighting the complexity of persuasion in online discourse. Subsequent studies have built on this dataset using more advanced techniques. Dutta, Das, and Chakraborty (2020) introduced an attention-based hierarchical Long Short-Term Memory model to classify persuasive conversations on CMV by analyzing argument-specific components, such as claims and premises. Similarly, Wei, Liu, and Li (2016) used karma scores as a proxy for success, finding that argumentative features outperformed surface-level cues like length and punctuation in predicting persuasiveness. These works demonstrate the promise of data-driven persuasion modeling, but also expose its dependence on platform-specific feedback, limiting broader applicability.

Recent work has shifted from predicting persuasive outcomes to detecting persuasive strategies, often using deep learning and transformer-based NLP models. Karki et al. (2022) classified Cialdini's six persuasion principles (Reciprocation, Consistency, Social Proof, Likeability, Authority, and Scarcity) in phishing emails using machine-learning transformers based on BERT. In a related direction, Wang et al. (2020) developed a personalized persuasive dialogue system by integrating sentence-level and contextual features, including turn position, sentiment, and character embeddings, enabling the classification of persuasion strategies in conversations. While these models perform well, they remain largely opaque and task-specific, offering limited insight into underlying persuasive mechanisms and weak connections to psychological theory, underscoring the need for models that are both interpretable and cognitively grounded.

Despite recent advances, traditional methods for persuasion detection continue to face challenges related to scalability, generalization, and personalization, often relying on large annotated datasets and lengthy text inputs. Large language models (LLMs) offer a promising alternative, leveraging their statistically grounded representations of human language to overcome these limitations. However, LLMs introduce new challenges – most notably, a lack of transparency. Operating as black-box systems, they are often difficult to interpret, which limits their usefulness in contexts requiring explainability. The current research project addresses these gaps by building on and extending three key areas of prior work: (1) models for predicting persuasive success, (2) studies identifying persuasive strategies and message characteristics, and (3) applications of LLMs in persuasion contexts. This integrative approach supports real-time applications such as influence campaign detection and misinformation mitigation, while also enabling greater flexibility in identifying and analyzing the persuasive elements within language.

3 Experimental Setup

Figure 1 outlines the workflow of our hybrid model for persuasion classification in target data such as the Winning Arguments dataset. The goal is to predict whether a response successfully changes the original poster's view by leveraging interpretable, theory-driven features in combination with scalable language model outputs. We begin by identifying a set of psychologically grounded features from the Truth Wins dataset (detailed below), including Attention, Interesting-If-True, Positive Emotion, Negative Emo-
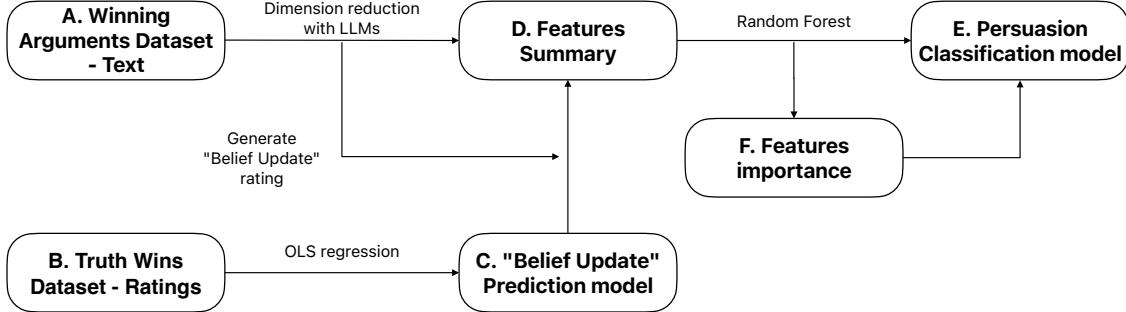
Figure 1: Flowchart showing the main steps of our hybrid model approach.

tion, Influential, Shareable, and Truthfulness (Figure 1.B). These features form the basis for a structured representation of persuasive content.

Next, comments from the Winning Arguments dataset (Figure 1.A) are passed to LLMs, which extract feature ratings based on the features derived from the Truth Wins dataset (Figure 1.D). To estimate the likelihood of belief change, we train a separate belief update regression model on annotated belief shift data from Truth Wins (Figure 1.C).

Together, the feature ratings generated by the LLMs and the belief update scores predicted by the regression model were input into a Random Forest classifier for binary classification (Figure 1.E). Performance of the Random Forest classifier is then measured using the test set of Winning Arguments. Using the Random Forest model's mechanism for evaluating feature importance enables the removal of noise or negatively contributing features, further refining the final model (Figure 1.F). However, while this refinement did little to improve accuracy, it did allow us to simplify the model and clarify the ranking of features.

3.1 Datasets

Winning Arguments Dataset (Figure 1.A) This is the primary dataset for this study, sourced from the Change My View subreddit where users invite others to challenge their opinions by presenting counterarguments (Tan et al. 2016). Persuasion is explicitly marked by the OP awarding a delta symbol (Δ) to comments that successfully change their view, thus distinguishing between successful persuasion and unsuccessful comments.

We focus on root replies, the first-level responses from challengers to the OP. For each post, we extract one positive (successful) and one negative (unsuccessful) comment. Since posts often receive multiple challenges, we control for content similarity by selecting comment pairs with the highest Jaccard similarity scores, in an effort to ensure that comparisons are based more on persuasiveness rather than content differences. The dataset consists of 3,456 posts for training and 807 for testing, each including the OP's text, a positive reply, and a negative reply. This offers a robust and ecologically valid foundation for studying persuasive dynamics

in online discourse.

Truth Wins Dataset (Figure 1.B) The Truth Wins dataset was collected from a series of human-subject experiments designed to study persuasion and attention-seeking in written messages across diverse topics (Fay et al. 2024). Participants were asked to generate messages either to persuade or to garner online attention. A second between-subjects manipulation varied whether participants were told to base their messages on true information, false information, or on any combination they wished.

Each message was rated by ten independent human annotators along multiple dimensions including: belief update (ranging from -100 to 100, reflecting the change between prior versus post belief score), attention, interesting, interesting-if-true, positive emotion, negative emotion, perceived effectiveness of influence, shareability, and truthfulness. Except for belief update, all ratings were recorded on a 5-point Likert scale: "not at all", "slightly", "somewhat", "very much" and "extremely". Following Fay et al. (2024), we assigned each item an integer value from 1 ("not at all") to 5 ("extremely"), and aggregated the ratings by averaging across raters to reduce individual variance while improving signal stability and generalizability. This aggregated structure aligns with our focus on identifying consistent predictors in successful persuasion.

These features are grounded in prior research. Epistemic emotions (interesting, interesting-if-true) promote deeper processing and belief change (Muis et al. 2018; Briñol and Petty 2012). Interestingness boosts engagement, though its effect on persuasion is context-dependent (Alwitt 2000; Murphy et al. 2005). Shareability captures perceived social attention, which supports deeper analysis and persuasive impact (Shteynberg et al. 2016; Shteynberg 2018). Emotional valence shapes attention and persuasion; positive emotions enhance message acceptance, while relevant negative emotions can also increase persuasion (Lerner et al. 2015; Petty, , and Briñol 2015; Huntsinger and Ray 2016). Truthfulness reflects perceived accuracy, which makes messages more persuasive and less likely to be dismissed. (Chaiken 1980; Eagly and Chaiken 1993; Kunda 1990; Petty and Cacioppo 1986; Tormala 2016).

While the original study explored the influence of truthfulness on persuasion and attention-seeking, our research examines broader factors affecting persuasiveness, acknowledging that attention-seeking messages can also carry persuasive elements. The controlled, human-labeled data from the Truth Wins can be used to help classify the more ecologically valid discourse found in the CMV-based Winning Arguments dataset, offering a comprehensive view of persuasive mechanisms across different contexts.

3.2 Baseline Models

To evaluate our proposed hybrid model, we compare it to three representative baselines: theory-driven inference, transparent surface-level text modeling, and black-box reasoning via large language models (LLMs). Together, these baselines form a spectrum of interpretability, linguistic abstraction, and data reliance.

Theory-Driven Model – Belief Update Regression (Figure 1.C) As a baseline, we modeled persuasion as belief change. Using the Truth Wins dataset, which includes human-annotated belief update ratings, we trained an ordinal least square (OLS) regression model to identify significant features predictive of belief update, using the other eight key features studied by Fay et al. (2024). Following a stepwise procedure, variables were added or removed based on p-value thresholds (0.05), with problematic features excluded to ensure model stability.

The resulting model was then applied to the Winning Arguments dataset to generate predicted belief update scores for each comment. These scores were then thresholded – tuned for accuracy on the training set – to assign binary persuasion labels. Though not optimized for classifying persuasive success, it offers a theory-informed perspective based on belief update scores. In our hybrid model, these belief update scores are also used as a feature, contributing to a psychologically grounded signal that complements language-based predictions. Note that belief update and successful persuasion are distinct constructs: not all persuasive responses result in explicit belief change, and not all belief changes are externally acknowledged.

Transparent Language Model — Logistic Regression on Term Frequencies This baseline uses a shallow, interpretable language model to classify persuasion based solely on surface-level lexical features. Each comment is tokenized and vectorized using term frequency (TF), omitting inverse document frequency (IDF) to better accommodate short-form text, where rare words may not carry more informational weight. We train a logistic regression with 10-fold cross-validation, tuned on regularization strength, and evaluated performance using ROC AUC. This model provides a transparent lexical benchmark, revealing how much signal can be captured from word-level frequency patterns alone.

Black-Box Language Model — Zero-Shot LLM Classification To assess the potential of large language models (LLMs), we evaluated three models: LLaMA3-70B (LLaMA3), Gemma2-9B (Gemma2), and Mixtral-8x7B (Mixtral) in a zero-shot persuasion task. Each model receives the OP's post and two responses: one that successfully persuaded the OP and one that did not. The models are then prompted to select the response they judge to be more persuasive based on the input. The full prompt is provided in Listing 1 in Appendix A. All models are run with temperature set to 0 and a fixed seed (42) to ensure high probability responses and reproducibility.

This approach demonstrates the scalability and adaptability of LLMs, which can be applied without fine-tuning or additional supervision. Although tools like Gemma Scope provide interpretability scaffolding, these models still operate with a high degree of opacity, and their predictions can be difficult to interpret or validate – especially in high-stakes contexts such as misinformation detection, where transparency and accountability are critical. To mitigate these limitations, we subsequently use the same LLMs to generate structured feature ratings. These outputs are integrated into a hybrid model to support more interpretable and feature-aware classification of persuasive language.

3.3 Hybrid Model

Feature Extraction Using LLMs (Figure 1.D) Using the same three LLMs used for the Black-Box model (LLaMA3, Gemma2, and Mixtral), we generated feature-by-feature ratings for comments in the Winning Arguments dataset. Given the original post along with a pair of comments (one labeled positive, one negative), each model was prompted to generate ratings for both comments in relation to the original post. The full prompt is provided in Listing 2 in Appendix A. The selected features rated by the LLMs were adapted from the Truth Wins dataset and are defined as follows:

1. *Influential*: How effectively the comment influences the reader's perspective.

2. *Interesting*: How engaging and attention-grabbing the comment is.

3. *Interesting-If-True*: Potential interest level if the comment's claims were true.

4. *Positive*: Overall positivity, focusing on an uplifting or encouraging tone.

5. *Negative*: Overall negativity, considering critical or discouraging aspects.

6. *Shareable*: Likelihood of the comment being shared.

7. *Truthfulness*: Apparent accuracy and credibility of the comment.

8. *Attention*: Ability to capture and maintain reader focus.

To avoid ambiguity, we replaced the original Fay et al. (2024) feature label *Persuasive* with *Influential*. This *Influential* feature captures the perceived effectiveness of a comment in influencing a reader's perspective, rather than directly measuring actual belief change. This distinction prevents confusion between the general notion of perceived "persuasiveness" as a feature and the task of predicting successful persuasion itself. Using the feature schema derived from the Truth Wins dataset, we obtained LLM-generated ratings for the same features in the Winning Arguments
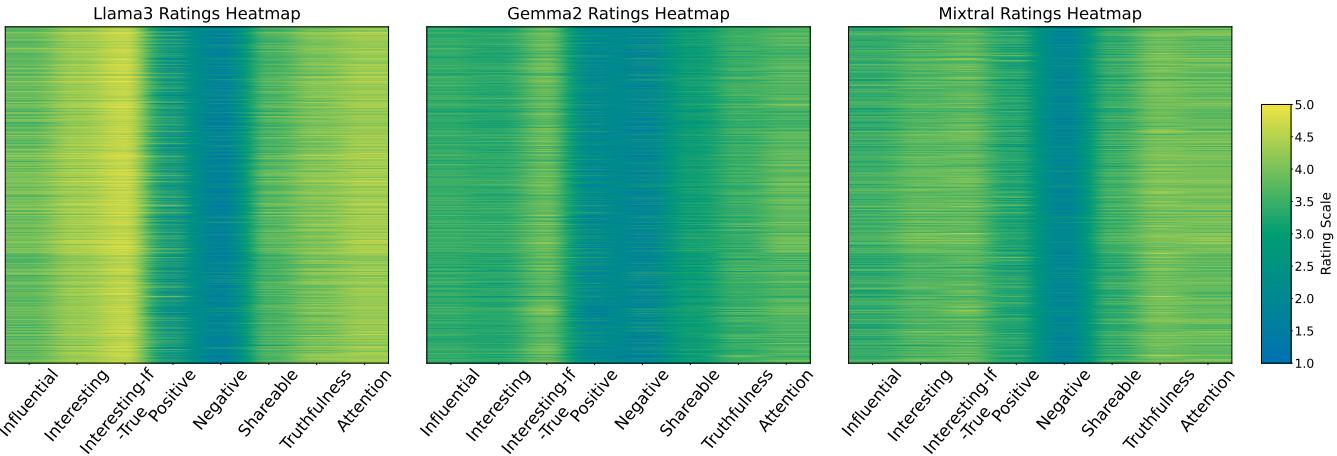
Figure 2: LLMs Generated Ratings Heatmap - This visualizes the ratings assigned by each LLM for each analysed feature (x-axis) for each argument in the Winning Arguments dataset (y-axis).

Model	Data-Driven	Theory-Driven	Hybrid Independent	Hybrid Interaction
Transparent (Logistic regression)	56.50%	-	-	-
Black-Box (LLaMA3)	64.93%	54.30%	73.17%	73.11%
Black-Box (Gemma2)	56.90%	63.80%	82.32%	82.69%
Black-Box (Mixtral)	61.71%	58.20%	72.55%	73.17%

Table 1: Accuracy comparison of models used for classifying successful persuasion (positive) vs. unsuccessful (negative) comments in the Winning Arguments dataset. **Transparent Data-Driven** refers to a logistic regression on term frequencies. **Black-Box Data-Driven** represents zero-shot classification from black-box LLMs. **Theory-Driven** uses thresholded belief update scores derived from regression models. **Hybrid Independent** and **Hybrid Interaction** refer to Random Forest models that combine features with belief update scores, using either independent terms only or including pairwise feature interactions, respectively.

dataset, providing a theory-driven input layer for persuasion classification. This conceptualization aligns closely with the probabilistic framework described by Wyer Jr. and Goldberg (1970), which distinguishes between a message's influence on a premise and the subsequent belief update. This approach enabled a structured analysis of persuasive strategies in real-world discourse, grounded in a controlled, theory-informed framework.

Independent vs. Interaction Terms   To assess the role of feature interactions in persuasion prediction, we developed two hybrid model variants: one using only independent features, and another incorporating feature interactions. The independent-term model evaluates each feature's contribution individually. In contrast, the interaction-term model includes pairwise combinations of features alongside their independent counterparts. For simplicity, interactions are limited to two-feature pairs. This approach enables us to determine whether interactions between features significantly enhance predictive performance or whether individual features alone are sufficient for effective persuasion classification.

Random Forest Classification (Figure 1.E)   We use a Random Forest (RF) classifier (Breiman 2001) for its strong predictive performance combined with the interpretability of decision trees. Our hybrid RF models are trained on LLM-generated features and belief update scores derived from regression models. The hybrid models are then evaluated on the test set. At this stage, the OP's post content is not taken into consideration. The RF-independent model combines LLM-generated features with belief update scores derived from the independent-term regression, while the RF-interaction model extends this by incorporating pairwise interaction features and scores from the interaction-term regression. Using RF's built-in permutation-based Variable Importance (VI), we assess which characteristics extracted by LLMs most meaningfully predict persuasive success, driving model performance.

VI assesses each feature by measuring decrease in classification accuracy when its values are randomly permuted, directly quantifies each feature's predictive contribution. Throughout this study, unless otherwise specified, "feature importance" specifically refers to VI. Using VI, we can remove noisy or redundant features, streamline the model, and mitigate potential overfitting without sacrificing accuracy. Although VI can be biased toward high-cardinality features (Strobl et al. 2007), our dataset consists exclusively of ordinal categorical features (rated 1 to 5) and continuous belief update scores, mitigating this bias. Our random forest classifier uses 300 trees, balancing complexity and generalizability. A fixed random seed (42) ensures reproducibility.
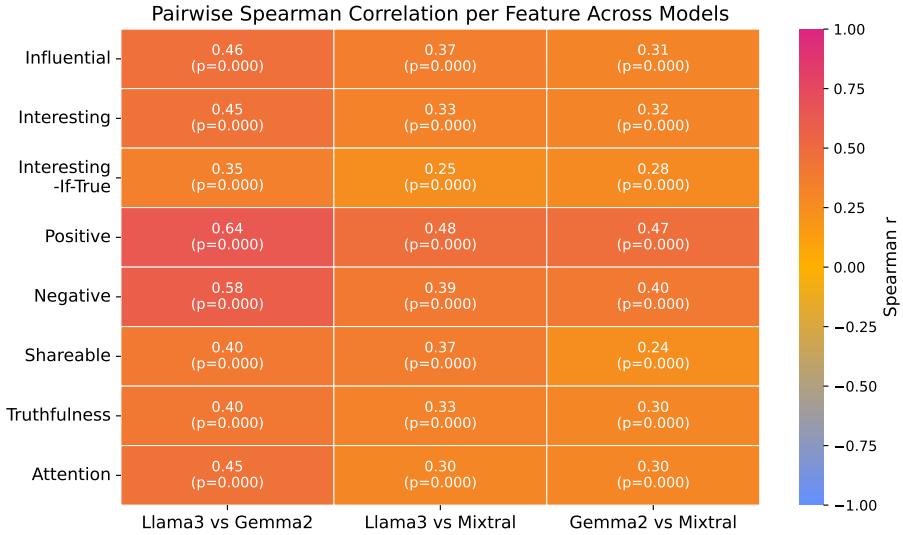
Figure 3: Spearman Rank Correlation between LLMs' generated ratings. Each cell represents the correlation between the ratings generated by the given pair of LLMs (x-axis) for a given feature (y-axis).

4 Results

4.1 Cross-Model Agreement in Generated Feature Ratings

To examine how different language models interpret identical instructions for feature ratings, we compared ratings generated by LLaMA3, Gemma2, and Mixtral on the full Winning Arguments dataset. Each model independently rated the following features: *Influential*, *Interesting*, *Interesting-If-True*, *Positive*, *Negative*, *Truthfulness*, *Attention*, and *Shareable*.

Heatmaps (Figure 2) reveal systematic rating differences across models. LLaMA3 consistently produced higher ratings for features such as *Influential*, *Interesting*, *Interesting-If-True*, and *Attention*, suggesting a relatively generous interpretation of these criteria. In contrast, Gemma2 and Mixtral tended toward lower ratings overall, with Gemma2 notably rating *Positive* lower and giving higher ratings for *Interesting-If-True* and *Attention*. Mixtral's ratings typically fell between those of LLaMA3 and Gemma2.

A pairwise correlation analysis (Figure 3) demonstrated consistent positive and statistically significant (p-value < 0.001) monotonic correlations between models, despite differences in absolute ratings. LLaMA3 and Gemma2 consistently showed the strongest correlations across features, with values reaching 0.64 for *Positive* and 0.58 for *Negative*, highlighting close agreement in valence judgments. In contrast, strong predictive features like *Interesting-If-True* showed weaker correlations across all pairs, pointing to greater ambiguity or model-specific interpretation.

These differences in both absolute scores and correlation structure help explain variation in downstream performance across models. At the same time, the consistent alignment in rating trends between LLMs, as shown in Figure 3 through statistically significant positive monotonic correlations, indicates a degree of consistency across models. This observed alignment suggests that the LLM-generated features may be suitable for use in downstream modeling, despite variation in individual rating scales.

4.2 Model Accuracy Performance

We evaluated model performance for classifying successful persuasion (positive) vs. unsuccessful (negative) comments in the Winning Arguments dataset, summarized in Table 1.

Hybrid classifiers integrating LLM-generated features outperformed both the theory-driven model, and transparent and black-box language model baselines. The Gemma2-based hybrid model achieved the highest accuracy, roughly 82% for both independent-term and interaction-term variants, approximately a 25% improvement over its black-box baseline. Hybrid models built on LLaMA3 and Mixtral followed closely, with accuracies around 73% and similarly small differences between the two feature configurations.

Interestingly, including feature interactions and removing negatively contributing features had little impact on predictive accuracy across all hybrid models. This indicates that independent features, as generated by LLMs, already sufficiently capture the predictive signals necessary for effective persuasion classification.

4.3 Feature Importance (Figure 1.F)

Figure 4 summarizes the most important model features. Despite differences amongst LLMs, all consistently identify a core set of important features: *Influential*, *Shareable*, *Interesting/Interesting-If-True*, and *Attention*. Specifically, in the independent-term models, LLaMA3 emphasized *Interesting-If-True*, Gemma2 prioritized *Attention*, and Mixtral strongly highlighted *Influential*. Interaction-term models revealed that combinations involving *Interesting-If-True*, *Interesting*, and *Influential* substantially enhanced predictive performance, indicating potentially informative in-
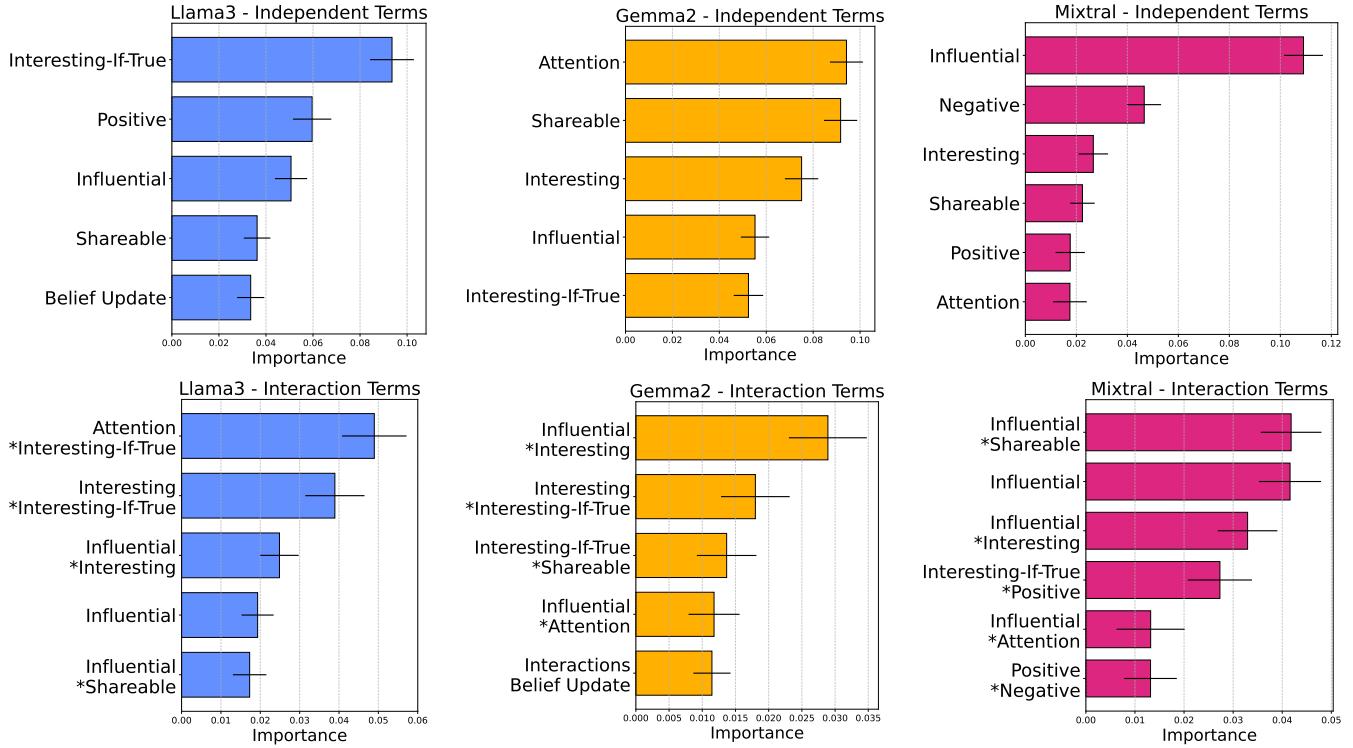
Figure 4: Top 5 Most Important Features in the Hybrid Model (Random Forest), with Independent Terms shown on the top row and Interaction Terms on the bottom row. Each bar represents the mean feature importance across 100 permutations, and the error line indicates the standard deviation. For the Mixtral-based models, six features are shown due to a tie in importance scores for the fifth position.

teractions among these features. Additionally, *Belief Update* was influential in Gemma2, suggesting its relevance in predicting persuasive success. Emotional features (*Positive* and *Negative*) showed inconsistent predictive power, warranting further investigation. Notably, *Truthfulness* consistently ranked among the least important features, which may reflect limited variation in the accuracy of information within CMV posts rather than a lack of relevance to persuasion.

5 Discussion

Overall, our findings show that hybrid models outperform both theory-driven and data-driven baselines, suggesting that combining LLMs with structured, theory-informed features improves the prediction of persuasive success. Notably, LLMs perform best when used to generate interpretable ratings for individual features, highlighting the value of integrating statistical language models with domain knowledge to better model persuasive success.

5.1 Theory-Driven Model

The belief update prediction model, trained on the Truth Wins dataset and applied to LLM-generated feature ratings from the Winning Arguments dataset, reveals critical insights into cross-domain generalizability. While the model offers a broad indication of belief change, its simplicity – using ordinal linear regression – fails to capture the com-

plexities of successful persuasion. Incorporating the belief update score into the final model adds some interpretability but does not significantly improve performance. This suggests that belief update scores, while useful, are not sufficient to fully capture the dynamics of persuasive impact.

What we learn here is that, although theory-driven models can provide valuable insights, they struggle to translate across domains and fail to account for the nuanced, dynamic nature of persuasion. This underscores the need for more sophisticated, hybrid models that combine theoretical insights with the strengths of machine learning to better capture the complexities of persuasion and improve generalizability across diverse data sources.

5.2 Transparent Data-Driven Model

The logistic regression model based on term frequency (TF) highlights the limitations of using word frequency alone for persuasion detection. TF models, particularly for short-form text like CMV comments, struggle to capture linguistic and contextual elements such as sentiment and interaction dynamics. This emphasizes the need for more theory-informed features and predictors that connect specific linguistic cues to successful persuasion – an area where the hybrid model, integrating both theory-driven insights and data-driven techniques, can offer a more comprehensive solution.

5.3 Black-Box Data-Driven Model - Leveraging LLMs

Zero-shot classifiers, based on LLMs, demonstrated an improvement over logistic regression, though directly prompting LLMs to select the more persuasive comment proved less effective. LLMs excel at converting linguistic features into ratings that capture statistical regularities in human communication. This ability to represent language numerically enhances the performance of downstream models, particularly when structured data is needed. Although some models performed better than others in zero-shot classification, the real value of LLMs lies in their capacity to encode linguistic features into structured data. This underscores the potential of LLMs to aid in tasks requiring large-scale, feature-rich representations.

Research highlights the potential of LLMs as a powerful tool for persuasion analysis and belief influence. Studies have shown that LLMs, such as GPT-3.5, align closely with human judgments on persuasiveness, with high correlation between human and LLM-generated ratings (Fay et al. 2024). In fact, LLMs have been found to outperform humans in assessing persuasiveness across various topics and demographics (Salvi et al. 2024) and can effectively influence opinions on both polarized and less polarized issues (Bai et al. 2023). Additionally, Costello, Pennycook, and Rand (2024) demonstrated that conversations with GPT-4 Turbo could reduce belief in conspiracy theories by roughly 20%, with effects lasting up to two months, showcasing LLMs' potential in belief change.

Given these capabilities, future research should explore comparisons between LLM-generated and human ratings for accuracy and consistency. Hybrid approaches could further enhance performance in complex tasks like influence campaign detection and misinformation analysis, where a deeper understanding of persuasive content and context is crucial.

5.4 Cognitive Features and Biases in Persuasion

The top predictive features identified across all three models were *Interesting-If-True*, *Influential*, and *Shareable*. In the context of the CMV forum, where the primary goal is to change the OP's viewpoint, these features become particularly relevant. Unlike everyday conversations, which may not always prioritize persuasion, the CMV forum encourages high-quality, persuasive arguments. As a result, less obvious features like curiosity (*Interesting*-If-True) and social sharing (*Shareable*) become crucial factors in shifting opinions.

In the Hybrid Independent models, *Interesting-If-True* ranks highest for LLaMA3, while *Attention* holds the top position for Gemma2. Both features also rank among the top five for Mixtral. These features are closely tied to epistemic emotions, which stimulate cognitive engagement (Muis et al. 2018). Curiosity fosters exploration and information seeking, encouraging individuals to gather evidence and critically evaluate messages to resolve their curiosity (Vogl et al. 2019). In the context of persuasion, a message that sparks curiosity – such as "That's interesting, is it true?" – may prompt the audience to scrutinize it more rigorously,

potentially verifying its credibility before accepting it.

Similarly, interest sustains attention and systematic processing. As part of the broader category of epistemic emotions, interest – like curiosity and surprise – has been shown to enhance message processing and facilitate persuasion by prompting deeper cognitive engagement (Muis et al. 2018; Briñol and Petty 2012). These emotions are also associated with increased message sharing and heightened perceptions of credibility, even when the underlying content lacks truthfulness (Rijo and Waldzus 2023). This may explain why *Shareable* is influential across all three models, alongside *Interesting-If-True*. Both enhance emotional engagement, mediating how information is perceived, evaluated, and disseminated online.

Truthfulness was a weaker predictor of persuasive success in CMV. Perceived truthfulness might be more important in environments with a wider mix of objectively true and false information. Research by Rijo and Waldzus (2023) suggests that fake news often elicits stronger epistemic emotional responses than true news, making it appear more credible and leading to wider dissemination. These findings highlight the dominance of perceived truth over objective truth in shaping belief formation and information spread. Moreover, regarding misinformation, Pennycook and Rand (2022) highlighted that while people can typically distinguish between true and false information when asked, this analytical distinction fades in informal contexts like social media. In these environments, social and affective factors – rather than truth-based analysis – primarily drive sharing behavior.

Given this focus on persuasion rather than truthfulness, future research could further explore the role and drivers of perceived truth quality in the Winning Arguments dataset to assess whether it plays a significant role in persuasion. For example, Fazio et al. (2015) demonstrated the illusory truth effect in news classification: repetition of a claim (even if false) can increase its perceived truth and likelihood of being shared. Investigating whether similar effects apply in the CMV context could offer further insights into how persuasion works, regardless of objective truth.

Emotion was also a significant feature, with *Positive* emotions ranking highly for LLaMA3 and *Negative* emotions for Mixtral. This aligns with research on how emotions influence persuasion. According to Naranowicz (2022), emotional state affects critical engagement: negative moods promote scrutiny of argument quality, while positive moods encourage reliance on credibility cues and general trust. Furthermore, Paletz et al. (2023) and Fay et al. (2024) found that positive emotions drive content sharing, reinforcing the role of emotional engagement in persuasion. These findings highlight the importance of emotional features in our model.

While our analysis captures emotional valence, it overlooks arousal, which plays a crucial role in persuasion. High-arousal emotions – whether positive or negative – amplify persuasive impact by encouraging fast, intuitive processing (Lühring et al. 2024), leading to snap judgments and reduced source verification. Since arousal is absent from our feature set, future research should examine its role in persuasion. Incorporating arousal-related features could provide deeper insights into decision-making, credibility assessments, and

content sharing, offering a more comprehensive understanding of emotional engagement in persuasion.

6 Conclusion

This study demonstrates the potential of large language models (LLMs) as effective tools for predicting successful persuasion in online discourse. By generating feature ratings on dimensions such as influence effectiveness, interest, and shareability, LLMs provide a scalable method for analyzing persuasive language, bridging the gap between manual annotation and automated analysis.

Beyond their practical applications, the findings support theoretical perspectives on persuasion. Feature importance analysis in the Random Forest model revealed significant patterns aligning with existing literature, reinforcing the role of linguistic and cognitive factors in persuasive content. This supports the validity of the feature set and highlights its relevance in computational persuasion research.

Trained on high-quality arguments from Change My View subreddit, the model represents a step forward in both theoretical and applied research. Automating the rating process for belief prediction offers valuable insights into opinion shifts in digital spaces. By integrating theory-driven insights with data-driven modeling, this work advances both our understanding of persuasive mechanisms and real-world applications, including influence campaign detection, misinformation mitigation, and content moderation.

Limitations

The Winning Arguments dataset, sourced from a forum focused on persuasion, may limit the generalizability of our findings. Since users on CMV are actively trying to change others' opinions, the dynamics may differ from more organic or varied settings. Future research should apply the model to other datasets where persuasive intent is less explicit, such as general social media platforms, news comments, or product reviews, to evaluate whether the model performs well across different online environments.

Another key limitation of this study is the potential bias introduced by relying on LLM-generated ratings. While LLMs offer a scalable way to assess textual dimensions, their consistency and accuracy relative to human annotations remain uncertain. Prior work has shown that LLMs can exhibit preference for specific token patterns (Zheng et al. 2024), produce less consistent ratings than human annotators (Stureborg, Alikaniotis, and Suhara 2024), and reflect cognitive biases learned from training data (Dillion et al. 2023). These issues can affect tasks that require nuanced or objective judgment. Although LLMs may align with human judgments in broad patterns, their outputs can still be biased or unstable, potentially impacting downstream modeling performance.

A related concern is the lack of validation against human-sourced ratings. The hybrid models in this study rely primarily on LLM-generated ratings for the classification task, but the absence of human annotations limits our ability to assess the reliability of those inputs. Future work could compare LLM-generated ratings with human-annotated counterparts for the same features. This would provide a form of sanity check, establish a ground truth for model evaluation, and help identify systematic rating discrepancies across models.

Relying on LLM-generated feature ratings may overlook important linguistic nuances such as grammar, sentence structure, and subtle rhetorical or emotional cues, which are critical for understanding persuasive success. Future work could incorporate more advanced natural language processing techniques, integrating syntactic and semantic analysis alongside LLM-generated features to capture a deeper, richer understanding of persuasion.

LLMs often function as black boxes, making their internal decision-making opaque. This creates challenges in understanding how they determine which features of an argument are persuasive. Future research should explore interpretability techniques to better understand the patterns and factors that drive the model's outputs. These tools can provide insights into the patterns LLMs use to generate persuasive judgments, enhancing model transparency and usability.

Ethical Consideration

This study draws upon two previous datasets: the Winning Arguments dataset and the Truth Wins dataset. Both datasets were collected in prior research and are used here under the assumption that the original studies adhered to appropriate ethical standards for data collection. Readers are encouraged to refer to the respective publications for further details on data sourcing, consent, and anonymization procedures.

A portion of the data in this study is generated using Large Language Models (LLMs). While these tools offer efficiency and scale, they also raise several ethical concerns. Most notably, LLMs function as black boxes. Users have limited insight into the data used during training, which may contain biases, or misinformation. Furthermore, the proprietary nature of the LLMs employed (accessed via the Groq Cloud API) means they are not open-source, and their behavior or availability can change without notice. These models are costly to retrain and may become outdated over time. Additional risks include the potential for output degradation, loss of coherence, and generation of repetitive or harmful content. These limitations must be taken into account when interpreting results involving LLM-generated data.

The development of a persuasion prediction model presents a dual-use dilemma. On one hand, this research aims to enhance our understanding of persuasive communication, with the broader goal of fostering safer and more constructive online environments. By identifying linguistic and structural features associated with successful persuasion, we hope to contribute to the creation of more responsible and effective messaging strategies. On the other hand, we recognize that these insights could be misused—to manipulate public opinion, spread misinformation, or exploit vulnerable audiences. As such, we stress that the findings of this study must be applied with caution and responsibility. Ethical safeguards, transparency in deployment, and continuous monitoring are essential to mitigate potential misuse.

Acknowledgements

References

Alwitt, L. F. 2000. Effects of Interestingness on Evaluations of TV Commercials. *Journal of Current Issues & Research in Advertising*, 22(1): 41–53. Publisher: Routledge _eprint: https://doi.org/10.1080/10641734.2000.10505100.

Arechar, A. A.; Allen, J.; Berinsky, A. J.; Cole, R.; Epstein, Z.; Garimella, K.; Gully, A.; Lu, J. G.; Ross, R. M.; Stagnaro, M. N.; Zhang, Y.; Pennycook, G.; and Rand, D. G. 2023. Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, 7(9): 1502–1513. Publisher: Nature Publishing Group.

Bai, M. H.; Voelkel, J. G.; Eichstaedt, j. C.; and Willer, R. 2023. Artificial Intelligence Can Persuade Humans on Political Issues.

Barberá, P. 2020. *Social Media, Echo Chambers, and Political Polarization*, 34–55. Cambridge University Press.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1): 5–32.

Briñol, P.; and Petty, R. E. 2012. A history of attitudes and persuasion research. In *Handbook of the History of Social Psychology*. Psychology Press. ISBN 978-0-203-80849-8. Num Pages: 38.

Chaiken, S. 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39(5): 752–766. Place: US Publisher: American Psychological Association.

Cialdini, R. B. 1993. Influence: The psychology of persuasion. *HarperCollins*.

Costello, T. H.; Pennycook, G.; and Rand, D. G. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714): eadq1814. Publisher: American Association for the Advancement of Science.

Dillard, J. P.; and Pfau, M. 2002. *The persuasion handbook: Developments in theory and practice*. Sage Publications.

Dillion, D.; Tandon, N.; Gu, Y.; and Gray, K. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7): 597–600.

Dutta, S.; Das, D.; and Chakraborty, T. 2020. Changing views: Persuasion modeling and argument extraction from online discussions. *Information Processing & Management*, 57(2): 102085.

Eagly, A. H.; and Chaiken, S. 1993. *The psychology of attitudes*. The psychology of attitudes. Orlando, FL, US: Harcourt Brace Jovanovich College Publishers. ISBN 978-0-15-500097-1. Pages: xxii, 794.

Fay, N.; Ransom, K.; Walker, B.; Howe, P.; Perfors, A.; and Kashima, Y. 2024. Truth Wins: True Information is More Persuasive and Shareable than Falsehoods.

Fazio, L. K.; Brashier, N. M.; Payne, B. K.; and Marsh, E. J. 2015. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5): 993–1002. Place: US Publisher: American Psychological Association.

Fogg, B. J. 2008. Mass Interpersonal Persuasion: An Early View of a New Phenomenon. In Oinas-Kukkonen, H.; Hasle, P.; Harjumaa, M.; Segerståhl, K.; and Øhrstrøm, P., eds., *Persuasive Technology*, 23–34. Berlin, Heidelberg: Springer. ISBN 978-3-540-68504-3.

Huntsinger, J. R.; and Ray, C. 2016. A flexible influence of affective feelings on creative and analytic performance. *Emotion*, 16(6): 826–837. Place: US Publisher: American Psychological Association.

Hwang, Y. 2012. Social Diffusion of Campaign Effects: Campaign-Generated Interpersonal Communication as a Mediator of Antitobacco Campaign Effects. *Communication Research*, 39(1): 120–141. Publisher: SAGE Publications Inc.

Karki, B.; Abri, F.; Namin, A. S.; and Jones, K. S. 2022. Using Transformers for Identification of Persuasion Principles in Phishing Emails. In *2022 IEEE International Conference on Big Data (Big Data)*, 2841–2848.

Kunda, Z. 1990. The case for motivated reasoning. *Psychological Bulletin*, 108(3): 480–498. Place: US Publisher: American Psychological Association.

Lerner, J. S.; Li, Y.; Valdesolo, P.; and Kassam, K. S. 2015. Emotion and Decision Making. *Annual Review of Psychology*, 66(Volume 66, 2015): 799–823. Publisher: Annual Reviews.

Lukin, S. M.; Anand, P.; Walker, M.; and Whittaker, S. 2017. Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion. ArXiv:1708.09085 [cs].

Lühring, J.; Shetty, A.; Koschmieder, C.; Garcia, D.; Waldherr, A.; and Metzler, H. 2024. Emotions in misinformation studies: distinguishing affective state from emotional response and misinformation recognition from acceptance. *Cognitive Research: Principles and Implications*, 9(1): 82.

Muis, K. R.; , C., Marianne; ; and Singh, C. A. 2018. The Role of Epistemic Emotions in Personal Epistemology and Self-Regulated Learning. *Educational Psychologist*, 53(3): 165–184. Publisher: Routledge _eprint: https://doi.org/10.1080/00461520.2017.1421465.

Murphy, P. K.; Holleran, T. A.; Long, J. F.; and Zeruth, J. A. 2005. Examining the complex roles of motivation and text medium in the persuasion process. *Contemporary Educational Psychology*, 30(4): 418–438.

Naranowicz, M. 2022. Mood effects on semantic processes: Behavioural and electrophysiological evidence. *Frontiers in Psychology*, 13. Publisher: Frontiers.

Paletz, S. B.; Johns, M. A.; Murauskaite, E. E.; Golonka, E. M.; Pandža, N. B.; Rytting, C. A.; Buntain, C.; and Ellis, D. 2023. Emotional content and sharing on Facebook:

A theory cage match. *Science Advances*, 9(39): eade9231. Publisher: American Association for the Advancement of Science.

Panagopoulos, C. 2012. Campaign Context and Preference Dynamics in U.S. Presidential Elections. *Journal of Elections, Public Opinion and Parties*, 22(2): 123–137. Publisher: Routledge _eprint: https://doi.org/10.1080/17457289.2012.662232.

Pennycook, G.; and Rand, D. G. 2022. Nudging Social Media toward Accuracy. *The ANNALS of the American Academy of Political and Social Science*, 700(1): 152–164. Publisher: SAGE Publications Inc.

Petty, R. E.; ; and Briñol, P. 2015. Emotion and persuasion: Cognitive and meta-cognitive processes impact attitudes. *Cognition and Emotion*, 29(1): 1–26. Publisher: Routledge _eprint: https://doi.org/10.1080/02699931.2014.967183.

Petty, R. E.; and Cacioppo, J. T. 1986. The Elaboration Likelihood Model of Persuasion. In Berkowitz, L., ed., *Advances in Experimental Social Psychology*, volume 19, 123–205. Academic Press.

Petty, R. E.; and Cacioppo, J. T. 2012. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer Science & Business Media.

Reardon, K. K. 1991. *Persuasion in practice*. Sage.

Rijo, A.; and Waldzus, S. 2023. That's interesting! The role of epistemic emotions and perceived credibility in the relation between prior beliefs and susceptibility to fake-news. *Computers in Human Behavior*, 141: 107619.

Salvi, F.; Ribeiro, M. H.; Gallotti, R.; and West, R. 2024. On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial. ArXiv:2403.14380 [cs].

Shteynberg, G. 2018. A collective perspective: shared attention and the mind. *Current Opinion in Psychology*, 23: 93–97.

Shteynberg, G.; Bramlett, J. M.; Fles, E. H.; and Cameron, J. 2016. The broadcast of shared attention and its impact on political persuasion. *Journal of Personality and Social Psychology*, 111(5): 665–673. Place: US Publisher: American Psychological Association.

Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; and Hothorn, T. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1): 25.

Stureborg, R.; Alikaniotis, D.; and Suhara, Y. 2024. Large Language Models are Inconsistent and Biased Evaluators. ArXiv:2405.01724 [cs].

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 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, 613–624. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4143-1.

Tormala, Z. L. 2016. The role of certainty (and uncertainty) in attitudes and persuasion. *Current Opinion in Psychology*, 10: 6–11.

Vogl, E.; Pekrun, R.; Murayama, K.; Loderer, K.; and Schubert, S. 2019. Surprise, Curiosity, and Confusion Promote Knowledge Exploration: Evidence for Robust Effects of Epistemic Emotions. *Frontiers in Psychology*, 10. Publisher: Frontiers.

Wang, X.; Shi, W.; Kim, R.; Oh, Y.; Yang, S.; Zhang, J.; and Yu, Z. 2020. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. ArXiv:1906.06725 [cs].

Weber, D.; Falzon, L.; Mitchell, L.; and Nasim, M. 2022. Promoting and countering misinformation during Australia's 2019–2020 bushfires: a case study of polarisation. *Social Network Analysis and Mining*, 12(1): 64.

Wei, Z.; Liu, Y.; and Li, Y. 2016. Is This Post Persuasive? Ranking Argumentative Comments in Online Forum. In Erk, K.; and Smith, N. A., eds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 195–200. Berlin, Germany: Association for Computational Linguistics.

Wyer Jr., R. S.; and Goldberg, L. 1970. A probabilistic analysis of the relationships among belief and attitudes. *Psychological Review*, 77(2): 100–120. Place: US Publisher: American Psychological Association.

Youk, S.; Malik, M.; Chen, Y.; Hopp, F. R.; and Weber, R. 2024. Measures of Argument Strength: A Computational, Large-Scale Analysis of Effective Persuasion in Real-World Debates. *Communication Methods and Measures*, 18(1): 7–29. Publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2023.2230866.

Zhao, X.; Strasser, A.; Cappella, J. N.; Lerman, C.; and Fishbein, M. 2011. A Measure of Perceived Argument Strength: Reliability and Validity. *Communication Methods and Measures*, 5(1): 48–75. Publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2010.547822.

Zheng, C.; Zhou, H.; Meng, F.; Zhou, J.; and Huang, M. 2024. Large Language Models Are Not Robust Multiple Choice Selectors. ArXiv:2309.03882 [cs].

A   LLMs Prompts

Listing 1: Prompt used for zero-shot classification with LLMs.

```
1   <|begin_of_text|><|start_header_id|>
        system<|end_header_id|>
2   You are a classifier evaluating the
        persuasiveness of two comments on an
        original post.
3   Identify which comment is more
        persuasive.
4   Label the more persuasive comment as "
        positive" (1) and the less persuasive
        as "negative" (0).
5
6   Input includes:
7   - "op_text": original post text
8   - "comment1": first comment
9   - "comment2": second comment
10
```

```
11  Provide the result as a JSON object for
        each comment. No explanation is
        required.
12  <|eot_id|><|start_header_id|>user<|
        end_header_id|>
13  Here is the context and comments: {
        context}
14  <|eot_id|><|start_header_id|>assistant<|
        end_header_id|>
```

Listing 2: Prompt used for LLM-based feature extraction. In the original prompt, the label *Persuasive* was used; however, we refer to it as *Influential* throughout the paper to avoid ambiguity, as discussed in Section 3.3

```
1   begin_of_text|><|start_header_id|>system
        <|end_header_id|>
2   You are an expert evaluator tasked with
        rating two comments based on their
        context within an original post.
3   Each feature is rated from "one" (lowest
        ) to "five" (highest).
4
5   Input includes: "op_text" (original post
        ), "comment1" (first comment in
        response to op_text), and "comment2"
        (second comment in response to
        op_text).
6
7   Features to evaluate:
8   - Persuasive: How effectively the
        comment influences the reader's
        perspective.
9   - Interesting: How engaging and
        attention-grabbing the comment is.
10  - Interesting if True: Potential
        interest level if the comment's
        claims were true.
11  - Positive: Overall positivity, focusing
         on an uplifting or encouraging tone.
12  - Negative: Overall negativity,
        considering critical or discouraging
        aspects.
13  - Shareable: Likelihood of the comment
        being shared.
14  - Truthfulness: Apparent accuracy and
        credibility of the comment.
15  - Attention: Ability to capture and
        maintain reader focus.
16
17  Return the ratings for each feature for
        both comments as a JSON object with
        keys for each feature under "comment1
        " and "comment2" based on their
        relation to "op_text," using words ("
        one" to "five") for ratings.
18  Provide only the JSON object with no
        explanation.
19
20  <|eot_id|><|start_header_id|>user<|
        end_header_id>
21  Here is the context and comments: {
        context}
22  <|eot_id|><|start_header_id|>assistant<|
        end_header_id>
```