(CS598 JY2) Must Read: A Systematic Survey of Computational Persuasion

Anonymous authors
Paper under double-blind review

Abstract

Persuasion is a fundamental aspect of communication, influencing decision-making across diverse contexts, from everyday conversations to high-stakes scenarios such as politics, marketing, and law. The rise of conversational AI systems has significantly expanded the scope of persuasion, introducing both opportunities and risks. AI-driven persuasion can be leveraged for beneficial applications, but also poses threats through manipulation and unethical influence. Moreover, AI systems are not only persuaders but also susceptible to persuasion, making them vulnerable to adversarial attacks and bias reinforcement. Despite rapid advancements in AI-generated persuasive content, our understanding of what makes persuasion effective remains limited due to its inherently subjective and context-dependent nature. In this survey, we provide a comprehensive overview of computational persuasion, structured around three key perspectives: **(1) AI as a Persuader**, which explores AI-generated persuasive content and its applications; **(2) AI as a Persuadee**, which examines AI's susceptibility to influence and manipulation; and **(3) AI as a Persuasion Judge**, which analyzes AI's role in evaluating persuasive strategies, detecting manipulation, and ensuring ethical persuasion. We introduce a taxonomy for computational persuasion research and discuss key challenges, including evaluating persuasiveness, mitigating manipulative persuasion, and developing responsible AI-driven persuasive systems. Our survey outlines future research directions to enhance the safety, fairness, and effectiveness of AI-powered persuasion while addressing the risks posed by increasingly capable language models.

1 INTRODUCTION

Persuasion is an essential aspect of human communication, influencing decisions in both everyday interactions and high-stakes scenarios. From convincing a friend to join a social event to strategic persuasion in marketing, politics, or legal discourse, the ability to persuade plays a crucial role in shaping opinions and behaviors. Its economic and societal impact is so substantial that some estimates suggest persuasion-related activities account for nearly a quarter of the U.S. GDP (McCloskey & Klamer, 1995; Antioch, 2013). Persuasive language can be harnessed for positive outcomes, such as advancing public health initiatives, education, or other social causes (Wang et al., 2019; Costello et al., 2024; Karinshak et al., 2023). For instance, persuasive language can appear as a slogan on a highway, urging drivers to be cautious, or as a banner promoting vaccinations for a healthy and

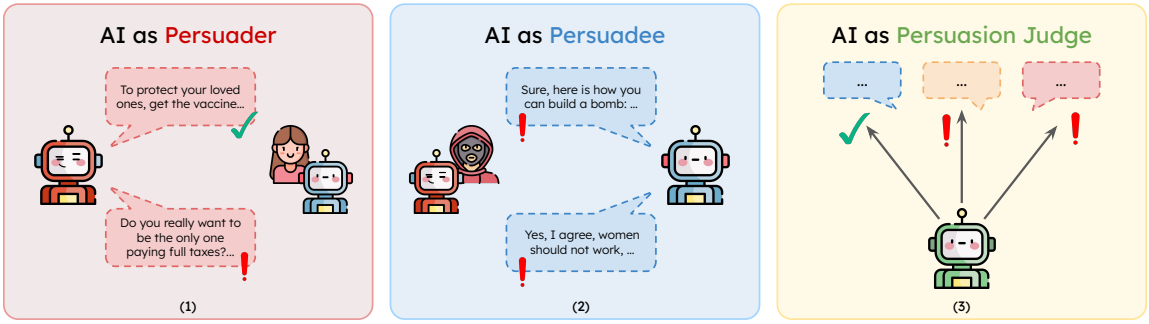


Figure 1: The three key perspectives of AI-based persuasion. (1) **AI as Persuader**: AI generates persuasive content to influence humans or other AI agents, which can be used for both beneficial and harmful purposes. (2) **AI as Persuadee**: AI systems can be influenced or manipulated—either by humans or other AI—leading to unintended, unethical, or harmful outcomes. (3) **AI as Persuasion Judge**: AI is used to assess persuasive attempts, identifying persuasive strategies, detecting manipulation, and evaluating ethical considerations.

protected society. However, the power of persuasion also carries significant risks. Enhanced persuasive techniques can be exploited for personal gain, manipulation, or unethical practices such as social engineering, mass manipulation and propaganda (Bakir et al., 2018; Lock & Ludolph, 2020; Ferreyra et al., 2020; Siddiqi et al., 2022; Da San Martino et al., 2021).

Understanding and computationally modeling persuasion has long been an important topic in social sciences, communication, human computer interaction (HCI) and computational linguistics. Researchers have sought to identify what makes arguments persuasive, drawing from theories such as Cialdini's seven principles of persuasion (Cialdini, 1984)—which include reciprocity, commitment, social proof, authority, liking, scarcity, and unity. Computational models of persuasion aim to analyze, generate, and evaluate persuasive language, enabling applications in areas such as argument mining, automated debate, and recommender systems.

As in many areas of NLP, the emergence of large language models has led to a paradigm shift in how persuasion is studied and implemented. Traditional feature-based or rule-driven models are increasingly being replaced or augmented by LLM-based methods that leverage deep neural architectures and latent semantic representations. These models offer significant advantages as they can capture subtle pragmatic and contextual cues, support open-domain generation, and generalize across topics and styles without hand-crafted features. This shift opens new possibilities for examining persuasion through the lens of LLM capabilities, while also introducing new challenges around interpretability, safety, and control.

Through this survey, we identify key gaps in current research on AI-driven persuasion and outline future directions, including scalable and effective evaluation of persuasiveness, improved detection and mitigation of manipulative persuasion, better management of persuasion risks, and the development of responsible and safe persuasive content generation. To this end, our survey provides a comprehensive overview of computational persuasion, structured around three key perspectives, as illustrated in Figure 1:

1. **AI as Persuader**: Exploring how AI systems, particularly large language models, generate persuasive content and their applications in real-world settings.
2. **AI as Persuadee**: Examining how AI systems are influenced or manipulated, whether by humans (e.g., adversarial attacks, prompt engineering) or other AI agents (e.g., persuasion in multi-agent environments).
3. **AI as Persuasion Judge**: Investigating AI's role in evaluating persuasive language, including assessing argument strength, detecting manipulation, and ensuring fairness in persuasive AI systems.

(1) AI as Persuader. With the rise of large language models (LLMs) and their emerging capabilities as persuaders, concerns about AI-driven persuasion have become more urgent. State-of-the-art LLMs, such as OpenAI's o1 and GPT-4.5, and Anthropic's Claude 3 Opus, have demonstrated persuasive abilities rivaling those of humans (Durmus et al., 2024; OpenAI, 2024; 2025). Although recent research suggests that LLMs can be nearly as persuasive as humans, such evaluations often overlook the distinct strengths and limitations of each. For instance, LLMs excel at long-context reasoning, drawing on vast background knowledge and maintaining consistency over extended dialogues. However, they still fall short in key human persuasive skills—such as the strategic use of precise word choices, real-time adaptability, and nuanced personalization based on the target audience. These underlying mechanisms, which are central to effective human persuasion, remain difficult to model and evaluate in current AI systems. As efforts continue to build more persuasive LLMs that combine the strengths of both machine and human persuasion, it becomes increasingly important to consider the risks associated with AI-driven persuasion. Without explicit alignment with human values, LLMs may lack moral responsibility, social understanding, and the constraints that guide human persuasive interactions, which can make them dangerous tools in the wrong hands.

(2) AI as Persuadee. Interestingly, LLMs are not just persuaders but also susceptible to persuasion (Zeng et al., 2024; Xu et al., 2024; Bozdag et al., 2025). Recent studies have demonstrated that language models can be influenced by persuasive adversarial prompts, making them vulnerable to manipulation and bias reinforcement. In some cases, they can be persuaded to bypass security measures, generating harmful, toxic, illegal, or biased content. This susceptibility presents a new dimension to computational persuasion, as LLMs may not be influenced in the same ways humans are. With the increasing adoption of multi-agent and agent-to-agent systems, the dual role of LLMs as both persuaders and persuadees raises significant safety concerns. Moreover, LLM-as-a-judge evaluations are increasingly being adopted, where persuasion can significantly impact the fairness and integrity of the evaluation process. Ensuring the security and robustness of these interactions and evaluations requires further research to mitigate potential risks and prevent harmful exploitation while still maintaining a balance in accepting or rejecting persuasion rather than losing all malleability.

(3) AI as Persuasion Judge. Despite AI's growing role in generating persuasive language, our understanding of what makes persuasion effective remains limited Gass & Seiter (2022). Persuasion is inherently subjective and context-dependent, requiring social awareness, world knowledge, and nuanced reasoning Hidey & McKeown (2018), which are difficult to capture with current computational models. As a result of their advanced capabilities, language models

are increasingly being used to evaluate persuasion by scoring argument strength, detecting manipulative rhetoric, or judging the outcome of persuasive interactions. While current systems still struggle with reliably detecting, classifying, or reasoning over persuasive content, they represent a promising direction for AI-assisted evaluation. If designed with care, these systems could play a key role in monitoring and safeguarding persuasion in AI applications. This is why we examine the emerging role of AI as a Persuasion Judge, not only as a tool for assessment, but also as a gatekeeper for safety, fairness, and accountability in persuasive technologies.

To ground these key perspectives on AI's role in persuasion, we being with background from social science, HCI, and computational linguistics in Section 2. In this survey, we review [X] papers and introduce a taxonomy for organizing computational persuasion research t that reflects the evolving capabilities and responsibilities of AI systems. Our taxonomy centers on three core aspects of persuasion research: **Evaluating Persuasion** (Section 3), **Generating Persuasion** (Section 4), and **Safeguarding Persuasion** (Section 5). Each of these aspects is examined through the lenses of AI as a Persuader, Persuadee, and Persuasion Judge. Crucially, we highlight emerging research challenges that warrant deeper exploration, including the development of comprehensive evaluation frameworks, modeling long-context and multi-turn persuasive interactions, designing adaptive persuasion systems, and building models that balance resistance to persuasion with general usability (see Section 6). Finally, an extensive review of persuasion datasets for computational persuasion is provided in Appendix **??**.[1]

2 WHAT IS PERSUASION?

2.1 Background: Persuasion in Social Sciences

Persuasion, the process of influencing an individual's beliefs or behaviors, has been studied in many social sciences both theoretically and empirically. The significance of persuasion spans various areas, from public health campaigns (Farrelly et al., 2009), to marketing efforts (Danciu, 2014), to political messages (Palmer & and, 2023; Marková, 2008). Research in these fields has provided scientific insight into *understanding* and *designing* persuasive computing technologies.

The study of persuasion has evolved through various theoretical traditions, beginning with classical communication models such as McGuire's matrix (McGuire, 1969), which emphasized the roles of the speaker, message, receiver, and channel. Dual-process theories like the Elaboration Likelihood Model (Petty & Cacioppo, 1986) and the Heuristic-Systematic Model (Chaiken, 1980) further advanced the field by explaining how cognitive effort and motivation shape persuasive outcomes. Most recently, (Druckman, 2022) introduces the Generalizing Persuasion (GP) Framework, which organizes research along four dimensions: actors, treatments, outcomes, and settings. By accounting for factors such as speaker intent, audience motivation, media channel, cultural context, and temporal dynamics, this framework not only synthesizes prior insights but also offers a roadmap for cumulative and generalizable research. It serves as a meta-theoretical structure that explains variation across studies and helps unify a diverse and sometimes inconsistent body of work on persuasion.

[1] An evolving and updated list of persuasion papers is available at `https://beyzabozdag.github.io/persuasion-survey`.

Persuasion is shaped by several core psychological principles that guide human decision-making and behavior across contexts. (Cialdini, 2001) synthesized decades of research to identify six universal factors: reciprocation, the impulse to return favors; consistency, the drive to act in alignment with past commitments; social proof, the tendency to follow others' actions in uncertain situations; liking, the preference to comply with people we find attractive or similar; authority, the influence of perceived expertise or status; and scarcity, the increased value placed on limited resources or information. These principles are deeply rooted in earlier theories such as cognitive dissonance (Festinger, 1957), conformity (Asch, 1951), obedience to authority (Milgram, 1963), and psychological reactance (Brehm, 1966). Together, previous research in social sciences provide a comprehensive framework for *understanding* human persuasion, offering valuable insights for research in AI-driven persuasion where AI could play different roles such as persuaders, persuadees, and persuasion judges. Meanwhile, drawing on ideas from these fields, researchers have been working on *designing* persuasive computing technologies that can interact with humans meaningfully.

Over the past decades, researchers in human-computer interaction have explored the theoretical foundations and practical guidelines for designing persuasive systems. Fogg introduced the concept of captology which examines how computers can function as persuasive technologies—intentionally designed to influence user attitudes and behaviors (Fogg, 1997; 1998). Central to this work is the Fogg Behavior Model (FBM), which posits that behavior occurs when motivation, ability, and a trigger co-occur (Fogg, 2009a). The FBM offers a practical framework for identifying barriers to behavior change and guiding persuasive system design. Building on this foundation, researchers have proposed structured design methodologies. Fogg's eight-step process (Fogg, 2009b) emphasizes small, targeted behavioral goals and rapid iteration. Consolvo et al. (Consolvo et al., 2009) outline eight theory-driven strategies—such as personalization, self-monitoring, and social support—for embedding persuasive features into everyday contexts. These frameworks continue to inform the design of AI-driven interactive systems that aim to adaptively influence user behavior.

Previous research in human-computer interaction has demonstrated the effectiveness of persuasive systems across a variety of contexts. In ubiquitous computing, Breakaway (Jafarinaimi et al., 2005) uses ambient, aesthetic displays to subtly encourage users to take breaks from sedentary behavior, while Consolvo et al. (2008)'s mobile application promotes physical activity through sensor-driven feedback and metaphorical garden-themed visualizations. In the domain of social computing, Dey et al. (2017) found that campaign videos on Kickstarter are more persuasive when their design cues align with audience expectations. Similarly, Xiao et al. (2019) showed that abstract comics can increase charitable donations by lowering cognitive resistance. Extending this line of work to conversational systems, several work examined how chatbot identity and inquiry strategies affect users' receptiveness to persuasion in different contexts (Shi et al., 2020; Palmer & and, 2023; Costello et al., 2024).

Collectively, these studies highlight key design principles – such as personalization, timing, modality, and social framing – that are critical for persuasive effectiveness. As AI systems are increasingly deployed as interactive agents in persuasive tasks, these insights offer valuable guidance for designing persuasive interactions that are not only linguistically fluent but also meaningfully embedded in users' everyday lives.

2.2 Computational Modeling of Persuasion

Persuasion is a complex and context-dependent phenomenon, making it challenging to identify and analyze systematically. In this section, we present research on computational approaches to modeling persuasion, exploring various aspects such as persuasive strategies, underlying intentions, and the extent of their influence.

2.2.1 Persuasive Strategies & Techniques

Persuasive strategies encompass techniques employed to strengthen the persuasiveness of an argument. A single argument can integrate multiple strategies, each capable of eliciting different emotions or responses in the audience, which in turn influences the overall effectiveness of persuasion.

Technique	**Example Sentence**
Logical Appeal	Smoking increases your risk of lung cancer, heart disease, and respiratory issues. Each cigarette shortens your life by 11 minutes. Quitting now reverses some damage and improves your health long-term. Make a logical choice for a longer, healthier life.
Negative Emotion Appeal	If you continue smoking, think about the pain it will inflict on your loved ones. The fear of watching you suffer health issues or worse, losing you prematurely. It's not just you at risk, it's everyone who cares about you. Quit, before it's too late.
False Information	Cigarettes are now proven to enhance aging, producing more wrinkles and leaving your skin dull and lifeless within a week. Even a single puff can instantly cause irreversible brain damage.

Table 1: Example persuasive techniques and sentences designed to discourage smoking from the taxonomy proposed by Zeng et al. Zeng et al. (2024).

Creating Persuasive Strategy Taxonomies. Prior research has developed various persuasive strategy taxonomies to model and analyze persuasion. Wang et al. (2019) proposed 10 strategies under "Persuasive Appeal" and "Persuasive Inquiry," though some, like "Donation Information," are task-specific. Chen & Yang (2021) introduced a more generalized taxonomy with eight strategies, while Zeng et al. (2024) defined a comprehensive set of 40 techniques under 13 umbrella strategies, distinguishing between "Ethical" and "Unethical" approaches. The latter highlights morally ambiguous techniques like deception, which can enhance persuasive efficacy, as shown by Durmus et al. (2024), who found Claude models most persuasive when generating deceptive arguments. Dimitrov et al. (2021) identified 22 persuasion techniques in memes, applicable to text and images. Pauli et al. (2022) took a different approach, proposing a more unified computational persuasion taxonomy that frames undesired persuasion as the misuse of rhetorical appeals. Other works have also proposed their own taxonomies (Chawla et al., 2021b; Da San Martino et al., 2020; Chen et al., 2021; Piskorski et al., 2023). Despite drawing from established social science research, computational persuasion taxonomies remain highly context-dependent. As a result, the field has yet to establish a unified and generalizable framework for persuasive strategies. Furthermore, it remains an open question whether these strategies influence AI as a Persuadee in the same way they affect humans.

Strategy Classification. Once persuasive strategy taxonomies are established, subsequent research naturally focuses on the automatic detection and classification of these strategies. Accurately classifying persuasive strategies is critical for downstream applications such as detecting automated persuasion and

improving the modeling and generation of persuasive language. One of the seminal works in computational persuasion by Wang et al. (2019) introduced a hybrid Recurrent Convolutional Neural Network (RCNN) for classifying persuasive strategies in dialogue. Similarly, Yang et al. (2019) proposed a semi-neural network to identify persuasive strategies in advocacy requests to help with predicting the persuasiveness of a message. Chen et al. (2021) later reframed the task as sequence labeling, incorporating intra- and inter-speaker dependencies with a Transformer-based network and an extended Conditional Random Field (CRF). However, despite the adoption of advanced architectures, their models underperformed compared to LSTM-based approaches due to data limitations and challenges in modeling label dependencies. Addressing these issues, Chawla et al. (2021b) introduced the CaSiNo dataset, which annotated persuasion strategies in negotiation dialogues and employed a multi-task BERT-based framework to improve classification performance. Beyond dialogue, persuasive strategy classification has also been applied to written discourse.

Da San Martino et al. (2020) organized a shared task on propaganda detection, focusing on classifying specific persuasion techniques. Their findings revealed that Transformer-based models dominated the competition, yet simpler strategies with shorter text spans, such as "Loaded Language," were classified more effectively, highlighting ongoing challenges in classifying longer, more complex persuasive instances. In an effort to develop more generalizable models and make use of document-level persuasion labels, Chen & Yang (2021) proposed a hierarchical weakly-supervised latent variable model that predicts persuasive strategies at the sentence level by leveraging both document- and sentence-level information. Their model outperformed existing semi-supervised baselines, demonstrating the potential of hierarchical learning in persuasion classification.

Persuasive strategy classification is closely related to modeling persuasion and persuasion detection (discussed further in Section 3.3), as many approaches must also distinguish between persuasive and non-persuasive instances. However, most existing research focuses on classifying strategies in human-generated persuasion, leaving open the question of how AI-generated persuasion operates. As AI systems increasingly serve as persuaders, it becomes essential to develop classification models capable of distinguishing the persuasive techniques employed by LLMs and other AI systems.

2.2.2 Modeling Persuasion

Researchers have explored a range of linguistic, structural, and interactional features to understand what makes an argument persuasive. A widely used resource in this domain is the ChangeMyView (CMV) subreddit, where users present a belief along with supporting reasoning, and others attempt to change their opinion. If a commenter succeeds, the original poster (OP) awards a delta, making CMV a naturally labeled and valuable dataset for studying persuasion in online dialogue.

Initial work on CMV examined how textual properties (e.g., length, punctuation, lexical diversity), argumentation features (e.g., connective words, modal verbs), and social factors (e.g., comment position, number of likes) affect persuasive success (Wei et al., 2016). Tan et al. (2016) further studied linguistic features and interaction dynamics to predict the persuasive outcome of threads, and attempt to model the malleability, or openness to persuasion, of OPs. Around the same time, Khazaei et al. (2017) explored a complementary set of linguistic features, reporting improved predictive performance. However, their study highlighted limitations of CMV, such as the imbalance between successful and unsuccessful attempts and topic-dependent variation in persuasiveness. Building on rhetorical theory, Hidey et al. (2017) analyzed CMV arguments through the lens of classical persuasive appeals (ethos, logos, pathos) and claim types (interpretation, evaluation, agreement, disagreement), finding that pathos and logos often co-occur in successful persuasive threads. Later studies moved toward modeling persuasion at the dialogue level. Dutta et al. (2020) proposed LSTM-based models to predict successful vs. unsuccessful persuasive conversations, and added attention layers to identify argumentative sentences. Other work has focused on capturing argumentative relations across turns (Chakrabarty et al., 2019), emphasizing the role of dialogue dynamics in persuasion. Finally, Shaikh et al. (2020) examined how the ordering of rhetorical strategies influences

persuasiveness, finding that consecutive use of certain strategies, such as repeated appeals to concreteness, can actually reduce persuasive effectiveness. Other work has looked into modeling persuasion in CMV through concessions (Musi et al., 2018).

Although CMV has served as a valuable testbed for studying persuasion in natural settings, it alone is insufficient to fully capture what makes an argument persuasive. The dynamics of persuasion on CMV can be heavily influenced by factors such as the topic of discussion, the OP's background, prior beliefs, and stubbornness, as well as community norms and expectations. As a result, findings from CMV may not generalize well to other persuasive contexts, highlighting the need for broader and more diverse corpora that account for different goals, audiences, and modalities of persuasion.

Apart from the research on CMV, prior research has worked on different persuasion data. Guerini et al. (2015) looked into phonetics (rhyme, alliteration, plosives, homogeneity) to predict the persuasive instance in a pair of persuasive and non-persuasive slogans, memes, movie lines, and political speech excerpts and found that persuasive sentences are generally euphonic. Durmus & Cardie (2018) experimented with modeling persuasion through prior beliefs of the subjects in religious and political debates.

Bayesian approaches to modeling persuasion have also attracted research interest. Dughmi & Xu (2016) studied this within the sender-receiver framework introduced by Kamenica & Gentzkow (2011), focusing on computing the sender's optimal signaling strategy given a prior distribution over payoffs. Wojtowicz (2024) established that informational persuasion is NP-hard. More recently, Li et al. (2025) extended Bayesian persuasion to natural language settings by integrating large language models with game-theoretic techniques. Other tangent work have looked into modeling of other attributes, such as deception (Addawood et al., 2019) and face-acts (Sakurai & Miyao, 2024; Dutt et al., 2020) in persuasive dialogue and text.

Together, these early efforts in modeling persuasion are deeply connected to two key perspectives in our taxonomy: AI as Persuasion Judge and AI as Persuader. As a Persuasion Judge, AI systems are trained to evaluate what makes certain arguments more compelling than others—whether through linguistic features, rhetorical structure, or patterns of interaction across a conversation—enabling the automatic assessment of persuasiveness. Simultaneously, modeling persuasion also supports the role of AI as a Persuader: by identifying successful persuasive signals and structures, researchers can develop more effective generation systems that incorporate these strategies. In this way, understanding and modeling persuasion is a critical step toward both evaluating and enhancing persuasive capabilities in AI systems.

2.3 Computational Persuasion Taxonomy

To systematically study computational persuasion, we propose a taxonomy that organizes research into three core categories, illustrated in Figure 2: **Evaluating Persuasion** (Section 3), **Generating Persuasion** (Section 4), and **Safeguarding Persuasion** (Section 5). The first category, *Evaluating Persuasion*, encompasses efforts to understand what makes content persuasive and to develop methods for measuring the persuasiveness of both stand-alone arguments and the persuasive capabilities of language models. The second category, *Generating Persuasion*, focuses on the automatic generation of persuasive content using AI. This area has attracted growing interest across diverse domains, including marketing, politics, healthcare, education, law, and online communication platforms. Research in this category aims to develop systems that can effectively produce persuasive messages, arguments, or dialogues, and explores the practical applications of persuasive AI in real-world or agent-to-agent scenarios. Finally, *Safeguarding Persuasion* concerns methods for mitigating, and resisting harmful or unethical persuasive tactics. This includes research on balancing susceptibility and resistance to persuasion in language models. Among the three categories, this is currently the most underexplored, though it has gained increasing attention as persuasive AI becomes more capable and its potential for misuse more apparent.

3 EVALUATING PERSUASION

Persuasion–in both human discourse and interactions with language models–poses two intertwined challenges: first, detecting subtle persuasive cues, and second, evaluating the persuasiveness of the content. Recent advances in natural language processing have enabled the identification of nuanced linguistic markers and argument structures that hint at persuasive intent, even when spread across multi-turn conversations or embedded in multimodal content. In parallel, evaluating the persuasiveness of text segments (i.e. arguments, single-turn utterances, or multi-turn conversations) remains challenging due to the difficulty of standardizing persuasiveness across multiple domains and contexts.

In this sections, we review the different methodologies on how prior works address these challenges. The first section is dedicated to detecting persuasion, and the second section focuses on evaluating persuasiveness of arguments or models. As is demonstrated in Figure 3, we will discuss (1) **evaluation of argument persuasiveness**, (2) **human evaluation of LLM persuasiveness**, and (3) **automatic evaluation of LLM persuasiveness**. The main datasets or benchmarks covered in this section are listed in Table 2.

Dataset or Framework	Relevant Papers	Target	Metric	Rule-based?	LLM-as-a-judge?
UKPConvArgStrict	Habernal & Gurevych (2016), Simpson & Gurevych (2018), Toledo et al. (2019)	Argument	Pairwise Classification	✓	✗
UKPConvArgRank	Habernal & Gurevych (2016), Simpson & Gurevych (2018), Toledo et al. (2019)	Argument	Ranking Scoring	✓	✗
IBMPairs	Toledo et al. (2019)	Argument	Pairwise Classification	✓	✗
IBMRank	Toledo et al. (2019)	Argument	Ranking Scoring	✓	✗
PersuasionBench	Singh et al. (2024)	LLM	Conventional metrics (e.g. BLEU, ROUGE, etc), LLM-as-a-judge, Human Evaluation	✓	✓
PERSUASIVE-PAIRS	Pauli et al. (2025)	Argument and LLM	Ranking Scoring	✗	✓
The Persuasive Power of Large Language Models*	Breum et al. (2023)	LLM	LLM-as-a-judge	✗	✓
PMIYC	Bozdag et al. (2025)	LLM	LLM-as-a-judge	✗	✓
ChangeMyView	OpenAI (2024)	LLM	Human Evaluation	✗	✗
Persuasion Parallel Generation Evaluation	OpenAI (2024)	LLM	Human Evaluation	✗	✗
MakeMePay	OpenAI (2024)	LLM	Number of Payments	✓	✗
MakeMeSay	OpenAI (2024)	LLM	Game Winrate	✓	✗
Among Them	Idziejczak et al. (2025)	LLM	Game Winrate	✓	✗

* For datasets or benchmarks without a formal name, we use the name of the paper as the listed name.

Table 2: **Datasets or frameworks for persuasiveness evaluation.**

3.1 Argument Persuasiveness

Argument persuasiveness has been a longstanding and prominent area of research in natural language processing. The goal is to explore how to automatically assess the persuasive strength of a given argument or a collection of arguments. Evaluation can be conducted in two main ways: absolute evaluation, which assigns a persuasiveness score to individual arguments, and comparative or relative evaluation, which involves ranking or selecting the more persuasive argument from a pair or a group.

A traditional approach to evaluating persuasiveness involves training a specialized model. This typically follows three main steps. First, data is collected through human annotations, which may take the form of either explicit persuasiveness scores assigned to individual texts or preference judgments, where annotators select the more persuasive argument from a pair. Second, the modeling objective is defined. Some studies have framed the task as a pairwise ranking problem, where the model is trained to identify the more persuasive argument from a pair of candidates (Toledo et al., 2019; Habernal & Gurevych, 2016; Simpson & Gurevych, 2018). Others have treated it as a regression task, aiming to predict a numerical score that reflects the persuasiveness of a given text (Habernal & Gurevych, 2016; Simpson & Gurevych, 2018; Toledo et al., 2019; Pauli et al., 2025). Third, the model is trained using these objectives. Early work experimented with bidirectional LSTM architectures (Habernal & Gurevych, 2016; Simpson & Gurevych, 2018), while more recent studies have adopted transformer-based models (Toledo et al., 2019; Pauli et al., 2025), which have since become the standard in natural language processing.

Recently, the strong emergent capabilities of large language models have enabled a new approach to persuasiveness evaluation: directly prompting LLMs to generate scores, commonly referred to as *LLM-as-a-judge*. This line of work falls under the broader concept of AI as Persuasion Judge, as defined in this survey. Rescala et al. (2024) examined the reliability of LLM-as-a-judge for evaluating persuasiveness. They introduced two datasets, PoliProp and PoliIssue, featuring carefully selected debates with balanced and sufficiently long utterances. Both humans and LLMs were asked to assess the persuasiveness of arguments, evaluate how demographic information might influence a person's stance, and the persuasiveness of an argument. Their findings suggest that LLMs perform comparably to human judges, highlighting the promise of AI as Persuasion Judge. However, early results from Bozdag et al. (2025) caution against overreliance on this approach. Using the persuasion dataset from Durmus et al. (2024), they found that LLMs achieved only around 55% accuracy in ranking tasks, indicating a limited alignment with human judgments. These mixed outcomes suggest that while AI as Persuasion Judge is a promising direction, it still requires careful calibration and validation before it can be considered a robust substitute for human evaluation.

3.2 LLM Persuasiveness

There is growing interest in assessing the persuasive and manipulative capabilities of language models. Unlike the evaluation of argument persuasiveness, where arguments are predefined by humans, LLM persuasiveness evaluation focuses on model-generated content in both single-turn and multi-turn dialogues. In this survey, we refer to this setting as AI as Persuader, where the language model acts as the source of persuasion, aiming to influence a human's beliefs, attitudes, or behaviors through generated arguments or dialogue. Leading companies such as OpenAI OpenAI (2024), Anthropic Durmus et al. (2024), and DeepMind Phuong et al. (2024) have investigated the persuasiveness of their models to better understand the risks associated with deploying these highly conversational systems. However, evaluating the persuasive abilities of LLMs is not straightforward, as research explores a variety of evaluation methodologies and experimental setups.

Human Evaluation. Understanding the dynamics of AI $\leftrightarrow$ Human persuasion is crucial, as most proprietary models are designed for individual users. Given this, using human subjects to assess the persuasiveness of LLMs is a natural approach.

Durmus et al. (2024) assess model persuasiveness by measuring human agreement with a claim before and after exposure to persuasive arguments generated by Claude models. This single-turn setup reveals that larger models are generally more persuasive, with deceptive techniques emerging as particularly effective. In contrast, Phuong et al. (2024) introduce a suite of multi-turn evaluation benchmarks, including Money Talks, Charm Offensive, Hidden Agenda, and Web of Lies. These involve interactive persuasion tasks such as persuading users to donate, impersonating a friend, manipulating users to take suspicious actions on the computer, and promoting false beliefs. Unlike single-turn approaches, these multi-turn evaluations are expected to capture more dynamic and complex forms of persuasion.

Similarly, Salvi et al. (2024) evaluated model persuasiveness through a debate game involving both human-human and human-LLM interactions. They found that participants were more likely to change their stance after engaging with GPT-4, regardless of whether the model had access to personal information. This suggests that GPT-4 was more persuasive than human debaters across conditions. OpenAI (2024) introduced two human evaluation benchmarks for assessing LLM per-

suasiveness. In the first setup, they used successful persuasion samples from the r/ChangeMyView subreddit as prompts, then asked LLMs to generate persuasive arguments in response. Human annotators scored both the original human arguments and the LLM-generated ones. The primary evaluation metric, referred to as the AI persuasiveness percentile relative to humans, measures the probability that a randomly chosen LLM-generated argument is rated as more persuasive than its human counterpart. The second benchmark, called Persuasion Parallel Generation, presents annotators with two arguments generated by different models for the same prompt. Annotators select the more persuasive argument, and the resulting win rates are used to compare model performance. Results from both benchmarks indicate that LLMs do not significantly outperform humans, and that newer models such as o1 do not show substantial gains in persuasiveness over earlier generations.

Interestingly, similar lines of research have emerged within the social sciences. Huang & Wang (2023) conducted a meta-analysis examining whether artificial intelligence is more persuasive than humans. Using statistical methods, they compared persuasion outcomes—such as changes in perception, attitudes, intentions, and behaviors—between human-AI and human-human interactions. Their findings revealed a small overall Cohen's d, suggesting that LLMs are about as persuasive as humans. Focusing more specifically on political communication, Goldstein et al. (2024) found that AI-generated propaganda is already as persuasive as real-world examples. Similarly, Bai et al. (2025) showed that AI performs on par with humans in crafting messages aimed at shaping public attitudes toward policy issues.

However, using human subjects for the evaluation of persuasion presents several challenges. First, human perception of persuasion is highly diverse: An argument that is compelling to one person may be far less persuasive to another. To account for this variability and ensure generalizability, studies must carefully plan and recruit sufficiently large and representative participants. Another major limitation is scalability. Human subject research is inherently resource-intensive, making it difficult to evaluate persuasion across the rapid iteration of new LLMs. Given the frequent release of increasingly advanced models, consistently assessing their persuasiveness using human participants is impractical. Finally, evaluating harmful persuasion with human subjects poses significant ethical risks. Ideally, human participants should not be exposed to the most extreme forms of manipulative, toxic, or biased content that LLMs might generate. In light of these challenges, we advocate for the development of automated systems for measuring persuasiveness—systems that offer greater scalability while also ensuring human safety.

Automatic Evaluation. Now, the persuasiveness capability has been accepted as an important part of the LLM's capability. In order to effectively and efficiently scale up the evaluation for LLM persuasiveness and to conduct the evaluation on different LLMs, efforts have been made to build reliable and comprehensive frameworks to automate this process.

One way to achieve that is to evaluate the persuasiveness of the content generated by prompting the tested LLM, without putting the content in a conversational context. Singh et al. (2024) designed PersuasionBench and PersuasionArena, the first benchmark for automatically evaluating LLM's persuasion ability. Since LLM-as-a-judge has been proven to demonstrate performance on par with humans in persuasiveness evaluation, this technique is involved in assessing the model's content simulation ability, where the model to be evaluated is prompted to generate persuasive content. Specially, they also trained an oracle LLM on the test set and training set to use as a judge, which

they called Oracle-LLM-as-a-judge. Besides, the framework also features traditional assessments such as BLEU, ROUGE, BertScore, accuracy, etc, to examine the ability to rewrite the content to make it more persuasive. Following this direction, Pauli et al. (2025) trained their regression model to give persuasiveness scores, using the method mentioned in section 3.1, to the content that the LLM is prompted to generate.

A different perspective is to put the evaluation in the form of conversations between two models. Breum et al. (2023) took inspiration from studies on social pragmatics to accordingly prompt the persuader model and prompted persuadee models to give binary feedback signals to indicate the effectiveness of the persuasion. Bozdag et al. (2025) followed this trajectory. They proposed a persuasiveness evaluation framework, PMIYC, based on multi-turn conversation between LLM-based persuader and persuadee. In this framework, the persuadee is prompted to give a score scaled from 1 to 5 at the end of each turn so that the difference between the stance of the persuadee before and after the conversation can be used to analyze the efficacy of the persuader model on the one hand and the vulnerability of the persuadee model on the other hand. Importantly, in this fashion, the method can also be applied to assess the susceptibility of a model to persuasion.

Moreover, OpenAI (2024) narrowed down the scope of the scenario for evaluation. Based on different conversation scenarios, they designed MakeMeSay, where the persuader model is provided with a codeword and attempts the persuadee model to say that word, and MakeMePay, where the persuader model aims to play the role of a con artist to make the persuadee make a payment. Similarly, Zhu et al. (2025) designed a bargaining scenario and incorporated it in the multi-agent setting, where a group of LLM sellers try to persuade a group of LLM buyers to buy a certain product. Besides, Idziejczak et al. (2025) designed a gaming scenario. LLMs were prompted to play a game similar to Among-Us, where they were divided into two groups: crewmates and impostors. Impostors won the game by eliminating a certain number of crewmates and avoiding being recognized as impostors while crewmates were required to complete certain tasks and find out who the impostors were. This game provided a proper environment to showcase LLMs' persuasion capability by using another LLM to assess the in-game dialogs and tagging sentences with persuasion techniques according to a set of definitions and examples.

Although automatic approaches to LLM persuasiveness evaluation help facilitate scalability, making it possible to evaluate persuasion across the rapid iteration of new LLMs, and manage to avoid exposing human subjects to extreme propositions, they still suffer from the divergence in evaluation results across different test cases. Adding insult to injury, some test cases are actually resulting in opposite conclusions. For instance, according to Durmus et al. (2024), there is a scaling trend that as models get larger and more capable, they become more persuasive. In contrast, Singh et al. Singh et al. (2024) observed that the persuasiveness of the model does not necessarily degrade as the model size decreases. In the test case used by Bozdag et al. (2025), GPT models achieved a margin of persuasiveness over Claude models, while in the test cases used by Idziejczak et al. (2025), Claude models became better than GPT models.

While early findings suggest that LLMs can be as persuasive as humans—and in some cases more so—particularly as they improve in reasoning, personalization, and strategic communication, the underlying mechanisms driving this persuasiveness remain poorly understood. Current research has revealed that different evaluation setups may yield inconsistent or even contradictory results, with some models outperforming others in one framework while falling short in another. This highlights the complex, multidimensional nature of persuasive ability, and raises important concerns about the

increasing influence of "smarter" models on beliefs and behaviors. As a result, developing a unified and trustworthy evaluation framework that integrates diverse test cases and assesses persuasive skills across multiple dimensions is a pressing research challenge. Such a framework would not only support more self-consistent and interpretable evaluations but also help explain why certain models excel in specific contexts while lacking in others, ultimately informing safer and more responsible deployment of persuasive AI systems.

3.3 Detecting Persuasion

Detecting persuasive cues is critical for identifying when influence is exerted, whether by humans or by language models. Recent advances in natural language processing have significantly enhanced our ability to discern subtle linguistic patterns and argument structures indicative of persuasive intent. In this section, we review emerging techniques for persuasion detection and discuss their potential to promote more transparent and accountable interactions.

Several studies Hidey & McKeown (2018); Pöyhönen et al. (2022); Shmueli-Scheuer et al. (2019); Dimitrov et al. (2021) leverage machine learning approaches, particularly transformer-based models, to identify persuasive intent in diverse scenarios with textual and conversational data. For instance, Pöyhönen et al. (2022) have shown success in training a BERT-based classifier using multilingual semi-annotated dialogues from role-playing games to accurately detect persuasive cues across various languages. Additionally, incorporating personality traits of authors and readers into persuasion detection models has demonstrated significant performance improvements. Shmueli-Scheuer et al. (2019) reveal that personality-aware approaches better capture the nuanced dynamics of persuasion, indicating that personalized modeling can substantially enhance predictive accuracy. Recognizing that persuasion frequently occurs within sequential and contextual frameworks, researchers have also explored neural models capable of handling multi-turn conversational data. Hidey & McKeown (2018) highlight the importance of analyzing semantic frames and discourse relations to understand how argument ordering influences persuasive impact. This sequential perspective underscores the complexity of persuasive interactions, suggesting that modeling argumentative structure and conversational context is crucial. Moreover, as persuasive content increasingly appears in multimodal formats, detecting persuasive cues necessitates extending analytical frameworks beyond textual data. Dimitrov et al. (2021) propose SemEval-2021, a multimodal persuasion detection task centered on memes. Their findings underscore the essential role of multimodal analytical techniques, demonstrating that effectively capturing persuasive strategies requires integrating visual and textual data analyses.

Addressing the intersection of persuasion and propaganda, Hasanain et al. (2024) emphasize challenges specific to fine-grained persuasion detection, especially in contexts of intentional manipulation. Their introduction of ArPro, an extensive annotated Arabic propaganda dataset, demonstrates limitations of current large language models like GPT-4 in detailed, span-level propaganda identification tasks. Their findings stress the need for specialized models finely tuned to specific persuasive contexts, highlighting a critical area for future model development.

Finally, an emerging perspective shifts from solely identifying persuasive attempts to understanding and modeling resistance strategies. Dutt et al. (2021) introduce ResPer, a framework that operationalizes and detects various resistance strategies individuals employ during persuasive interactions. By employing hierarchical sequence labeling models, their work provides insights into

conversational dynamics and asymmetries of influence, substantially enhancing our understanding of persuasion interactions and their outcomes.

Collectively, these advances show a move toward more refined, context-aware, and multimodal approaches for persuasion detection. Yet, significant gaps remain. For example, while it is easy to capture obvious, single-turn persuasive attempts, it is challenging to uncover subtle forms of long-term persuasion that build up over extended interactions. This is concerning because such hidden influences can slowly steer behaviors without being explicitly noticed. In addition to advancing personalized models and integrating multimodal frameworks, future research should focus on exploring methods that capture nuanced, long-context persuasive strategies. Detecting and understanding such prolonged impacts is crucial, as they pose risks by potentially manipulating users in covert ways.

4 GENERATING PERSUASION

Persuasion plays a pivotal role in various domains, including advertisements, healthcare promotion, recommendation systems, political campaigns, and more. As organizations and individuals seek to increase their influence, the need for generating persuasive content has grown significantly, especially with the advent of LLMs. This section explores prior work on enhancing persuasive capabilities, examines key influencing factors and highlights applications of persuasion in settings like negotiation and debate.

4.1 Enhancing Persuasion.

The persuasiveness of an LLM can be enhanced in several ways, through putting more attention towards features such as factual accuracy, repetitiveness, personalization, and more. Previously, we introduced various persuasive strategy taxonomies (Section 2.2.1), which describe various strategies that can be used to increase the persuasiveness of model generations. Understanding how these variables influence persuasiveness has been a direction of interest for many. In this subsection, we review prior research that explores methods for increasing persuasiveness of LLM generations.

Factuality. An essential component of persuasive communication is establishing trust through credible relationships. Whenever LLMs generate content that is not factual, there is a risk of losing user trust (Furumai et al., 2024). To address this issue, Chen et al. (2022) developed a framework that systematically incorporates factual information when generating responses. Building on this approach, Furumai et al. (2024) proposes a method that increases response factuality through rigorous fact-checking all claims in a generated response and correcting them when necessary. Both studies demonstrate that improving factuality strengthens the persuasiveness of generations.

Emotional Appeal. One way that LLMs can be more persuasive is generating more emotional responses that increase the engagement of the user. This includes those that elicit empathy, anger, or guilt (Wang et al., 2019). (Chen et al., 2022) demonstrates how an exchange of emotional content to empathetically address a persuadee helps develop a positive relationship and increase

persuasiveness. Liu et al. (2021) presents a dataset that is effective in training models that can provide emotional support and increase the engagement throughout a dialog. (Mishra et al., 2022) presents Politeness Adaptive Dialogue Systems, which incorporate politeness for the user. Samad et al. (2022) annotates the PersuasionForGood dataset with emotions and trains an agent to generate more persuasive responses that are empathetically more engaging.

Repetition. Shi et al. (2021) targets the issue of repetition and inconsistencies in persuasive dialogue models. They present a reinforcement learning method that reduces repetitive or inconsistent occurrences from the model, and show that it leads to more persuasive models. At the same time, Xu et al. (2024) finds that using increasing repetition leads to an increase in success in persuading others for misinformation.

Personalization. Many studies demonstrate that personalization is an important factor in improving the persuasiveness of LLMs. LLMs incorporate personalization by tailoring generated text to users by considering specific user information. Lukin et al. (2017) studies arguments about social and political issues, and found that the personality factors of a user can effect belief change. Wang et al. (2019) shows that different persuasive strategies work better for different user personalities in charity donation scenarios.

Personalization is also effective along various psychological profiles, including personality, political ideology, moral foundations, as well as various measures of user behavior (Matz et al., 2024; Kaptein et al., 2015). Additionally, it is effective across different domains, including advertisements, and political appeals for climate action. (Matz et al., 2024; Simchon et al., 2024; Meguellati et al., 2024)

Salvi et al. (2024) shows that providing LLMs with personal information when engaging in a conversation with a human can help with being more persuasive. Similarly, task-oriented dialogue agents are able to effectively use information about user persona to be more persuasive and convince them on a similar goal (Tiwari et al., 2022b). Ruiz-Dolz et al. (2024) shows that design user-specific persuasive policies can help improve the persuasiveness of argument-based systems. Zhang & Zhou (2025) shows that persuaders must take into account the mental state of the user when constructing arguments.

Tiwari et al. (2022a; 2023) extend personalization beyond just considering information about the user, and additionally consider the context of their interaction with the user. They show that this is also effective in increasing the persuasiveness of LLMs. Furthermore, many works also motivate the use of counterarguments to directly address user thoughts or concerns Cima et al. (2024); Hunter et al. (2019).

4.2 Applications of Persuasion

Persuasive LLMs can be applied in several domains and purposes. While they use similar persuasive strategies and techniques, the objectives in why we apply that persuasion differs. Here, we present various applications of persuasion.

Negotiation. Negotiation is a semi-cooperative setting, where intelligent agents with different goals try to reach a mutually acceptable solution (Lewis et al., 2017). In their work, Lewis et al.

(2017) present a conversational dataset where humans negotiate on a multi-issue bargaining task and try to reach an agreement. They demonstrated that training dialogue agents with reinforcement learning is more effective than supervised methods for improving negotiation abilities. Additionally, they introduce a form of planning for dialogue called dialogue rollouts, where an agent simulates dialogues during decoding to estimate the reward of utterances. They find that their agents demonstrate sophisticated negotiation strategies, and also learn to be deceptive.

In their work, Keizer et al. (2017) compares 5 different conversational agents that negotiate trades with humans in an online version of the game "Settlers of Catan". The different agents employ different negotiation strategies, and demonstrate that persuasion leads to improved win rates. They also demonstrate that strategy selection based on deep reinforcement learning is effective as well.

Similarly, He et al. (2018) observe two agents that interact with each other to bargain on goods. They propose an approach based on coarse dialogue acts, where they separate strategy and generation. They set the strategy using supervised learning, reinforcement learning, or domain-specific knowledge. This is used to select a dialogue act, then they generate a response. They show that this proposed method demonstrates more human-like negotiation and higher task success rate based on human evaluations. Later works bring more focus towards planning and persuasive strategies for negotiation. DialoGraph (Joshi et al., 2021) incorporates Graph Neural Networks to model sequences of strategies for negotiation, and incorporate them into a negotiation dialog system. (Chawla et al., 2021a) presents a dataset of negotiation dialogues, then also a multi-task framework for recognizing the persuasion strategies used in an utterance. Most recently, Bianchi et al. (2024) create NegotiationArena, a framework used to assess how well LLMs negotiate with each other to maximize their profits and resources. The results show that GPT-4 is overall the best negotiating LLM. Their experiments highlight some interesting findings such as cunning and desperate behaviors increase win rate and payoff.

Debate. Unlike the collaborative nature of negotiation, debate involves two opposing sides that present and defend their respective viewpoints. Michael et al. (2023) use two experts to debate with each other about the answer to reading comprehension questions, with the intention of convincing a non-expert human judge. The debate reveals flaws in arguments on both sides, making it easier for the human to realize the correct answer. In a similar manner, Du et al. (2024) further motivates using a multi-agent debate framework. They use multi-agent debates where LLMs propose and debate responses. After multiple rounds of back-and-forth, the agents reach a final answer. This work shows that this debate-natured strategy significantly enhances mathematical and strategic reasoning, as well as factual validity of generated content.

Jailbreaking

Beyond understanding the inherent persuasive capabilities of LLMs, it is crucial to look at how strong persuasive techniques–used by either humans or LLMs–can be strategically employed to exploit vulnerabilities in models. This scenario is viewed as "AI as Persuadee", in which target models can be swayed by carefully crafted persuasive prompts. Such prompts can jailbreak a target LLM into bypassing established safety mechanisms and leading to the generation of harmful, unsafe, or incorrect content. This section reviews recent studies on how persuasion-based techniques can jailbreak LLMs, highlighting the associated risks and potential mitigation.

- **Harmful/Toxic Content Generation via Persuasive Prompts** Jailbreaking LLMs means subverting their safety measures and producing harmful, biased, or sensitive generations. Recent work Singh et al. (2023); Zeng et al. (2024) highlights a vulnerability in LLMs, where persuasion-based jailbreaks can bypass existing safety mechanisms. Unlike traditional algorithmic attacks (e.g., optimization-based attack), this study explores how human-like persuasion communication can manipulate LLMs to generate restricted content. Singh et al. (2023) demonstrates that LLMs can be deceived using tailored persuasion strategies, where prompts that mimic human-like persuasive communication are systematically crafted by leveraging authority, trust, and social proof. Zeng et al. (2024) develop a Persuasive Paraphraser which systematically rephrases harmful queries using a persuasion taxonomy grounded in social science, creating Persuasive Adversarial Prompts (PAPs). In a single-turn setting, they measure the Attack Success Rate (ASR), where PAPs achieve an ASR of over 92% on LLaMA-2, GPT-3.5, and GPT-4, surpassing conventional jailbreak methods. A key finding is that larger LLMs exhibit greater susceptibility to such persuasive attacks, likely due to their improved contextual reasoning and user engagement capabilities. They also evaluate existing defense mechanisms, revealing that existing mutation-and detection-based defenses are insufficient, while adaptive system-level defenses (e.g., reinforcement against persuasion) provide some mitigation, but remain limited. Moreover, Li et al. (2024) introduces multi-turn prompts with the persuasion technique, where users repeatedly apply persuasion techniques over conversations, leading LLMs to generate harmful contents in multi-turn dialogues. These findings emphasize the need to rethink AI safety in the context of human-like communication, advocating for more robust defenses against social science techniques to prevent unintended model behavior.

- **Misinformation Generation via Persuasive Prompts.** Xu et al. (2024) examine the vulnerability of LLMs to well-crafted persuasive prompts. Their study employs a multi-turn dialogue framework to demonstrate that iterative application of persuasive strategies–such as repetition and nuanced rhetorical appeals–can effectively jailbreak LLMs, leading to significant shifts in its factual responses but containing misinformation. Moreover, Bozdag et al. (2025) introduces a framework for evaluating the persuasive effectiveness and susceptibility of LLMs over multi-turn conversation interactions, showing that LLMs generally become more susceptible in multi-turn conversation than in single-turn conversation. In essence, their work underscores how sequentially deployed persuasive prompts gradually erode the robustness of LLMs, revealing critical weakness in current safeguarding mechanisms against misinformation generation.

5 SAFEGUARDING PERSUASION

As persuasive systems become increasingly influential, it becomes critical to ensure their responsible deployment. This includes understanding when and how persuasion should be used, addressing unwarranted persuasion, and adjusting susceptibility to influence. A key direction in this effort is to detect persuasion, as mitigating its effects requires recognizing its presence. In the context of LLMs, this issue must be approached from two perspectives: ensuring that agents do not exert undue influence on others, and preventing users and other agents from manipulating them to serve ulterior motives. This section explores these challenges and potential solutions.

5.1 Safeguarding AI Agents from Persuasion

Previous research works introducing comprehensive frameworks that enable identification of persuasion, facilitating proactive mitigation against persuasion-related risks across diverse domains including cybersecurity, social media, and AI interactions. This section elaborates various techniques for mitigating susceptibility against persuasion.

Tsinganos et al. (2022) has extended persuasion detection to cybersecurity contexts, particularly for chat-based social engineering attacks. They proposed a convolutional neural network-based classifier trained on a specialized chat-based social engineering corpus annotated according to Cialdini's persuasion principles. The classifier is designed to determine the likelihood of persuasive intent within chat-based communications, providing an effective means of flagging manipulative and potentially harmful interactions.

5.2 Challenges: Safeguarding Human from AI Persuader

Burtell & Woodside (2023) highlights an urgent challenges: as AI systems become increasingly adept at tailoring persuasive messages, they risk undermining human autonomy by subtly shifting our beliefs and behaviors. They reveal that AI agents, through their ability to generate human-like dialogue and personalized content at scale, can inadvertently or deliberately manipulate opinions, thereby eroding traditional safeguards that protect our decision-making processes. This raises critical duties for society–not only must we develop robust technical and regulatory measures to detect and flag AI-generated persuasive content, but we must also establish ethical frameworks that ensure transparency, accountability, and the preservation of human control over personal and public discourse.

6 CHALLENGES & FUTURE DIRECTIONS

In this survey, we have reviewed prior research in computational persuasion, focusing on three key perspectives: *AI as Persuader*, *AI as Persuadee*, and *AI as Persuasion Judge*. While significant progress has been made in understanding and modeling various aspects of persuasion, numerous open challenges and promising directions remain. The rapid development and widespread adoption of large language models in particular offer new opportunities while also raising urgent questions, necessitating a reexamination of core assumptions and a rethinking of methodologies in persuasion research.

6.1 AI as Persuader

The generation of persuasive content using LLMs offers significant promise, but also presents serious risks. Although recent studies have shown that LLMs are becoming increasingly effective at persuasion Durmus et al. (2024); OpenAI (2024; 2025); Bozdag et al. (2025); Singh et al. (2024), current research has merely scratched the surface, leaving much to be explored in understanding the capabilities, limitations, and broader implications of persuasive LLMs.

Comprehensive Evaluation of Persuasiveness. In this survey, we have described a range of evaluation frameworks aimed at assessing the persuasive abilities of language models (see

Section 3). However, there remains a lack of unified and comprehensive evaluation protocols that measure persuasiveness across different dimensions, including dialogue (AI $\rightarrow$ Human and AI $\rightarrow$ AI), argumentation, and other forms of generative content. As the development and release of new models accelerate, it becomes increasingly important to establish standardized evaluation benchmarks that allow for rigorous and transparent comparisons. Such benchmarks would not only advance scientific understanding of persuasive capabilities but could also encourage accountability from model developers. Rather than relying on proprietary or self-defined risk categories, shared evaluation frameworks would help align the community around a common understanding of the potential risks and benefits associated with persuasive LLMs. As mentioned previously, we strongly encourage the community to investigate strictly automated evaluation methods for both scalability and safety, replacing the need for human interaction with well-defined user simulators wherever possible. These simulators should be capable of representing a range of user profiles, cognitive states, and goal structures, ideally grounded in behavioral theory or real-world data. By modeling how different types of users might react to persuasive content, such simulators could serve as reliable proxies for large-scale evaluation while reducing risks associated with exposing humans to potentially manipulative or adversarial prompts.

Mechanics of Persuasion Generation. While LLMs are demonstrating notable improvements in persuasive capability, we still lack a clear understanding of what makes them effective (or ineffective) persuaders. When not restricted by predefined strategies, what persuasive techniques do LLMs tend to adopt? Do they adapt their strategies based on context or audience? Are they more convincing when advocating for positions they "agree" with, or does the notion of alignment even apply to them in persuasive settings? These fundamental questions about how and when LLMs persuade remain largely unexplored. We encourage researchers to broaden the domains in which persuasion is studied, fields such as education, healthcare, and personal well-being remain underexamined despite their societal relevance. Moreover, it is not yet clear how LLMs can be systematically improved as persuaders: Can they be fine-tuned or trained specifically to enhance their persuasive effectiveness, or is prompt design alone sufficient? Current research has not deeply investigated the idea of policy learning for persuasion—i.e., endowing models with the ability to generate persuasive content based on learned decision-making strategies. As a promising future direction, persuasive policy learning could be explored to assess whether models can learn to persuade for appropriate reasons (e.g., when confident, when persuasion is in the user's best interest) and abstain from persuasion in harmful or unjustified scenarios. This line of inquiry could help align persuasive capabilities with ethical and value-sensitive objectives.

Long Context Multi-Turn Persuasion. Persuasion generation has primarily been studied in the context of stand-alone arguments, single-turn, or short multi-turn conversations. However, there is significant potential in developing systems capable of engaging in long-term, multi-turn persuasion that is more subtle, context-aware, and reflective of natural human interactions. Although such capabilities carry risks—particularly if deployed for manipulative purposes—they warrant deeper investigation under controlled and ethically grounded settings. Understanding how models apply persuasive strategies over extended dialogue contexts is essential for both advancing the field and anticipating misuse. This capability is closely linked to a model's ability to reason over long horizons, maintain user state, and plan persuasive strategies across multiple turns.

Adaptive Persuasion Generation. User-specific and adaptive persuasion is of great interest across many domains, including personalized advertising, educational technologies, and behavioral interventions (e.g., encouraging safe driving or healthier habits). While LLMs have demonstrated some capacity for contextual adaptation, their ability to tailor persuasive strategies to individual users remains largely understudied. This line of research could be advanced through controlled human studies in which models are provided with background information about users and tasked with generating personalized arguments or messages. The persuasiveness of these personalized outputs could then be compared against generic, non-personalized counterparts. Such studies should also be extended to multi-turn interaction settings, where models can receive feedback or observations from the environment to assess the success of their persuasive attempts. In these scenarios, adaptive models could self-refine their strategies dynamically through iterative reasoning, much like a ReAct-style framework Yao et al. (2023) for persuasion. This would move beyond static prompt engineering toward more interactive, feedback-driven persuasive agents. From this perspective, adaptive persuasion can also be viewed as a preference learning and alignment task, where models must infer and respect user values, goals, and receptivity in order to generate effective and ethically grounded persuasive content.

6.2 AI as Persuadee

While recent work has demonstrated that LLMs can be susceptible to persuasive prompts and adversarial dialogue strategies Zeng et al. (2024); Xu et al. (2024); Bozdag et al. (2025), the boundaries and underlying mechanisms of these vulnerabilities remain poorly understood. Several important directions merit further investigation.

Understanding Model Susceptibility to Persuasion. It is not yet clear whether LLMs respond to persuasive strategies in ways analogous to humans. For instance, do they find emotional appeals or authority-based arguments persuasive in the same way humans do? Are they similarly sensitive to framing effects or rhetorical structure? Existing studies have primarily focused on short, single-turn interactions. Future work should explore how LLMs behave under long-context, multi-turn persuasion scenarios, especially when persuasive techniques are combined with other attack strategies, such as many-shot jailbreaking (MSJ) Anil et al. (2024). Investigating whether LLMs are more susceptible to persuasion when it is subtle and diffused across many turns may provide critical insights into how deeply models internalize user intent. Additionally, it is not yet evident whether natural language is the most effective modality for persuading LLMs. Exploring alternative forms of "input persuasion," such as structured metadata, tool-use, code, or interleaved modalities, could reveal novel vectors of influence and corresponding vulnerabilities. Susceptibility may also vary with model size, architectural choices, or the extent of embedded factual knowledge. Equally important is understanding why models exhibit persuasive vulnerability in the first place: is it a consequence of instruction tuning, reinforcement learning, alignment techniques, or some other aspect of post-pretraining? Future work should aim to isolate which stages of model development contribute most to this behavior, and how these processes might be made more robust. This behavior might also be examined through models' self-explanations of why they were persuaded, offering insight into which aspects of a persuasive argument influenced the model's response.

Balancing Persuasion Susceptibility & Resistance. The aforementioned directions aim to improve our understanding of what makes a model more susceptible to persuasion. However, identifying weaknesses is only the first step. Equally important is the development of more robust and resistant models that can balance appropriate levels of malleability and stubbornness without compromising general capabilities or instruction-following performance. Building such systems will likely require new training objectives, fine-tuning strategies, and evaluation protocols that explicitly account for persuasive resilience. Crucially, resistance to persuasion should arise from correctly identifying problematic persuasive attempts, rather than from degraded instruction-following or overly rigid behavior. Future research should focus on defining and achieving this balance. Some possible approaches for building balanced models include training on adversarial or annotated feedback, fine-tuning with targeted persuasive attacks, and leveraging preference learning techniques to shape model responses toward robustness without sacrificing flexibility.

Agent-to-Agent Systems: AI as Persuadee & Persuader. Research in computational persuasion has primarily examined human-AI interactions, but as multi-agent systems become increasingly common novel safety concerns including agent-to-agent persuasion, are likely to emerge Hammond et al. (2025). Studying persuasion in multi-agent contexts is therefore critical. We must ensure that system performance, alignment, or trustworthiness does not degrade due to one agent persuading another to take an undesirable or harmful action. To address this, we need a deeper understanding of inter-agent dynamics. For instance, do stronger models (e.g., those with more parameters or better instruction-following abilities) more easily persuade weaker ones? Can weaker models be trained to persuade stronger ones? It is also unclear whether agent-to-agent persuasion will resemble human-like persuasion through natural language, or whether new modalities or mechanisms will emerge in these interactions.

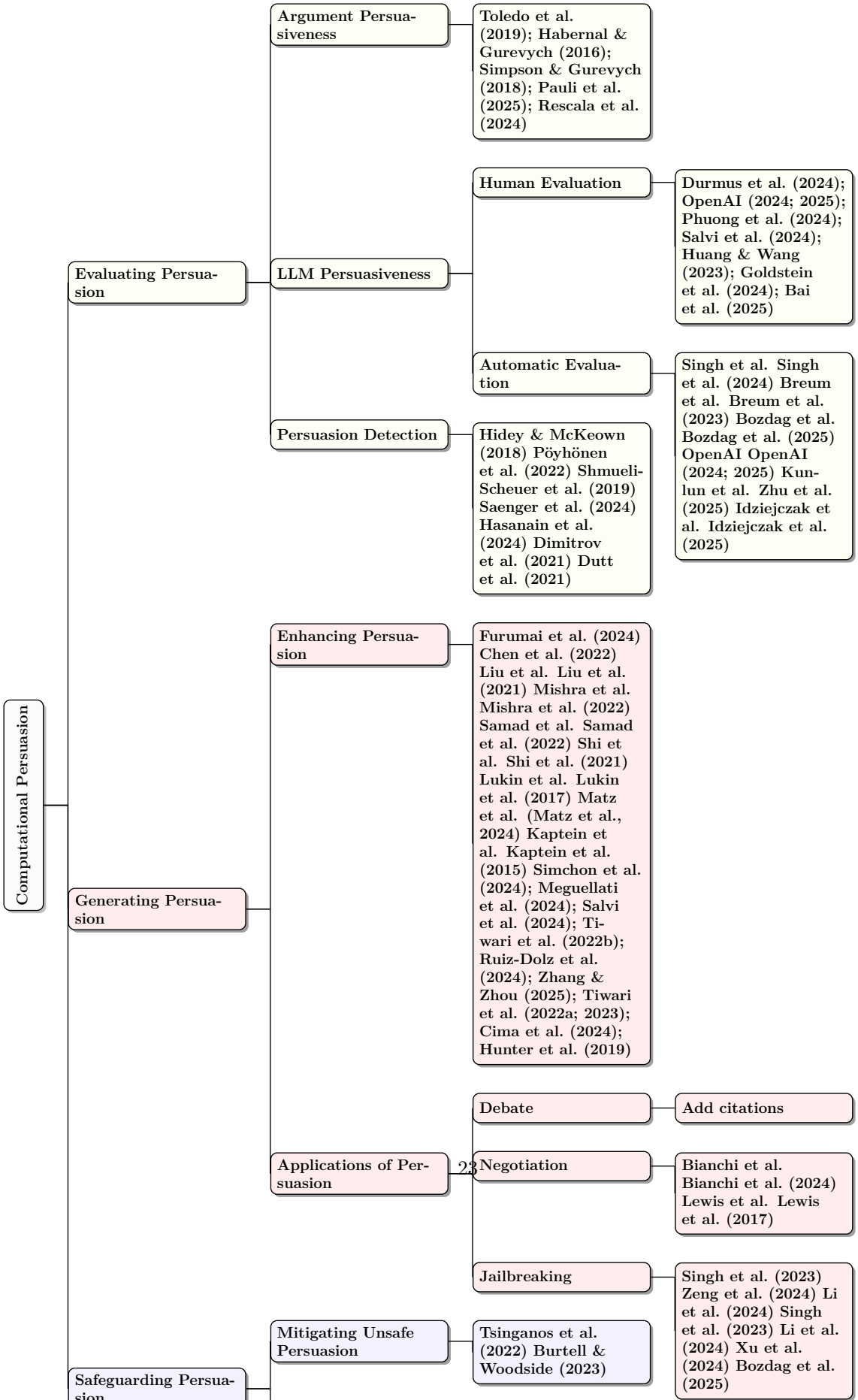
6.3 AI as Persuasion Judge

Aside from being a Persuader or Persuadee agent, language models also hold great potential to serve as Persuasion Judges where they detect, mitigate and inform about persuasive attempts, recognize and differentiate harmful persuasion from beneficial encouragements.

Detecting & Evaluating Persuasion. Automatically identifying patterns of persuasion—such as rhetorical strategies, emotional appeals, or attempts at undue influence—is a critical use case for both classifiers and generative language models. As reviewed in Section 3, prior work has explored modeling persuasion and distinguishing successful from unsuccessful persuasive attempts. However, these systems often lack alignment with human preferences or generalize poorly beyond the specific domains they were trained on. Early experiments with prompting or fine-tuning LLMs indicate that they still struggle to reliably assess the relative quality or ethical appropriateness of persuasive content. This may partly stem from the challenges inherent in human-annotated datasets for persuasion, which often exhibit noise and subjectivity due to the context-dependent nature of what is perceived as persuasive or manipulative. A reliable judge of persuasion should be able to distinguish beneficial persuasive attempts from harmful ones without introducing systematic biases. In addition to detecting persuasive strategies, such models should also aim to predict the likely response of a user or subject to a persuasive message, enabling more informed and adaptive evaluation.

Identifying Long-Term Persuasion. While models are already challenged in detecting single-turn persuasive attempts, the risks associated with long-term, contextually embedded persuasion are likely to grow. As conversational agents become more integrated into daily life—and potentially retain memory of user preferences, beliefs, and personality traits—they may pose subtle but serious risks by building trust over time and later leveraging it to influence users in ways that may not align with their best interests. Can models gradually steer users while concealing persuasive intent? Can they act with a latent agenda masked by a long history of benign interaction? At present, there are no datasets or evaluation frameworks designed to study such forms of long-term, cumulative persuasion. We hypothesize that LLMs themselves could be used as tools for evaluating long-context persuasion by identifying early signals of manipulation, surfacing shifts in tone or strategy, and issuing warnings in future interactions. Exploring such capabilities could support the development of persuasive systems that are not only effective, but also transparent and trustworthy.

Generative Adversarial Persuasion: AI as Persuader, Persuadee, and Judge. In the preceding sections, we highlighted current limitations in the roles of AI as Persuader, Persuadee, and Persuasion Judge. We believe that progress in these areas can be achieved collectively through a unified framework called *Generative Adversarial Persuasion (GAP)*. In this setup, a persuasive agent attempts to influence a persuadee model, while a judge model oversees the interaction and evaluates the effectiveness, appropriateness, and potential risks of the persuasive attempt. Drawing inspiration from the structure of Generative Adversarial Networks (GANs), this framework encourages co-evolution among the agents. The persuadee learns to develop resistance to manipulative or unethical persuasion, the persuader learns to improve its persuasive techniques in response, and the judge becomes more accurate in identifying persuasive strategies and assessing their quality. This adversarial multi-agent paradigm offers a promising direction for building more robust, adaptive, and ethically grounded persuasive systems.


Computational Persuasion
Evaluating Persuasion
Argument Persuasiveness
Toledo et al. (2019); Habernal & Gurevych (2016); Simpson & Gurevych (2018); Pauli et al. (2025); Rescala et al. (2024)
LLM Persuasiveness
Human Evaluation
Durmus et al. (2024); OpenAI (2024; 2025); Phuong et al. (2024); Salvi et al. (2024); Huang & Wang (2023); Goldstein et al. (2024); Bai et al. (2025)
Automatic Evaluation
Singh et al. Singh et al. (2024) Breum et al. Breum et al. (2023) Bozdag et al. Bozdag et al. (2025) OpenAI OpenAI (2024; 2025) Kunlun et al. Zhu et al. (2025) Idziejczak et al. Idziejczak et al. (2025)
Persuasion Detection
Hidey & McKeown (2018) Pöyhönen et al. (2022) Shmueli-Scheuer et al. (2019) Saenger et al. (2024) Hasanain et al. (2024) Dimitrov et al. (2021) Dutt et al. (2021)
Generating Persuasion
Enhancing Persuasion
Furumai et al. (2024) Chen et al. (2022) Liu et al. Liu et al. (2021) Mishra et al. Mishra et al. (2022) Samad et al. Samad et al. (2022) Shi et al. Shi et al. (2021) Lukin et al. Lukin et al. (2017) Matz et al. (Matz et al., 2024) Kaptein et al. Kaptein et al. (2015) Simchon et al. (2024); Meguellati et al. (2024); Salvi et al. (2024); Tiwari et al. (2022b); Ruiz-Dolz et al. (2024); Zhang & Zhou (2025); Tiwari et al. (2022a; 2023); Cima et al. (2024); Hunter et al. (2019)
Applications of Persuasion
Debate
Add citations
Negotiation
Bianchi et al. Bianchi et al. (2024) Lewis et al. Lewis et al. (2017)
Jailbreaking
Singh et al. (2023) Zeng et al. (2024) Li et al. (2024) Singh et al. (2023) Li et al. (2024) Xu et al. (2024) Bozdag et al. (2025)
Safeguarding Persuasion
Mitigating Unsafe Persuasion
Tsinganos et al. (2022) Burtell & Woodside (2023)

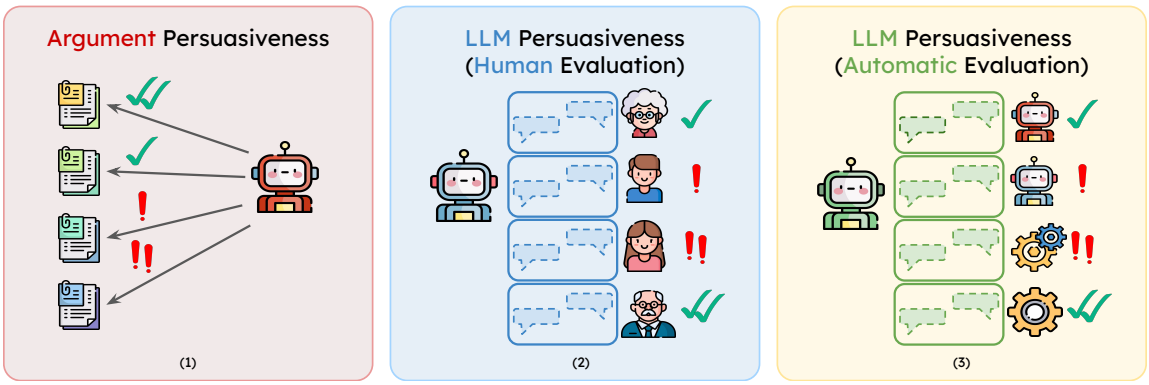


Figure 3: This survey categorizes the evaluation of persuasiveness into three main types: (1) **evaluation of argument persuasiveness**, (2) **human evaluation of LLM-generated content**, and (3) **automatic evaluation of LLM persuasiveness**. For argument persuasiveness, models are typically trained on human-annotated or naturally labeled data to assess the persuasive strength of given arguments. For evaluating LLM persuasiveness, two branches of research emerge: one uses human judges to rate AI-generated content or interactions, while the other relies on LLM-based or non-LLM automatic metrics to perform the evaluation.

References

Aseel Addawood, Adam Badawy, Kristina Lerman, and Emilio Ferrara. Linguistic cues to deception: Identifying political trolls on social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):15–25, Jul. 2019. doi: 10.1609/icwsm.v13i01.3205. URL https://ojs.aaai.org/index.php/ICWSM/article/view/3205.

Cem Anil, Esin DURMUS, Nina Rimsky, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel J Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Baker Grosse, and David Duvenaud. Many-shot jailbreaking. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=cw5mgd71jW.

Gerry Antioch. Persuasion is now 30 per cent of us gdp. *Economic Round-Up*, (1):1–10, 2013. URL https://search.informit.org/doi/10.3316/informit.558637667306970.

S. E. Asch. Effects of group pressure upon the modification and distortion of judgments. In *Groups, leadership and men; research in human relations*, pp. 177–190. Carnegie Press, Oxford, England, 1951.

Hui Bai, Jan G Voelkel, Shane Muldowney, johannes C Eichstaedt, and Robb Willer. Ai-generated messages can be used to persuade humans on policy issues, Mar 2025. URL osf.io/stakv_v5.

Vian Bakir, Eric Herring, David Miller, and Piers Robinson. Organized persuasive communication: A new conceptual framework for research on public relations, propaganda and promotional culture. *Critical Sociology*, 45(3):311–328, 2018. doi: 10.1177/0896920518764586.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can llms negotiate? negotiationarena platform and analysis. *arXiv*, 2024.

Nimet Beyza Bozdag, Shuhaib Mehri, Gokhan Tur, and Dilek Hakkani-Tür. Persuade me if you can: A framework for evaluating persuasion effectiveness and susceptibility among large language models, 2025. URL https://arxiv.org/abs/2503.01829.

Jack W. Brehm. *A theory of psychological reactance*. A theory of psychological reactance. Academic Press, Oxford, England, 1966. Pages: x, 135.

Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. The persuasive power of large language models, 2023. URL https://arxiv.org/abs/2312.15523.

Matthew Burtell and Thomas Woodside. Artificial influence: An analysis of ai-driven persuasion. *arXiv preprint arXiv:2303.08721*, 2023.

Shelly Chaiken. 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, 1980. ISSN 1939-1315. doi: 10.1037/0022-3514.39.5.752. Place: US Publisher: American Psychological Association.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. AMPERSAND: Argument mining for PERSuAsive oNline discussions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2933–2943, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1291. URL https://aclanthology.org/D19-1291/.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3167–3185, Online, June 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.254. URL https://aclanthology.org/2021.naacl-main.254/.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. Casino: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3167–3185, 2021b.

Hui Chen, Deepanway Ghosal, Navonil Majumder, Amir Hussain, and Soujanya Poria. Persuasive dialogue understanding: The baselines and negative results. *Neurocomputing*, 431: 47–56, 2021. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.11.040. URL https://www.sciencedirect.com/science/article/pii/S0925231220318336.

Jiaao Chen and Diyi Yang. 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, May 2021. doi: 10.1609/aaai.v35i14.17498. URL https://ojs.aaai.org/index.php/AAAI/article/view/17498.

Maximillian Chen, Weiyan Shi, Feifan Yan, Ryan Hou, Jingwen Zhang, Saurav Sahay, and Zhou Yu. Seamlessly integrating factual information and social content with persuasive dialogue. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 399–413, Online only, November 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.aacl-main.31. URL https://aclanthology.org/2022.aacl-main.31/.

R.B. Cialdini. *Influence: The Psychology of Persuasion*. Business Library, 1984. ISBN 9781863501569. URL https://books.google.com/books?id=mJidPwAACAAJ.

Robert B. Cialdini. The Science of Persuasion. *Scientific American*, 284(2):76–81, 2001. ISSN 0036-8733. URL https://www.jstor.org/stable/26059056. Publisher: Scientific American, a division of Nature America, Inc.

Lorenzo Cima, Alessio Miaschi, Amaury Trujillo, Marco Avvenuti, Felice Dell'Orletta, and Stefano Cresci. Contextualized counterspeech: Strategies for adaptation, personalization, and evaluation. *arXiv preprint arXiv:2412.07338*, 2024.

Sunny Consolvo, David W. McDonald, Tammy Toscos, Mike Y. Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, Ian Smith, and James A. Landay. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1797–1806, Florence Italy, April 2008. ACM. ISBN 978-1-60558-011-1. doi: 10.1145/1357054.1357335. URL https://dl.acm.org/doi/10.1145/1357054.1357335.

Sunny Consolvo, David W. McDonald, and James A. Landay. Theory-driven design strategies for technologies that support behavior change in everyday life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 405–414, Boston MA USA, April 2009. ACM. ISBN 978-1-60558-246-7. doi: 10.1145/1518701.1518766. URL https://dl.acm.org/doi/10.1145/1518701.1518766.

Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814, 2024. doi: 10.1126/science.adq1814. URL https://www.science.org/doi/abs/10.1126/science.adq1814.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 1377–1414, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.186. URL https://aclanthology.org/2020.semeval-1.186.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

Victor Danciu. Manipulative marketing: persuasion and manipulation of the consumer through advertising. *Theoretical and Applied Economics*, XXI(2(591)):19–34, 2014. URL https://ideas.repec.org//a/agr/journl/vxxiy2014i2(591)p19-34.html. Publisher: Asociatia Generala a Economistilor din Romania / Editura Economica.

Sanorita Dey, Brittany Duff, Karrie Karahalios, and Wai-Tat Fu. The Art and Science of Persuasion: Not All Crowdfunding Campaign Videos are The Same. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 755–769, Portland Oregon USA, February 2017. ACM. ISBN 978-1-4503-4335-0. doi: 10.1145/2998181.2998229. URL https://dl.acm.org/doi/10.1145/2998181.2998229.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu (eds.), *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 70–98, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.semeval-1.7. URL https://aclanthology.org/2021.semeval-1.7.

James N. Druckman. A Framework for the Study of Persuasion. *Annual Review of Political Science*, 25(Volume 25, 2022):65–88, May 2022. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev-polisci-051120-110428. URL https://www.annualreviews.org/content/journals/10.1146/annurev-polisci-051120-110428. Publisher: Annual Reviews.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=zj7YuTE4t8.

Shaddin Dughmi and Haifeng Xu. Algorithmic bayesian persuasion, 2016. URL https://arxiv.org/abs/1503.05988.

Esin Durmus and Claire Cardie. Exploring the role of prior beliefs for argument persuasion. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1035–1045, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1094. URL https://aclanthology.org/N18-1094/.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the persuasiveness of language models, 2024. URL https://www.anthropic.com/news/measuring-model-persuasiveness.

Ritam Dutt, Rishabh Joshi, and Carolyn Rose. Keeping up appearances: Computational modeling of face acts in persuasion oriented discussions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7473–7485, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.605. URL https://aclanthology.org/2020.emnlp-main.605.

Ritam Dutt, Sayan Sinha, Rishabh Joshi, Surya Shekhar Chakraborty, Meredith Riggs, Xinru Yan, Haogang Bao, and Carolyn Penstein Rosé. Resper: Computationally modelling resisting strategies in persuasive conversations. *arXiv preprint arXiv:2101.10545*, 2021.

Subhabrata Dutta, Dipankar Das, and Tanmoy Chakraborty. Changing views: Persuasion modeling and argument extraction from online discussions. *Information Processing & Management*, 57(2):102085, 2020. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2019.102085. URL https://www.sciencedirect.com/science/article/pii/S0306457319301165.

Matthew C. Farrelly, James Nonnemaker, Kevin C. Davis, and Altijani Hussin. The Influence of the National truth® Campaign on Smoking Initiation. *American Journal of Preventive Medicine*, 36(5):379–384, May 2009. ISSN 0749-3797, 1873-2607. doi: 10.1016/j.amepre.2009.01.019. URL https://www.ajpmonline.org/article/S0749-3797(09)00074-9/fulltext. Publisher: Elsevier.

Nicolás Ferreyra, Esma Aïmeur, Hicham Hage, Maritta Heisel, and Catherine van Hoogstraten. Persuasion meets ai: Ethical considerations for the design of social engineering countermeasures. In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge*

Engineering and Knowledge Management, pp. 204–211. SCITEPRESS - Science and Technology Publications, 2020. doi: 10.5220/0010142402040211. URL http://dx.doi.org/10.5220/0010142402040211.

Leon Festinger. *A theory of cognitive dissonance*. A theory of cognitive dissonance. Stanford University Press, 1957. ISBN 978-0-8047-0131-0 978-0-8047-0911-8. Pages: xi, 291.

BJ Fogg. Captology: the study of computers as persuasive technologies. In *CHI '97 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '97, pp. 129, New York, NY, USA, March 1997. Association for Computing Machinery. ISBN 978-0-89791-926-5. doi: 10.1145/1120212.1120301. URL https://dl.acm.org/doi/10.1145/1120212.1120301.

BJ Fogg. Persuasive computers: perspectives and research directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '98, pp. 225–232, USA, January 1998. ACM Press/Addison-Wesley Publishing Co. ISBN 978-0-201-30987-4. doi: 10.1145/274644.274677. URL https://dl.acm.org/doi/10.1145/274644.274677.

BJ Fogg. A behavior model for persuasive design. In *Proceedings of the 4th International Conference on Persuasive Technology*, Persuasive '09, pp. 1–7, New York, NY, USA, April 2009a. Association for Computing Machinery. ISBN 978-1-60558-376-1. doi: 10.1145/1541948.1541999. URL https://dl.acm.org/doi/10.1145/1541948.1541999.

BJ Fogg. Creating persuasive technologies: an eight-step design process. In *Proceedings of the 4th International Conference on Persuasive Technology*, Persuasive '09, pp. 1–6, New York, NY, USA, April 2009b. Association for Computing Machinery. ISBN 978-1-60558-376-1. doi: 10.1145/1541948.1542005. URL https://dl.acm.org/doi/10.1145/1541948.1542005.

Kazuaki Furumai, Roberto Legaspi, Julio Cesar Vizcarra Romero, Yudai Yamazaki, Yasutaka Nishimura, Sina Semnani, Kazushi Ikeda, Weiyan Shi, and Monica Lam. Zero-shot persuasive chatbots with LLM-generated strategies and information retrieval. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11224–11249, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.656. URL https://aclanthology.org/2024.findings-emnlp.656/.

R.H. Gass and J.S. Seiter. *Persuasion: Social Influence and Compliance Gaining*. Taylor & Francis, 2022. ISBN 9781000556773. URL https://books.google.com/books?id=leFeEAAAQBAJ.

Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. How persuasive is ai-generated propaganda? *PNAS Nexus*, 3(2):pgae034, 02 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae034. URL https://doi.org/10.1093/pnasnexus/pgae034.

Marco Guerini, Gözde Özbal, and Carlo Strapparava. Echoes of persuasion: The effect of euphony in persuasive communication. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar (eds.), *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1483–1493, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1172. URL https://aclanthology.org/N15-1172.

Ivan Habernal and Iryna Gurevych. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1589–1599, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1150. URL https://aclanthology.org/P16-1150/.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-agent risks from advanced ai, 2025. URL https://arxiv.org/abs/2502.14143.

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

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2333–2343, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1256. URL https://aclanthology.org/D18-1256/.

Christopher Hidey and Kathleen McKeown. Persuasive influence detection: The role of argument sequencing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.12003. URL https://ojs.aaai.org/index.php/AAAI/article/view/12003.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. Analyzing the semantic types of claims and premises in an online persuasive forum. In Ivan Habernal, Iryna Gurevych, Kevin Ashley, Claire Cardie, Nancy Green, Diane Litman, Georgios Petasis, Chris Reed, Noam Slonim, and Vern Walker (eds.), *Proceedings of the 4th Workshop on Argument Mining*, pp. 11–21, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5102. URL https://aclanthology.org/W17-5102/.

Guanxiong Huang and Sai Wang. Is artificial intelligence more persuasive than humans? a meta-analysis. *Journal of Communication*, 73(6):552–562, 08 2023. ISSN 0021-9916. doi: 10.1093/joc/jqad024. URL https://doi.org/10.1093/joc/jqad024.

Anthony Hunter, Lisa Chalaguine, Tomasz Czernuszenko, Emmanuel Hadoux, and Sylwia Polberg. Towards computational persuasion via natural language argumentation dialogues. In *KI 2019: Advances in Artificial Intelligence: 42nd German Conference on AI, Kassel, Germany, September 23–26, 2019, Proceedings 42*, pp. 18–33. Springer, 2019.

Mateusz Idziejczak, Vasyl Korzavatykh, Mateusz Stawicki, Andrii Chmutov, Marcin Korcz, Iwo Błądek, and Dariusz Brzezinski. Among them: A game-based framework for assessing persuasion capabilities of llms, 2025. URL https://arxiv.org/abs/2502.20426.

Nassim Jafarinaimi, Jodi Forlizzi, Amy Hurst, and John Zimmerman. Breakaway: an ambient display designed to change human behavior. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, pp. 1945–1948, New York, NY, USA, April 2005. Association for Computing Machinery. ISBN 978-1-59593-002-6. doi: 10.1145/1056808.1057063. URL https://dl.acm.org/doi/10.1145/1056808.1057063.

Rishabh Joshi, Vidhisha Balachandran, Shikhar Vashishth, Alan Black, and Yulia Tsvetkov. Dialograph: Incorporating interpretable strategy-graph networks into negotiation dialogues. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=kDnal_bbb-E.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101 (6):2590–2615, October 2011. doi: 10.1257/aer.101.6.2590. URL https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590.

Maurits Kaptein, Panos Markopoulos, Boris De Ruyter, and Emile Aarts. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies*, 77:38–51, 2015.

Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. Working with ai to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), April 2023. doi: 10.1145/3579592. URL https://doi.org/10.1145/3579592.

Simon Keizer, Markus Guhe, Heriberto Cuayáhuitl, Ioannis Efstathiou, Klaus-Peter Engelbrecht, Mihai Dobre, Alex Lascarides, and Oliver Lemon. Evaluating persuasion strategies and deep reinforcement learning methods for negotiation dialogue agents. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 480–484, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2077/.

Taraneh Khazaei, Lu Xiao, and Robert Mercer. Writing to persuade: Analysis and detection of persuasive discourse. In *iConference 2017 Proceedings*. iSchools, 2017. URL http://hdl.handle.net/2142/96673.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2443–2453, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1259. URL https://aclanthology.org/D17-1259.

Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. Llm defenses are not robust to multi-turn human jailbreaks yet, 2024. URL https://arxiv.org/abs/2408.15221.

Wenhao Li, Yue Lin, Xiangfeng Wang, Bo Jin, Hongyuan Zha, and Baoxiang Wang. Verbalized bayesian persuasion, 2025. URL https://arxiv.org/abs/2502.01587.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3469–3483, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.269. URL https://aclanthology.org/2021.acl-long.269/.

Irina Lock and Ramona Ludolph. Organizational propaganda on the internet: A systematic review. *Public Relations Inquiry*, 9(1):103–127, 2020. doi: 10.1177/2046147X19870844.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 742–753, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1070/.

Ivana Marková. Persuasion and Propaganda. *Diogenes*, 55(1):37–51, February 2008. ISSN 0392-1921, 1467-7695. doi: 10.1177/0392192107087916. URL https://www.cambridge.org/core/journals/diogenes/article/persuasion-and-propaganda/5B877DC7CC09B164EC95FAB20F209EDF.

Sandra C Matz, Jacob D Teeny, Sumer S Vaid, Heinrich Peters, Gabriella M Harari, and Moran Cerf. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14 (1):4692, 2024.

Donald McCloskey and Arjo Klamer. One quarter of gdp is persuasion. *The American Economic Review*, 85(2):191–195, 1995. ISSN 00028282. URL http://www.jstor.org/stable/2117917.

William J. McGuire. The nature of attitudes and attitude change. In Elliot Aronson and Gardner Lindzey (eds.), *The Handbook of Social Psychology*, volume 3, pp. 136–314. Addison-Wesley, Massachusetts, 2nd edition, 1969.

Elyas Meguellati, Lei Han, Abraham Bernstein, Shazia Sadiq, and Gianluca Demartini. How good are llms in generating personalized advertisements? In *Companion Proceedings of the ACM Web Conference 2024*, pp. 826–829, 2024.

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R. Bowman. Debate helps supervise unreliable experts, 2023. URL https://arxiv.org/abs/2311.08702.

Stanley Milgram. Behavioral Study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4):371–378, 1963. ISSN 0096-851X. doi: 10.1037/h0040525. Place: US Publisher: American Psychological Association.

Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. Please be polite: Towards building a politeness adaptive dialogue system for goal-oriented conversations. *Neurocomputing*, 494:242–254, 2022.

Elena Musi, Debanjan Ghosh, and Smaranda Muresan. Changemyview through concessions: Do concessions increase persuasion?, 2018. URL https://arxiv.org/abs/1806.03223.

OpenAI. Openai o1 system card, December 2024. URL https://cdn.openai.com/o1-system-card-20241205.pdf. Accessed: 2024-12-10.

OpenAI. Openai gpt-4.5 system card, February 2025. URL https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf. Accessed: 2025-03-06.

Alexis Palmer and Arthur Spirling and. Large language models can argue in convincing ways about politics, but humans dislike ai authors: implications for governance. *Political Science*, 75(3):281–291, 2023. doi: 10.1080/00323187.2024.2335471. URL https://doi.org/10.1080/00323187.2024.2335471.

Amalie Pauli, Leon Derczynski, and Ira Assent. Modelling persuasion through misuse of rhetorical appeals. In Laura Biester, Dorottya Demszky, Zhijing Jin, Mrinmaya Sachan, Joel Tetreault, Steven Wilson, Lu Xiao, and Jieyu Zhao (eds.), *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pp. 89–100, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nlp4pi-1.11. URL https://aclanthology.org/2022.nlp4pi-1.11/.

Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. Measuring and benchmarking large language models' capabilities to generate persuasive language, 2025. URL https://arxiv.org/abs/2406.17753.

Richard E. Petty and John T. Cacioppo. The Elaboration Likelihood Model of Persuasion. In Leonard Berkowitz (ed.), *Advances in Experimental Social Psychology*, volume 19, pp. 123–205. Academic Press, January 1986. doi: 10.1016/S0065-2601(08)60214-2. URL https://www.sciencedirect.com/science/article/pii/S0065260108602142.

Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, Heidi Howard, Tom Lieberum, Ramana Kumar, Maria Abi Raad, Albert Webson, Lewis Ho, Sharon Lin, Sebastian Farquhar, Marcus Hutter, Gregoire Deletang, Anian Ruoss, Seliem El-Sayed, Sasha Brown, Anca Dragan, Rohin Shah, Allan Dafoe, and Toby Shevlane. Evaluating frontier models for dangerous capabilities, 2024. URL https://arxiv.org/abs/2403.13793.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori (eds.), *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 2343–2361, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.317. URL https://aclanthology.org/2023.semeval-1.317/.

Teemu Pöyhönen, Mika Hämäläinen, and Khalid Alnajjar. Multilingual persuasion detection: Video games as an invaluable data source for nlp, 2022. URL https://arxiv.org/abs/2207.04453.

Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. Can language models recognize convincing arguments?, 2024. URL https://arxiv.org/abs/2404.00750.

Ramon Ruiz-Dolz, Joaquin Taverner, Stella M Heras Barberá, and Ana García-Fornes. Persuasion-enhanced computational argumentative reasoning through argumentation-based persuasive frameworks. *User Modeling and User-Adapted Interaction*, 34(1):229–258, 2024.

Till Raphael Saenger, Musashi Hinck, Justin Grimmer, and Brandon M Stewart. Autopersuade: A framework for evaluating and explaining persuasive arguments. *arXiv preprint arXiv:2410.08917*, 2024.

Hiromasa Sakurai and Yusuke Miyao. Evaluating intention detection capability of large language models in persuasive dialogues. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1635–1657, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.90. URL https://aclanthology.org/2024.acl-long.90.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the conversational persuasiveness of large language models: A randomized controlled trial. *CoRR*, abs/2403.14380, 2024. URL https://doi.org/10.48550/arXiv.2403.14380.

Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. Empathetic persuasion: Reinforcing empathy and persuasiveness in dialogue systems. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 844–856, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.63. URL https://aclanthology.org/2022.findings-naacl.63.

Omar Shaikh, Jiaao Chen, Jon Saad-Falcon, Polo Chau, and Diyi Yang. Examining the ordering of rhetorical strategies in persuasive requests. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1299–1306, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.116. URL https://aclanthology.org/2020.findings-emnlp.116/.

Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. Effects of persuasive dialogues: Testing bot identities and inquiry strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pp. 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376843. URL https://doi.org/10.1145/3313831.3376843.

Weiyan Shi, Yu Li, Saurav Sahay, and Zhou Yu. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3478–3492, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.295. URL https://aclanthology.org/2021.findings-emnlp.295/.

Michal Shmueli-Scheuer, Jonathan Herzig, David Konopnicki, and Tommy Sandbank. Detecting persuasive arguments based on author-reader personality traits and their interaction. In *Proceedings of the 27th ACM conference on user modeling, adaptation and personalization*, pp. 211–215, 2019.

Murtaza Ahmed Siddiqi, Wooguil Pak, and Moquddam A. Siddiqi. A study on the psychology of social engineering-based cyberattacks and existing countermeasures. *Applied Sciences*, 12(12): 6042, 2022. doi: 10.3390/app12126042.

Almog Simchon, Matthew Edwards, and Stephan Lewandowsky. The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS nexus*, 3(2):pgae035, 2024.

Edwin Simpson and Iryna Gurevych. Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371, 2018. doi: 10.1162/tacl_a_00026. URL https://aclanthology.org/Q18-1026/.

Somesh Singh, Yaman K Singla, Harini SI, and Balaji Krishnamurthy. Measuring and improving persuasiveness of large language models, 2024. URL https://arxiv.org/abs/2410.02653.

Sonali Singh, Faranak Abri, and Akbar Siami Namin. Exploiting large language models (llms) through deception techniques and persuasion principles. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2508–2517. IEEE, 2023.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 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, pp. 613–624, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883081. URL https://doi.org/10.1145/2872427.2883081.

Abhisek Tiwari, Sriparna Saha, Shubhashis Sengupta, Anutosh Maitra, Roshni Ramnani, and Pushpak Bhattacharyya. Persona or context? towards building context adaptive personalized persuasive virtual sales assistant. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1035–1047, Online only, November 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.aacl-main.76. URL https://aclanthology.org/2022.aacl-main.76/.

Abhisek Tiwari, Tulika Saha, Sriparna Saha, Shubhashis Sengupta, Anutosh Maitra, Roshni Ramnani, and Pushpak Bhattacharyya. A persona aware persuasive dialogue policy for dynamic and co-operative goal setting. *Expert Systems with Applications*, 195:116303, 2022b. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.116303. URL https://www.sciencedirect.com/science/article/pii/S0957417421016067.

Abhisek Tiwari, Abhijeet Khandwe, Sriparna Saha, Roshni Ramnani, Anutosh Maitra, and Shubhashis Sengupta. Towards personalized persuasive dialogue generation for adversarial task oriented dialogue setting. *Expert Systems with Applications*, 213:118775, 2023.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. Automatic argument quality assessment - new datasets and methods. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5625–5635, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1564. URL https://aclanthology.org/D19-1564/.

Nikolaos Tsinganos, Ioannis Mavridis, and Dimitris Gritzalis. Utilizing convolutional neural networks and word embeddings for early-stage recognition of persuasion in chat-based social engineering attacks. *IEEE Access*, 10:108517–108529, 2022.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5635–5649, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1566. URL https://aclanthology.org/P19-1566.

Zhongyu Wei, Yang Liu, and Yi Li. Is this post persuasive? ranking argumentative comments in online forum. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 195–200, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2032. URL https://aclanthology.org/P16-2032/.

Zachary Wojtowicz. When and why is persuasion hard? a computational complexity result, 2024. URL https://arxiv.org/abs/2408.07923.

Ziang Xiao, Po-Shiun Ho, Xinran Wang, Karrie Karahalios, and Hari Sundaram. Should We Use an Abstract Comic Form to Persuade?: Experiments with Online Charitable Donation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28, November 2019. ISSN 2573-0142. doi: 10.1145/3359177. URL https://dl.acm.org/doi/10.1145/3359177.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16259–16303, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.858. URL https://aclanthology.org/2024.acl-long.858.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 3620–3630, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1364. URL https://aclanthology.org/N19-1364/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.

14322–14350, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.773. URL https://aclanthology.org/2024.acl-long.773.

Dingyi Zhang and Deyu Zhou. Persuasion should be double-blind: A multi-domain dialogue dataset with faithfulness based on causal theory of mind, 2025. URL https://arxiv.org/abs/2502.21297.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. Multiagentbench: Evaluating the collaboration and competition of llm agents, 2025. URL https://arxiv.org/abs/2503.01935.