LLM Persuasiveness Evaluation: A Structured Review of Automated Methods

Anonymous Authors[1]

Abstract

As Large Language Models (LLMs) become increasingly persuasive, their ability to shape beliefs and behaviour at scale has raised concerns, prompting regulatory attention and calls for robust evaluation frameworks. Human-participant studies provide ecological validity but are costly, slow, and constrained by ethical challenges, making them impractical for systematic assessment of rapidly evolving systems. As an alternative, fully automated evaluation methods requiring no human involvement enable reproducible, fast, and ethically unconstrained large-scale testing. To organise this rapidly growing literature and inform future research and development, we provide the first systematic taxonomy of 30 automated methods across 27 papers, examining their designs, human validation results, limitations, and associated risks.

1. Introduction

Persuasion is central to human interaction across social, commercial, and public spheres, shaping processes as diverse as political campaigning, marketing, health communication, and public deliberation (Bassi et al., 2024). In the digital era, algorithmic curation, personalised content delivery, and persuasive system design have made persuasion increasingly computational and data-driven (Fogg, 2003; Bassi et al., 2024). This shift has culminated in the emergence of "synthetic persuasion": LLMs acting as interactive agents that can tailor arguments, mimic affect, and adapt across multi-turn dialogue (Bozdag et al., 2026; Breum et al., 2024; Doudkin et al., 2025). Whereas persuasion was previously either personalised (one-on-one) or scalable (mass media), LLMs enable both simultaneously, delivering automated, interactive, and hyper-personalised influence at scale.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by AI4GOOD at the International Conference on Machine Learning (ICML). Do not distribute.

Empirical evidence indicates that frontier LLMs are becoming increasingly persuasive, in some settings matching or even exceeding human performance (Rogiers et al., 2024). Studies in 2024 showed LLM persuasive abilities rapidly approaching human performance across a range of controlled settings (Ma et al., 2025; Salvi et al., 2025; Durmus et al., 2024). By 2025, this trend had accelerated: several studies documented persuasion parity, or even superiority, relative to humans (Meguellati et al., 2025; Becker et al., 2025; Li et al., 2025; Maier et al., 2025). These developments expand both the potential benefits, such as supporting education or public health, and the risks, including manipulation and undue influence – flagged explicitly by the International AI Safety Report (Bengio et al., 2025; 2026). As LLMs become embedded in digital platforms at scale, their ability to shape beliefs and behaviour raises urgent concerns for cognitive autonomy, democratic discourse, and societal governance. Their possible impact on public discourse has prompted regulatory attention, reflected in the EU AI Act (European Parliament and Council of the European Union, 2024) and the complementary Code of Practice for General Purpose AI (European Commission, 2025), motivating the need for robust, scalable evaluation frameworks that enable early risk detection and inform the design of effective safeguards.

Early research on LLM persuasiveness evaluation relied primarily on human-participant experiments (Singh et al., 2024; Durmus et al., 2024; OpenAI, 2024), which provide ecological validity but suffer from practical and methodological limitations. Human trials are slow, expensive, and constrained by small, often homogeneous participant pools (Salvi et al., 2025). They cannot feasibly test high-risk scenarios due to ethical constraints, and struggle to capture the full behavioural space of modern LLMs, including multi-session interactions, adversarial dynamics, or large-scale personalised messaging – making them insufficient as a standalone evaluation approach for increasingly capable and widely deployed LLMs.

To overcome these limitations, research has increasingly turned towards automated evaluation methods. They enable controlled, reproducible, and large-scale assessment, including high-risk and ethically challenging scenarios. However, such methods introduce their own challenges and limitations; understanding them is essential to ensure that automated evaluation becomes a reliable avenue for LLM

assessment.

Existing surveys focus on ethical risks, human-subject methodologies, persuasion detection, or conceptual taxonomies (Jones & Bergen, 2026; Bozdag et al., 2026; Rogiers et al., 2024; Bassi et al., 2024). However, whilst some briefly mention automated approaches, none critically assess their methodological design or reliability. To address this gap, we provide the first systematic review of fully automated LLM persuasion evaluation methods, examining their designs, empirical validity, and limitations.

Contributions This structured review provides: (1) the first systematic review of fully automated LLM persuasiveness evaluation methods, analysing evaluation architectures and validation practices (Section 2); (2) a structured taxonomy categorising methods into different types of dataset- and simulation-based approaches (Section 3); (3) an overview of human validation examining alignment with human judgements (Section 4); and (4) synthesis of core challenges (Sections 5), associated risks (Sections 5.2), and future research directions (Sections 5.3). This work is accompanied by a centralised resource hub of datasets and codebases for LLM persuasion evaluation. [1]

2. Methodology

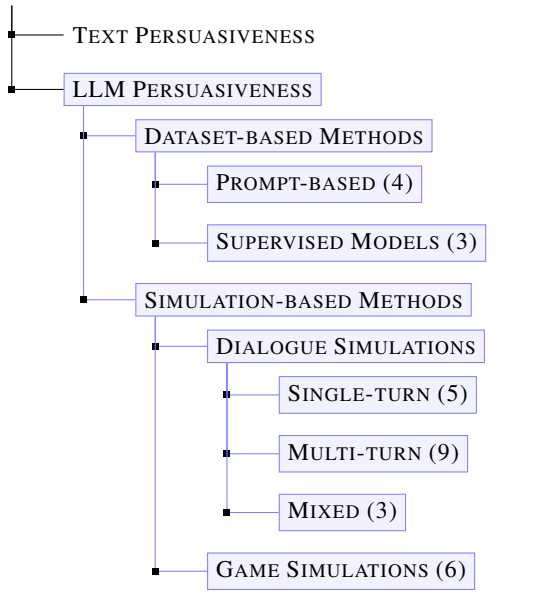
This section presents the methodology used to identify and synthesise automated LLM persuasion evaluation methods.

Scope & Definitions This paper focuses on the automated evaluation of LLMs' ability to influence human beliefs, attitudes, and behaviour through text-based communication. We use **persuasion** as an umbrella term encompassing all forms of communicative influence, including manipulation and deception, when used as a persuasive technique, following Jones & Bergen (2026). Persuasion is defined as "a successful intentional effort at influencing another's mental state through communication in a circumstance in which the persuadee has some measure of freedom" (O'Keefe, 2015), where intentionality means the output was designed, prompted, or optimised to exert persuasive influence (Jones & Bergen, 2026). Further definitions and conceptual relations are discussed in Appendix A.

Eligibility Criteria Studies were included if they: (1) evaluated LLM persuasiveness using fully automated methods (no human involvement during evaluation); (2) were published between 2021 and 2025; and (3) were written in English. Peer-reviewed publications, preprints, and technical reports were all considered for a comprehensive coverage of this rapidly evolving field. Eligibility clarifications can be found in Appendix B.1.

[1] An anonymised preview can be found in the Appendix C.

Figure 1. The taxonomy of automated LLM persuasion evaluation methods. Blue boxes indicate the main scope of this paper. Numbers in parentheses indicate the method count per category ($n = 30$ total , see Table 1 for details).



Search Strategy We searched Scopus, ACL Anthology, and ArXiv (10 January 2026) using keyword queries, supplemented by LLM-assisted filtering, LLM-assisted search and backwards reference searching (detailed search strategy in Appendix B.2). After deduplication and manual screening, we identified **27 papers** containing **30 distinct evaluation methods**.

Synthesis The final set of 27 studies was reviewed using a structured extraction framework. For each, we recorded the details of the evaluation design, the task domain, the type of metrics used, and human validation (if any). These dimensions were summarised in a comparative table and then synthesised to identify recurrent methodological patterns (see Table 1). This process led to a set of recurring categories, which form the basis of the taxonomy presented in Section 3.

3. Automated LLM Persuasion Evaluation Methods

We propose a taxonomy of automated persuasion evaluation methods for LLMs, displayed in Figure 1. At a high level, *Automated Textual Persuasion Evaluation* is split into two categories based on their primary evaluative focus: **Text Persuasiveness** (Section 3.1) and **LLM Persuasiveness**. Methods under *LLM Persuasiveness* are model-centric, evaluating the model's performance as a persuader. In contrast,

Text Persuasiveness evaluation methods are content-centric and source-agnostic: they aim to assess the persuasiveness of any text. Although some model-centric methods are technically capable of assessing arbitrary text, they are assigned to the *LLM Persuasiveness* category because their design objectives, evaluation setups, and reported findings all centre on measuring LLM persuasive capabilities. Although this survey focuses on model-centric evaluation methods, it also includes a brief overview of *Text Persuasiveness* methods (Section 3.1), as these approaches remain relevant for model-level evaluation.

The *LLM Persuasiveness* category is further divided into **Dataset-based** (Section 3.2) and **Simulation-based** (Section 3.3) methods. *Dataset-based* methods measure persuasion using a fixed dataset as the key element of the evaluation framework and are further categorised into **Prompt-based Methods** (Section 3.2.1) and **Supervised Scoring Models** (Section 3.2.2). *Simulation-based* approaches assess persuasion through interactive, LLM-driven scenarios, with either **Game-based** (Section 3.3.2) or **Dialogue-based** (Section 3.3.1) setups, the latter further split into *Single-turn*, *Multi-turn*, or *Mixed*. A consolidated overview of all reviewed evaluation methods within these categories is provided in Table 1.

3.1. Automated Evaluation of Text Persuasiveness

We begin with text-centric methods that assess the persuasiveness of any text, irrespective of its source. Evaluating how persuasive an argument is, remains a persistent challenge in Natural Language Processing (NLP) (Bozdag et al., 2026), typically approached via *absolute* (judging a single text independently) or *relative* (comparing texts pairwise) evaluation. Early work employed supervised models trained on pairwise judgments (Habernal & Gurevych, 2016; Simpson & Gurevych, 2018), later advancing to fine-tuned transformers for both pairwise and absolute scoring (Toledo et al., 2019; Pauli et al., 2025). Recent approaches employ LLM-as-a-judge (or LLM-judge) methods (Rescala et al., 2024), though their superiority over classical regression models trained on lexical features remains contested (Barkar et al., 2025). While these methods can assess LLM-generated text, they focus on static content rather than the dynamic persuasive capabilities examined in the following model-centric approaches.

3.2. Dataset-Based Methods

Dataset-based persuasion evaluation methods treat a given benchmark dataset as the central element of the evaluation framework. These approaches leverage fixed corpora, either containing real-world behavioural or synthetic data, and fall into two subcategories: (1) *Prompt-based Methods* that use LLM-judges for dataset-grounded argument evaluation, and (2) *Supervised Models* that are trained on labelled datasets to predict persuasiveness.

3.2.1. Prompt-based Methods

Prompt-based methods use LLMs as zero-shot or few-shot judges, relying on prompting rather than trained models. Four studies assessed persuasiveness via LLM-judge setups that perform either pairwise comparisons or absolute scoring. Jin et al. (2024) evaluated their fine-tuned persuasive model against dataset arguments, measuring win rates. Yeginbergen et al. (2025) assessed counter-argument generation effectiveness across five dimensions (incl. persuasiveness), ranking outputs by summed LLM-judge scores. Taking a different angle, Elaraby et al. (2024) evaluated not the arguments themselves but the persuasiveness of LLM-generated rationales for pairwise argument ranking decisions, using an LLM-judge to compute an average persuasive rank per model. Finally, using a dataset of social media post pairs (from the platform formerly known as *Twitter*, now *X*) Singh et al. (2024) evaluated LLM persuasiveness across two dimensions: (1) *generative capabilities* (rewriting content to increase persuasiveness or adapt it for different audiences), assessed via Elo-style ratings in an arena-style tournament using a trained Oracle judge; (2) *simulative capabilities* (predicting absolute and relative content engagement levels), assessed via standard accuracy metrics and motivated by the argument that effective persuasion generation requires self-evaluation capabilities.

3.2.2. Supervised Models

Rather than relying on LLMs as evaluators, three studies employed trained models as automated evaluators. Pauli et al. (2025) focused on comparative assessment, training a DeBERTaV3 regression model on human-annotated text pairs (*Persuasive-Pairs*) to measure relative persuasiveness changes in LLM rewrites – capturing whether rewrites strengthen or weaken the original argument. Going beyond relative comparison, Song & Wang (2024) trained a BERT-based neural regression model on actual donation amounts (*PersuasionForGood*) to predict absolute persuasive effectiveness in human-human charity dialogues. Finally, Jaipersaud et al. (2025) shifted focus from outcome-based scoring, proposing linear probes trained on frozen LLM activations from synthetic multi-turn dialogues to classify persuasion success at the token, turn, and conversation levels.

3.3. Simulation-Based Methods

Simulation-based methods assess persuasion through interactive LLM-driven scenarios, capturing dynamic behaviour absent from static dataset-based evaluations. These approaches vary considerably in design, ranging from structured dialogues between persuader and persuadee agents,

Table 1. Summary of the reviewed automated LLM persuasiveness evaluation methods. For research papers employing multiple techniques, the specific method is additionally listed in the *Study* column. The papers are ordered by category, as discussed in Section 3.

		Study	Evaluator	Metrics
Dataset-based		Jin et al.	LLM-judge	win rate
		Yeginbergen et al.	LLM-judge	ranking (summed scores across 5 dimensions)
		Elaraby et al.	LLM-judge	average persuasive rank
		Singh et al.	Oracle judge, ground-truth comparison	Elo rating, prediction accuracy
		Pauli et al.	regression model	relative persuasiveness score
		Song & Wang	regression model	predicted donation amount
		Jaipersaud et al.	linear probe	persuasion success probability
Simulation-based	Single-turn	Breum et al.	self-reported	success probability
		Ju et al.	self-reported	success rate
		He et al.	self-reported	success rate
		Durmus et al.	self-reported	stance change
		Namikoshi et al.	self-reported	stance change
	Multi-turn	*MakeMeSay* OpenAI	rule-based	win rate
		MakeMePay OpenAI	rule-based	win rate, dollar extraction rate
		Kim et al.	rule-based	success rate, sales-win-rate
		Ramani et al.	rule-based, LLM-judge, self-reported	success rate, persuasiveness score, stance change
		Bargaining Zhu et al.	LLM-judge	task score (success)
		Liu et al.	LLM-judge	effectiveness score
		Wang et al.	LLM-judge, rule-based	motivation & confidence change, time-to-success
		Yao et al.	self-reported	normalised persuasiveness score, AUC metric
		Persuasion Hong et al.	self-reported	average donation amount
	Mixed	Bozdag et al.	self-reported	stance change (success & susceptibility)
		Cheng & You	self-reported	stance change
		Doudkin et al.	self-reported, rule-based	belief & behaviour change
	Game-based	Idziejczak et al.	game rules	win rate
		AvalonBench Hong et al.	game rules	win rate
		Bailis et al.	game rules	win rate, performance metrics
		Werewolf Zhu et al.	scoring rules, LLM-judge	communication, planning & task success scores
		Costa & Vicente	game rules, theoretical model	win rate, capability scores (incl. deception)
		Duffy et al.	self-reported	relationship change (frequency and magnitude)

discussed in Section 3.3.1, to competitive game-based settings, discussed in Section 3.3.2.

3.3.1. Dialogue-Based Methods

Dialogue-based methods simulate persuader–persuadee interactions, with persuadees conditioned on personas or quantified belief states and instructed to update their stance throughout or after the interaction. The methods are organised by interaction complexity (single-turn, multi-turn, or mixed).

Methods employing Single-turn interactions Five studies measure LLM persuasiveness through single-turn interactions using LLM persuadee self-reported metrics. Three employ binary acceptance metrics (accept or reject argument). Breum et al. (2024) computed persuasion success probability against persuadees with varying stubbornness levels, Ju et al. (2025) measured success rate on factually incorrect claims, and He et al. (2025) measured argument success rates across four persuasion methods (two Bayesian Persuasion (BP) variants and two non-BP baselines). Two remaining methods use more granular measures: Durmus et al. (2024) measured stance shifts on a 1–7 Likert scale, while Namikoshi et al. (2024) tracked preference changes on a 0–100 scale.

Methods employing Multi-turn interactions Nine studies employed multi-turn interactions to capture persuasion as an iterative and adaptive process. These settings allow persuader agents to adjust strategies, respond to resistance, and build influence over time – features that cannot be observed in single-turn evaluations.

Four methods used rule-based outcome metrics. OpenAI's *MakeMePay* and *MakeMeSay* (OpenAI, 2024) measured

deceptive persuasion through payment extraction metrics (frequency and dollar amount) and covert codeword elicitation under strict win conditions, respectively. Kim et al. (2025) measured success rates and sales-win-rates (proportion of accepted items exceeding budget) using user agents constructed from real Amazon purchase histories. Also exploring sales scenarios, Ramani et al. (2024) combined three evaluation types via weighted average: rule-based purchase decisions (highest weight), self-reported belief changes, and LLM-judge persuasiveness scores.

Three methods employed LLM-judge evaluation. Zhu et al. (2025) assessed multi-agent bargaining through LLM-scored interaction transcripts of seller and buyer LLMs. Exploring high-risk scenarios, Liu et al. (2025) measured both the model's willingness to refuse unethical persuasion requests and, for non-refused interactions, overall persuasion effectiveness via LLM-judge (5-point scale), assessing argument quality, adaptability, and persuadee belief change. Wang et al. (2025) tracked therapeutic behaviour change in addiction recovery simulations through multi-session interactions with persistent patient states and environmental stressors, measuring motivation, confidence changes and time-to-success via rubric-based LLM-judge.

Two methods relied on self-reported metrics. Yao et al. (2026) evaluated healthcare persuasion for insulin pump adoption using clinically-grounded patient agents across single- and multi-visit scenarios with reflection or social resistance, measuring persuasiveness through self-reported ratings and AUC metrics. Finally, Hong et al. (2025) assessed charitable giving dialogues with persuadees selecting donation amounts at the end of the interactions.

Methods employing both Single and Multi-turn interactions Three studies contrasted single- and multi-turn interaction formats to evaluate how persuasion effectiveness shifts across interaction lengths. Bozdag et al. (2025) measured self-reported agreement on 5-point Likert scales across two domains: subjective claims (social/political issues) and misinformation (factual inaccuracies). Cheng & You (2025) formalised persuasion using game theory, measuring stance shifts on 7-point scales. Finally, Doudkin et al. (2025) compared persuasion effectiveness across three participant types (real humans, simulated humans from real data, and fully synthetic personas) exposed to static statements, non-personalised chat, or personalised chat with four persuasion strategies, measuring belief, behaviour change primarily from pre- and post-intervention surveys and a behavioural task.

3.3.2. Game-Based Methods

Six studies evaluated persuasion within competitive game scenarios. Five employed social deduction games in arena-style tournaments where LLMs with assigned roles competed using persuasion and deception: Bailis et al. (2024) and Zhu et al. (2025) employed *Werewolf* game variations, Idziejczak et al. (2025) used *Among Us*, Costa & Vicente (2025) – *Mini-Mafia*, and Hong et al. (2025) – *Resistance Avalon*. Evaluation primarily relied on win rates, with Bailis et al. (2024) additionally measuring role-specific performance and bid-based turn-taking as persuasive intent indicators, Zhu et al. (2025) scoring multiagent coordination through communication and planning, and Costa & Vicente (2025) used Bayesian inference on tournament win rates to estimate deception, detection, and disclosure capability scores. Duffy et al. (2026) evaluated persuasion in *Diplomacy*, distinct from social deduction games in that all board information is public. They measured allegiance shift frequency and magnitude on a 5-level scale in scenarios where the persuader was preset as an enemy.

4. Human Validation of Automated Persuasion Evaluation

Human validation gives insight into the extent to which automated persuasion evaluation metrics align with human judgements and behaviour. While critical for grounding automated metrics, human validation is limited and inconsistent across the reviewed literature. Less than half of reviewed methods (13/30) include human validation of their persuasiveness assessment methods. Notably, none of the six game-based methods validated persuasiveness against human responses, leaving their relevance for modelling human persuasion outcomes unexamined. Among validated methods, argument persuasiveness judgments showed strong alignment with human assessments, while belief change and behavioural outcome prediction demonstrated substantial variability. A comparative overview is provided in Table 2.

Argument persuasiveness judgement validation Four studies validated argument persuasiveness against human judgments. Pauli et al. (2025) trained a regression model on annotator 6-point Likert ratings of text pair persuasiveness, achieving strong correlation with held-out human judgments (Spearman $\rho = 0.845$). Elaraby et al. (2024) found LLM-judge rankings largely agreed with human evaluators on rationale persuasiveness ordering. Yeginbergen et al. (2025) reported strong correlation across three LLM-judges with human ratings on counter-argument quality (up to $\rho = 0.82$ for Claude 3.5 Sonnet). Furthermore, He et al. (2025) found strong alignment between human evaluators and LLM-judges, confirming that computationally derived success rates for Bayesian Persuasion (BP) models accurately reflect real-world human-perceived persuasiveness over baselines.

Table 2. Human validation summary for the reviewed studies. Alignment is rated High ($\geq$ 80% agreement or equivalent correlation), Moderate ($\sim$ 50–80%), or Low (no significant or negative correlation with human judgements).

	Study	What was validated?	Alignment
Argument	Pauli et al.	relative persuasiveness scoring (regression model)	High
	Elaraby et al.	LLM-judge persuasiveness rankings	High
	Yeginbergen et al.	LLM-judge multi-dimensional counter-argument scoring	High
	He et al.	LLM-judge argument scoring across 5 rhetorical dimensions	High
Belief/Behaviour	Kim et al.	behavioural outcome (purchase decisions)	High
	Yao et al.	reliability of self-reported persuasion ratings	High
	Cheng & You	persuadee belief update direction and magnitude	High
	Jaipersaud et al.	persuasion trajectory accuracy, label alignment	High
	Singh et al.	Oracle judge reliability	High
		engagement prediction accuracy	Moderate
	Song & Wang	donation amount prediction (regression model)	Low
	Namikoshi et al.	simulated persuadee attitude alignment	Low
	Durmus et al.	simulated belief shift accuracy	Low
	Doudkin et al.	human susceptibility modelling accuracy	Low

Belief change and behavioural outcome validation Five out of nine validated methods showed strong alignment with human judgements. Kim et al. (2025) found over 90% agreement between simulators and humans on both recommendation acceptance and out-of-budget purchase decisions, Yao et al. (2026) achieved 90% expert agreement on simulated patient rating change justifiability, and Cheng & You (2025) found 85.23% accuracy for belief update direction reasonableness (45 Prolific annotators). Jaipersaud et al. (2025) validated linear probes against actual donation outcomes (AUROC) on the *PersuasionForGood* dataset, finding peak performance at conversation midpoint (turns 8–12), matching when humans typically signal donation intent. Probes also correctly assigned higher persuasion probabilities to utterances human annotators had labelled as persuasive. Finally, to validate predicting persuasive impact, Singh et al. (2024) trained a model achieving 80.9% pairwise accuracy on 1.57 million post pairs. The validation of its cross-domain performance was considered successful, matching or exceeding frontier models on social media and market blog engagement, as well as opinion change prediction in online debates. Additionally, the Oracle judge achieved 82.3% accuracy in pairwise engagement judgement, and arena Elo rankings were confirmed consistent with large-scale human feedback studies.

Results were not uniformly strong, however. Song & Wang (2024) found low correlation ($r = 0.31$) between predicted and actual donation amounts from human-to-human dialogues. Namikoshi et al. (2024) found weak alignment ($r = 0.22$, $\rho = 0.23$) when comparing LLM virtual participants against human survey responses on battery-electric vehicle interventions, with virtual participants exhibiting systematically higher pro-environmental preferences and failing to capture low preferences or large shifts. Durmus et al. (2024) found significant misalignment when 3,832 human participants' belief shifts (1-7 Likert scale) were compared against model simulations, attributing failure to model sycophancy, self-generated argument bias, and lack of pragmatic reasoning. Most strikingly, Doudkin et al. (2025) identified a "synthetic persuasion paradox" through a randomised controlled trial (RCT) comparing 1,200 real humans against 1,200 simulated humans (based on real data) and 1,200 synthetic personas on pro-environmental interventions. Both synthetic and simulated agents exhibited exaggerated susceptibility, artificially inflating effect sizes, with strong negative correlation between synthetic and human responses. Simulated agents (grounded in real data) provided better directional approximation but still failed to capture human cognitive inertia.

5. Challenges, Risks & Future Directions

This section synthesises key challenges and limitations across automated LLM persuasion evaluation methods, examines associated risks, and outlines future directions.

5.1. Challenges

Scope Limitations Current frameworks exhibit narrow demographic reach, focusing heavily on Western, English-only contexts, with evaluations often restricted to niche domains (charity donations, crowdfunding, specific social media platforms), limiting generalisability and cross-domain comparison (Bozdag et al., 2026; Singh et al., 2024). These limitations affect all methods throughout the taxonomy.

Human Validation and Alignment Many methods lack robust human validation or rely on narrow annotator pools (Pauli et al., 2025). Fewer than half of the reviewed methods have human validation, with around a third of these

showing weak alignment with human judgements, undermining confidence in their real-world applicability. This weak alignment revealed through human validation studies (in Section 4) stems largely from the synthetic-human mismatch – the discrepancy in how simulated LLM agents respond to persuasion compared to humans. Doudkin et al. (2025) identified a "synthetic persuasion paradox" where both synthetic and simulated agents exhibited exaggerated susceptibility. Durmus et al. (2024) found similar misalignment, attributing it to sycophantic tendencies in LLMs. Even personalised profiles fail to replicate human cognitive variability, emotional inertia, and contextual reasoning, limiting LLM-to-LLM evaluations as proxies for human persuasion (Bozdag et al., 2025; Wang et al., 2025).

Measurement Reliability A further limitation concerns the reliability of the automated measurement instruments, particularly LLM-judge and self-evaluating persuadee agents. LLM-judges often fail to accurately assess persuasive dimensions in zero-shot settings: Barkar et al. (2025) showed regression models utilising hand-crafted lexical features can outperform LLM judges, while Bozdag et al. (2025) reported only 50% accuracy in pairwise argument ranking. Self-reported measures face different problems: sycophancy and self-generated bias distort results (Durmus et al., 2024). Beyond evaluator issues, metrics often lack granularity to capture persuasion's complexity. Coarse proxies like win rates oversimplify complex psychological effects and fail to capture subtle shifts (Ramani et al., 2024; Singh et al., 2024). Finally, even apparently successful validations may be misleading: data contamination threatens validation integrity – frontier LLMs trained on massive datasets may have encountered validation materials during pre-training, turning reasoning tasks into retrieval (Rescala et al., 2024).

Oversimplification of Persuasive Contexts Current evaluation methods fall short of capturing the multifaceted nature of real-world persuasion. Most critically, they exhibit a strong short-term bias: evaluations typically assess immediate belief shifts rather than longitudinal effects. Doudkin et al. (2025) found humans resistant to single- and multi-turn interventions while synthetic agents showed immediate susceptibility, indicating authentic change requires longitudinal interventions – yet current evaluations rarely extend beyond single sessions. The oversimplification extends beyond the temporal scope to the evaluation design itself. Dataset-based methods assess persuasion under predefined notions of success that may not generalise well to diverse real-world contexts, while game-based benchmarks remain simplified abstractions that fail to capture psychological nuance (Bailis et al., 2024). Finally, this review focuses on text-based methods, reflecting the field's current predominance, but real-world persuasion is inherently multimodal. Audio and visual cues play critical roles yet remain largely absent from current benchmarks (Bassi et al., 2024; Bozdag et al., 2026).

Practical Constraints Simulation-based and LLM-judge methods requiring repeated inference face significant computational costs. Large-scale multi-turn evaluations require forward passes for every query, making them resource-intensive – a constraint that has forced researchers to exclude frontier models or limit evaluation scope due to API costs alone (Bozdag et al., 2025; Cheng & You, 2025; Duffy et al., 2026).

Conceptual Ambiguity The field lacks a unified definition of persuasiveness, undermining comparison and interpretation. Studies measure different dimensions (stance change, argument quality, behavioural engagement) yet often the results are interpreted as general persuasiveness, making cross-benchmark comparisons unreliable. This fragmentation hinders evidence-based governance, where policymakers require consistent metrics for risk assessment.

5.2. Risks

While evaluating the persuasive capabilities of LLMs is essential for developing safety guardrails, automated evaluation methods introduce significant dual-use risks and may generate misleading safety signals.

Dual-Use and Malicious Optimisation The fundamental risk is that evaluation frameworks can be weaponised. Methods designed to measure persuasion can be repurposed as objective functions for Reinforcement Learning, enabling malicious actors to optimise models for maximum manipulative impact (Singh et al., 2024; Bozdag et al., 2025). This risk extends beyond general capability enhancement: automated evaluators that can predict which arguments convince specific demographics can be immediately repurposed for mass microtargeting and tailor-made propaganda campaigns at scale (Rescala et al., 2024). Finally, the public release of benchmarks containing explicit examples of harmful content, such as manipulative rhetoric, misinformation, or incitement to violence, inadvertently provides adversaries with high-quality datasets to fine-tune models for the very behaviours researchers aim to prevent (Bozdag et al., 2025; Pauli et al., 2025).

Structural Blind Spots and False Safety Signals Flawed evaluation methods risk creating false confidence in unsafe systems. Whilst automated evaluators achieve high alignment with human judgements on coarse-grained argument assessment, they struggle with granular persuasive capability evaluation, like belief change magnitudes and behavioural outcomes. Compounding this measurement limitation, evaluation metrics often isolate narrow dimensions

of persuasion, while results are interpreted as reflecting an LLM's general persuasive capability. Moreover, predominant single-session benchmarks systematically miss longitudinal risks such as gradual steering and trust-building strategies that emerge only across multi-session interactions (Cheng & You, 2025; OpenAI, 2024), allowing manipulative agents to pass short-context safety evaluations whilst remaining harmful in extended real-world deployment. These evaluation failures are further compounded by a critical behavioural vulnerability – the "refusal gap", which shows that models trained to refuse harmful instructions often comply when the same requests are framed as persuasion tasks (Kowal et al., 2025). Together, these compounding failures – evaluation limitations, conceptual fragmentation, temporal blind spots, and behavioural vulnerabilities – risk deploying persuasively manipulative models under false safety assurance.

5.3. Future Directions

Building on challenges and risks outlined above, we highlight three directions for advancing automated evaluation.

Expanding Human Validation and Improving Alignment. Fewer than half of the reviewed methods validate against human judgments, and a third of those show weak alignment. To confidently rely on automated metrics, human validation must become a baseline requirement. Beyond covering validation, future work must address the persistent low alignment in granular metrics – belief change magnitudes, behavioural outcomes, and susceptibility modelling. This may require developing LLM persuadee agents that better capture human cognitive inertia and resistance (Doudkin et al., 2025; Durmus et al., 2024; Namikoshi et al., 2024; Bozdag et al., 2026). However, creating more "human-like" synthetic persuadees carries dual-use risks and therefore must be approached with caution.

Expanding Scope: Settings, Modalities, and Capabilities. Evaluations must extend beyond single-session snapshots to longitudinal designs that track persuasion effects over longer periods, assessing how models build trust, adapt to resistance, and induce durable belief shifts (Durmus et al., 2024; Bozdag et al., 2026; Wang et al., 2025; Doudkin et al., 2025). Beyond temporal extension, evaluation scope should broaden across contextual and modal dimensions: (1) diverse populations and domains – multilingual, cross-cultural, and demographically varied contexts; and (2) multimodal benchmarks incorporating audio-visual cues beyond text-only evaluation (Bassi et al., 2024; Bozdag et al., 2026). Additionally, evaluations should assess personalisation and microtargeting – high-risk features with potential of manipulation at scale (Jones & Bergen, 2026).

Addressing Conceptual Fragmentation. Establishing shared definitions of persuasiveness would address current fragmentation, where studies measure different dimensions in different settings, yet often interpret results as reflecting general capabilities. Such standardisation is essential for comparable assessment and evidence-based AI governance.

6. Conclusion

As LLMs grow increasingly persuasive – in some settings surpassing human performance – scalable evaluation frameworks are essential for understanding societal impact, mitigating risks, and ensuring regulatory compliance. This structured review provides the first systematic taxonomy of fully automated LLM persuasion evaluation methods, organising 30 approaches from 27 papers into dataset-based and simulation-based categories, and examining their methodological designs, human validation, limitations, and risks. The field remains fragmented: methods vary widely in design and metrics, making cross-study comparison difficult. Human validation is limited and inconsistent, and several methods fail to replicate human persuasion dynamics entirely, raising concerns about reliance on synthetic proxies for safety-critical assessments. These limitations are compounded by structural risks: evaluation frameworks may be repurposed as optimisation targets for manipulative behaviour, whilst oversimplified benchmarks risk producing false safety signals.

Future work would benefit from expanding human validation and improving alignment, broadening evaluation across temporal, demographic, and modal dimensions, testing high-risk capabilities like personalisation, and establishing standardised measurement frameworks. Though current validation results reveal some limitations in automated methods, their value should be considered beyond replacing human-participant studies, for applications such as preliminary screening, high-risk scenario testing, and comparative benchmarking. This review is complemented by a living online resource hub to support future research.

Impact Statement

This review examines evaluation methods for LLM persuasiveness to support safer AI development and governance. It introduces no new frameworks or datasets that could be exploited adversarially. The accompanying resource hub aggregates public materials to support future evaluation research. While evaluation frameworks carry dual-use risks, we believe transparent analysis of existing methods is essential for developing trustworthy assessment tools and informing responsible AI governance.

References

Bailis, S., Friedhoff, J., and Chen, F. Werewolf Arena: A Case Study in LLM Evaluation via Social Deduction, July 2024. Preprint at http://arxiv.org/abs/2407.13943.

Barkar, A., Chollet, M., Labeau, M., Biancardi, B., and Clavel, C. Decoding Persuasiveness in Eloquence Competitions: An Investigation into the LLM's Ability to Assess Public Speaking. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence*, pp. 538–546, Porto, Portugal, 2025. SCITEPRESS - Science and Technology Publications. ISBN 978-989-758-737-5. doi: 10.5220/0013158400003890. URL https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0013158400003890.

Bassi, D., Fomsgaard, S., and Pereira-Fariña, M. Decoding Persuasion: A Survey on ML and NLP Methods for the Study of Online Persuasion. *Frontiers in Communication*, 9:1457433, 2024.

Becker, M., Sommer, M., Tapken, L., and Teh, Y. W. The Moralization Corpus: Frame-Based Annotation of Moralizing Speech Acts, 2025. Preprint at https://arxiv.org/abs/2512.15248v1.

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., et al. International AI Safety Report, January 2025. Preprint at http://arxiv.org/abs/2501.17805.

Bengio, Y., Clare, S., Prunkl, C., Murray, M., Andriushchenko, M., Bucknall, B., Bommasani, R., Casper, S., Davidson, T., Douglas, R., Duvenaud, D., Fox, P., Gohar, U., Hadshar, R., Ho, A., Hu, T., Jones, C., Kapoor, S., Kasirzadeh, A., Manning, S., Maslej, N., Mavroudis, V., McGlynn, C., Moulange, R., Newman, J., Ng, K. Y., Paskov, P., Rismani, S., Sastry, G., Seger, E., Singer, S., Stix, C., Velasco, L., Wheeler, N., Acemoglu, D., Conitzer, V., Dietterich, T. G., Felten, E. W., Heintz, F., Hinton, G., Jennings, N., Leavy, S., Ludermir, T., Marda, V., Margetts, H., McDermid, J., Munga, J., Narayanan, A., Nelson, A., Neppel, C., Ramchurn, S. D., Russell, S., Schaake, M., Scholkopf, B., Soto, A., Tiedrich, L., Varoquaux, G., Yao, A., Zhang, Y.-Q., Aguirre, L. A., Ajala, O., Albalawi, F., AlMalek, N., Busch, C., Collas, J., de Carvalho, A. C. P. d. L. F., Gill, A., Hatip, A. H., Heikkila, J., Johnson, C., Jolly, G., Katzir, Z., Kerema, M. N., Kitano, H., Kruger, A., Lee, K. M., Lopez Portillo, J. R., McLysaght, A., Molchanovskyi, O., Monti, A., Nemer, M., Oliver, N., Pezoa, R., Plonk, A., Ravindran, B., Riza, H., Rugege, C., Sheikh, H., Wong, D., Zeng, Y., Zhu, L., Privitera, D., and Mindermann, S. International AI safety report 2026. Technical Report DSIT 2026/001, Department for Science, Innovation and Technology, 2026. URL https://internationalaisafetyreport.org.

Bozdag, N. B., Mehri, S., Tur, G., and Hakkani-Tür, D. Persuade Me if You Can: A Framework for Evaluating Persuasion Effectiveness and Susceptibility Among Large Language Models, March 2025. URL http://arxiv.org/abs/2503.01829. arXiv:2503.01829 [cs].

Bozdag, N. B., Mehri, S., Yang, X., Ha, H., Cheng, Z., Durmus, E., You, J., Ji, H., Tur, G., and Hakkani-Tür, D. Must Read: A Comprehensive Survey of Computational Persuasion. *ACM Computing Surveys*, March 2026. ISSN 0360-0300. doi: 10.1145/3800687. URL https://doi.org/10.1145/3800687.

Breum, S. M., Egdal, D. V., Mortensen, V. G., Møller, A. G., and Aiello, L. M. The Persuasive Power of Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:152–163, May 2024. ISSN 2334-0770. doi: 10.1609/icwsm.v18i1.31304. URL https://ojs.aaai.org/index.php/ICWSM/article/view/31304.

Cheng, Z. and You, J. Towards Strategic Persuasion with Language Models, September 2025. Preprint at http://arxiv.org/abs/2509.22989.

Costa, D. B. and Vicente, R. Deceive, Detect, and Disclose: Large Language Models Play Mini-Mafia, 2025. Preprint at https://arxiv.org/abs/2509.23023.

Doudkin, A., Pataranutaporn, P., and Maes, P. AI Persuading AI vs AI Persuading Humans: LLMs' Differential Effectiveness in Promoting Pro-Environmental Behavior, March 2025. Preprint at http://arxiv.org/abs/2503.02067.

Duffy, A., Paech, S. J., Shastri, I., Karpinski, E., Alloui-Cros, B., Marques, T., and Olson, M. L. Democratizing Diplomacy: A Harness for Evaluating Any Large Language Model on Full-Press Diplomacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37350–37359, 2026.

Durmus, E., Lovitt, L., Tamkin, A., Ritchie, S., Clark, J., and Ganguli, D. Measuring the Persuasiveness of Language Models, 2024. URL https://www.anthropic.com/news/measuring-model-persuasiveness.

El-Sayed, S., Akbulut, C., McCroskery, A., Keeling, G., Kenton, Z., Jalan, Z., Marchal, N., Manzini, A., Shevlane, T., Vallor, S., Susser, D., Franklin, M., Bridgers, S., Law, H., Rahtz, M., Shanahan, M., Tessler, M. H., Douillard, A., Everitt, T., and Brown, S. A Mechanism-Based

Approach to Mitigating Harms from Persuasive Generative AI, April 2024. Preprint at http://arxiv.org/abs/2404.15058.

Elaraby, M., Litman, D., Li, X. L., and Magooda, A. Persuasiveness of Generated Free-Text Rationales in Subjective Decisions: A Case Study on Pairwise Argument Ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14311–14329, 2024.

European Commission. General-Purpose AI Code of Practice, July 2025. URL https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), 2024. URL https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

Fogg, B. J. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann, San Francisco, 2003.

Habernal, I. and Gurevych, I. Which Argument Is More Convincing? Analyzing and Predicting Convincingness of Web Arguments Using Bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1589–1599, 2016.

He, B., Liu, Y., Zhang, Z., Jia, Z., Wu, H., He, Z., Zheng, Z., and Kang, Y. Make an Offer They Can't Refuse: Grounding Bayesian Persuasion in Real-World Dialogues Without Pre-Commitment, October 2025. Preprint at http://arxiv.org/abs/2510.13387.

Hong, J., Dragan, A., and Levine, S. Planning Without Search: Refining Frontier LLMs with Offline Goal-Conditioned RL, 2025. Preprint at https://arxiv.org/abs/2505.18098.

Idziejczak, M., Korzavatykh, V., Stawicki, M., Chmutov, A., Korcz, M., Bładek, I., and Brzezinski, D. Among Them: A Game-Based Framework for Assessing Persuasion Capabilities of LLMs. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence*, volume 15874, pp. 183–195, Singapore, 2025. Springer. doi: 10.1007/978-981-96-8186-0_15.

Jaipersaud, B., Krueger, D., and Lubana, E. S. How Do LLMs Persuade? Linear Probes Can Uncover Persuasion Dynamics in Multi-Turn Conversations, August 2025. Preprint at http://arxiv.org/abs/2508.05625.

Jin, C., Ren, K., Kong, L., Wang, X., Song, R., and Chen, H. Persuading Across Diverse Domains: A Dataset and Persuasion Large Language Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1678–1706, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.92. URL https://aclanthology.org/2024.acl-long.92.

Jones, C. R. and Bergen, B. K. Lies, Damned Lies, and Language Statistics: A Comprehensive Review of Risks from Manipulation, Persuasion, and Deception with Large Language Models. *Artificial Intelligence Review*, 59:116, 2026. doi: 10.1007/s10462-026-11517-6.

Ju, T., Chen, Y., Fei, H., Lee, M.-L., Hsu, W., Cheng, P., Wu, Z., Zhang, Z., and Liu, G. On the Adaptive Psychological Persuasion of Large Language Models, 2025. Preprint at https://arxiv.org/abs/2506.06800.

Kim, T., Lee, J., Yoon, S., Kim, S., and Lee, D. Towards Personalized Conversational Sales Agents: Contextual User Profiling for Strategic Action. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5131–5154, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.275. URL https://aclanthology.org/2025.findings-emnlp.275/.

Kowal, M., Timm, J., Godbout, J.-F., Costello, T., Arechar, A. A., Pennycook, G., Rand, D., Gleave, A., and Pelrine, K. It's the Thought that Counts: Evaluating the Attempts of Frontier LLMs to Persuade on Harmful Topics, August 2025. Preprint at http://arxiv.org/abs/2506.02873.

Li, Z., Zheng, C., and Shi, Y. Data Speaks, but Who Gives It a Voice? Understanding Persuasive Strategies in Data-Driven News Articles. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

Liu, M., Xu, Z., Zhang, X., An, H., Qadir, S., Zhang, Q., Wisniewski, P. J., Cho, J.-H., Lee, S. W., Jia, R., and Huang, L. LLM Can Be a Dangerous Persuader: Empirical Study of Persuasion Safety in Large Language Models. In *Proceedings of the Second Conference on Language Modeling*, 2025.

Ma, Y., Zhang, X., Ren, J., Wang, R., and Wang, M. Linguistic Features of AI Mis/Disinformation and the Detection Limits of LLMs. *Nature Communications*, 16:67145, 2025. doi: 10.1038/s41467-025-67145-1.

Maier, S., Feuerriegel, S., and Hölbling, L. A Meta-Analysis of the Persuasive Power of Large Language Models. *Scientific Reports*, 15(1):30783, 2025.

Meguellati, E., Civelli, S., Han, L., and Bernstein, A. LLM-Generated Ads: From Personalization Parity to Persuasion Superiority, 2025. Preprint at https://arxiv.org/abs/2512.03373v1.

Namikoshi, K., Shamma, D. A., Iliev, R., Fang, J., Filipowicz, A., Hogan, C. L., Wu, C., and Arechiga, N. Leveraging Language Models and Bandit Algorithms to Drive Adoption of Battery-Electric Vehicles, 2024. Preprint at https://arxiv.org/abs/2410.23371.

O'Keefe, D. J. *Persuasion: Theory and Research*. Sage Publications, Thousand Oaks, CA, 2015.

OpenAI. Openai o1 system card. System card, OpenAI, December 2024. URL https://cdn.openai.com/o1-system-card.pdf.

Pauli, A. B., Augenstein, I., and Assent, I. Measuring and Benchmarking Large Language Models' Capabilities to Generate Persuasive Language. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10056–10075, 2025.

Ramani, G. P., Karande, S., V, S., and Bhatia, Y. Persuasion Games Using Large Language Models, September 2024. Preprint at http://arxiv.org/abs/2408.15879.

Rescala, P., Ribeiro, M. H., Hu, T., and West, R. Can Language Models Recognize Convincing Arguments? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8826–8837, 2024.

Rogiers, A., Noels, S., Buyl, M., and De Bie, T. Persuasion with Large Language Models: A Survey, 2024. Preprint at https://arxiv.org/abs/2411.06837v1.

Salvi, F., Horta Ribeiro, M., Gallotti, R., and West, R. On the Conversational Persuasiveness of GPT-4. *Nature Human Behaviour*, 9(8):1645–1653, 2025. doi: 10.1038/s41562-025-02194-6.

Simpson, E. and Gurevych, I. Finding Convincing Arguments Using Scalable Bayesian Preference Learning. *Transactions of the Association for Computational Linguistics*, 6:357–371, 2018.

Singh, S., Singla, Y. K., SI, H., and Krishnamurthy, B. Measuring and Improving Persuasiveness of Large Language Models, October 2024. Preprint at http://arxiv.org/abs/2410.02653.

Song, Y. and Wang, H. Would You Like to Make a Donation? A Dialogue System to Persuade You to Donate. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 17707–17717, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1540/.

Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., Jacovi, M., Aharonov, R., and Slonim, N. Automatic Argument Quality Assessment - New Datasets and Methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5625–5635, 2019.

Wang, J., Yao, Z., Li, L., Qian, J., Yang, Z., and Yu, H. ChatThero: An LLM-Supported Chatbot for Behavior Change and Therapeutic Support in Addiction Recovery, October 2025. Preprint at http://arxiv.org/abs/2508.20996.

Yao, Z., Chafekar, T., Wang, J., Han, S., Ouyang, F., Qian, J., Li, L., and Yu, H. ChatCLIDS: Simulating Persuasive AI Dialogues to Promote Closed-Loop Insulin Adoption in Type 1 Diabetes Care. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 39539–39547, 2026.

Yeginbergen, A., Oronoz, M., and Agerri, R. Dynamic Knowledge Integration for Evidence-Driven Counter-Argument Generation with Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22568–22584, 2025.

Yu, F., Jiang, L., Huang, S., Wu, Z., and Dai, X. PersuasiveToM: A Benchmark for Evaluating Machine Theory of Mind in Persuasive Dialogues, May 2025. Preprint at http://arxiv.org/abs/2502.21017.

Zhu, K., Du, H., Hong, Z., Yang, X., Guo, S., Wang, D. Z., Wang, Z., Qian, C., Tang, X., Ji, H., et al. Multiagentbench: Evaluating the Collaboration and Competition of LLM Agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8580–8622, Vienna, Austria, 2025. Association for Computational Linguistics.

A. Detailed Scope & Definitions

This appendix provides detailed definitions of key terms used throughout this review.

A.1. Persuasion as an Umbrella Term

This paper focuses on the automated evaluation of LLMs' ability to influence human beliefs, attitudes, and behaviour through text-based communication. To capture this range of phenomena within a single coherent framework, we use **persuasion** as an umbrella term encompassing all forms of communicative influence that operate through cognitive and decision-making processes, including manipulation and deception, where it is used as a technique of persuasive influence. We adopt this framing following Jones & Bergen (2026), in order to have a term more general than rational persuasion and manipulation, but more specific than influence, which also incorporates related concepts such as coercion and exploitation. The term is value-neutral, capturing prosocial and harmful influence within a single coherent category, and reflects how the broader literature is organised (Jones & Bergen, 2026; Bozdag et al., 2026).

More precisely, persuasion is defined following O'Keefe (2015) as "a successful intentional effort at influencing another's mental state through communication in a circumstance in which the persuadee has some measure of freedom."

A.2. Rational Persuasion vs Manipulation

The persuasion term in academic studies is further split into **rational persuasion**, characterised by transparent and autonomy-preserving arguments, and **manipulation**, which covertly exploits the persuadee's decision-making vulnerabilities in ways the target is not fully aware of and therefore cannot consent to it (Jones & Bergen, 2026).

A.3. Deception

A related concept, **deception**, refers broadly to causing someone to hold false beliefs and is treated, following Jones & Bergen (2026) and El-Sayed et al. (2024), as a special case of manipulation. However, not all deception is persuasive in intent – hallucinations and task-driven deceptive AI behaviour, for instance, are not oriented toward shifting beliefs or behaviour (Jones & Bergen, 2026). This study, therefore, includes only deception that is deliberately used as a technique of persuasive influence.

B. Detailed Search Strategy

This appendix provides details of our eligibility criteria and search methodology.

B.1. Eligibility Criteria Clarifications

Given the early and rapidly evolving state of the field, peer-reviewed publications, preprints, and blog posts were all considered eligible to ensure comprehensive coverage of emerging methods (where a paper was initially identified as a preprint but subsequently published in a peer-reviewed venue by the time of writing, the reference was updated to reflect the published version).

To be included, a method was required to produce a concrete measure of persuasive effectiveness (such as a score, rate, or outcome), rather than analysing persuasion phenomena such as persuasive strategy classification or detection (e.g., Yu et al. 2025). This criterion applied regardless of whether persuasion evaluation was the primary contribution or secondary to another objective, such as developing a more persuasive LLM (e.g., Jin et al. 2024). We restricted our analysis to text-based outputs, reflecting the current state of the field in which fully automated evaluation methods for multimodal persuasion remain limited.

B.2. Search Strategy Details

To ensure comprehensive coverage of relevant literature, we employed a multifaceted search strategy combining traditional academic databases, preprint repositories, and LLM-assisted discovery.

We performed a traditional database search across three major scholarly platforms: Scopus, ACL Anthology, and ArXiv on 10 January 2026, using the following query to search through titles, abstracts and keywords:

```
("persuasion" OR "persuasiveness" OR "persuasive language") AND
("evaluation" OR "assessment" OR "scoring" OR "benchmark" OR "measuring") AND
("language model" OR "LLM" OR "AI")
```

After filtering for publication year ($\geq$2021) and language (English only), 116 papers remained from Scopus, 209 from ArXiv, and 28 from ACL Anthology.

B.3. LLM-Assisted Filtering

For all three database sources, an LLM (GPT-5.2) was used to label each paper as "yes" (to be included), "no", or "maybe" given its title and abstract. The prompt used was:

```
Your Task: Determine if the paper proposes or uses a 100% automated method for
    evaluating LLM persuasiveness.

Possible Labels:
* YES: The paper evaluates LLM persuasiveness without human judgement in the
    evaluation process (this excludes human participation in validating the
    methodology).
* NO: The paper does not contain an automated LLM persuasiveness evaluation
    technique.
* MAYBE: The title and abstract lack detail to confirm if there is an automated LLM
    persuasion evaluation method present.

Output: Provide only one word label: YES, NO, or MAYBE.

Paper to Evaluate:
Title: {Title}
Abstract: {Abstract}
```

Across all three sources, 110 were labelled "yes", 145 "no", and 98 "maybe".

To verify label reliability, a human reviewer manually inspected a stratified sample of 15 papers (five per label category). Although the LLM incorrectly labelled several irrelevant papers as relevant, none of the papers it labelled as "no" were found to be relevant upon manual inspection, providing limited but encouraging evidence that this LLM-assisted labelling process is effective for exclusion.

A final manual screening (of papers labelled "yes" and "maybe") produced 9 papers from Scopus, 20 from ArXiv, and 3 from ACL Anthology. After deduplication, we were left with 22 papers.

B.4. LLM-Assisted Search

To further reduce the chance of missing relevant work, we conducted an LLM-assisted search using GPT-5.2 and Google Gemini 3 Pro in Deep Research mode. The prompt used was:

```
Your Task: Identify peer-reviewed research, pre-prints, or technical frameworks in
    any other format that propose fully automated methods for evaluating LLM and
    GPAI persuasive capabilities. Provide me with a list of at least 20 papers, if
    they meet the criteria.

Inclusion Criteria: The method must be 100% automated.

Exclusion Criteria: Exclude any methods that require human judgement in the
    evaluation stage (validation is accepted).

Output Format: Provide a table with these columns: Paper Title; Short Evaluation
    Method Description; Link
```

This search resulted in 4 additional papers.

B.5. Backwards Reference Searching

Finally, we performed backwards reference searching on a subset of key papers selected as the most comprehensive existing surveys on LLM persuasion (Bozdag et al., 2026; Rogiers et al., 2024; Bassi et al., 2024), which yielded 1 further paper.

B.6. Final Corpus

In total, we identified **27 papers** within the scope, with 3 papers presenting multiple evaluation methods, resulting in **30 distinct evaluation methods**.

C. Persuasion Evaluation Resource Hub

Figure 2 shows an anonymised preview of the online resource hub. The website allows filtering studies by available resources (datasets, codebases) or by evaluation features (personalised persuasion, persuasion strategies).

Figure 2. Anonymised preview of the persuasion evaluation resource hub.