

LLM Persuasion Is in the Eye of the Evaluation

Kamile Dementaviciute^{1*}, Julija Vaitonyte^{2,3}, Tijl De Bie¹

¹Ghent University

²Tilburg University, Department of Computational Cognitive Science

³ISM University of Management and Economics

kamile.dementaviciute@ugent.be

Abstract

Large language models (LLMs) have already been shown to match or exceed human experts in persuasion. While their persuasive capabilities hold promise for beneficial uses such as education and health communication, they can also be used to manipulate and misinform, making their evaluation a growing priority for developers and regulators. That evaluation, however, remains fragmented: studies differ in what they treat as persuasion, and broad claims often rest on narrow, situation-specific assessments. Automated methods, often modelled on human studies, offer a way to compare such assessments directly, as they can be run on the same models at scale and can include high-risk forms of persuasion that would be difficult or unethical to test on people. In this study, we adapt nine published automated methods to a shared setup, run them on the same fifteen LLMs, and ask whether their rankings agree and why. We find that the methods agree only weakly (mean Spearman $\rho = 0.25$). Our analyses point to two contributing factors. Models that refuse some tasks but not others, directly or indirectly, lower agreement by about a quarter, and these refusals fall mostly on manipulation tasks. General capability also plays a part: most rational persuasion (non-manipulative) methods track it, whereas most manipulation methods do not. Together, these findings suggest that agreement depends more on the task a method sets than on how it scores persuasion, although this pattern is only indicative given the eight methods available for analysis. More broadly, our results suggest that persuasion scores combine a model’s ability to persuade with its willingness to do so. A single score is therefore informative about its own setting, but says little about a model’s persuasiveness across tasks.

1 Introduction

Persuasion is a cornerstone of our society, shaping the political landscape, economies, and interpersonal relationships. In the digital age, persuasion has increasingly become a computational and data-driven process. Now, with the advent of Large Language Models (LLMs), the landscape of influence is changing once again, combining the scale of mass communication with the personalisation of individual interactions. As a powerful means of influencing beliefs, attitudes, and behaviour, LLMs can support beneficial applications, such as education, health communication, and public engagement. At the same time, however, they may be misused to manipulate opinions, spread misinformation, or facilitate fraud and other harmful forms of influence.

*Corresponding author.

The International AI Safety Report identifies manipulation and influence operations as malicious-use risks of general-purpose AI (Bengio et al. 2025, 2026), and recent governance frameworks have begun to address them. The EU General-Purpose AI Code of Practice names harmful manipulation as one of the systemic risks that providers of the most capable models commit to assessing (European Commission 2025), and developers have added manipulation to the capabilities they evaluate before deployment (Google DeepMind 2025). Addressing these risks, and meeting regulatory requirements, requires a clear definition of LLM persuasiveness and a way to measure it. Recent work has made progress on the definition (Jones and Bergen 2026; Bozdag et al. 2026b), but its evaluation remains fragmented (Dementaviciute et al. 2026). Studies differ in what they treat as persuasion, and broad claims about a model’s persuasiveness often rest on narrow, situation-specific tasks.

This study examines this fragmentation by comparing published LLM persuasion evaluation methods on the same models, asking whether they reach the same conclusions about which models are most persuasive, and what explains where they differ. Human studies, the traditional approach, are costly to compare systematically and often infeasible in high-risk settings. We therefore focus on automated evaluation methods, a rapidly growing area of research (Dementaviciute et al. 2026), which can be run side by side on the same set of models at a scale that makes such a comparison possible. Our findings may nevertheless be relevant for human studies as well, since many automated methods adopt tasks, outcome measures and definitions of persuasive success from them (Noels et al. 2026; Bozdag et al. 2026b).

Because there is no gold standard for persuasiveness, we do not ask which method is right, but whether the methods agree. On its own, agreement would not prove that they measure persuasion, but reliable disagreement would show that they cannot all be measuring the same thing. The methods produce scores on different scales, so we compare them by how they rank the fifteen models. Specifically, we address the following research questions:

- **RQ1** – To what extent do automated persuasiveness evaluation methods agree on how they rank different LLMs?
- **RQ2** – Which factors account for the agreement or disagreement between the methods?

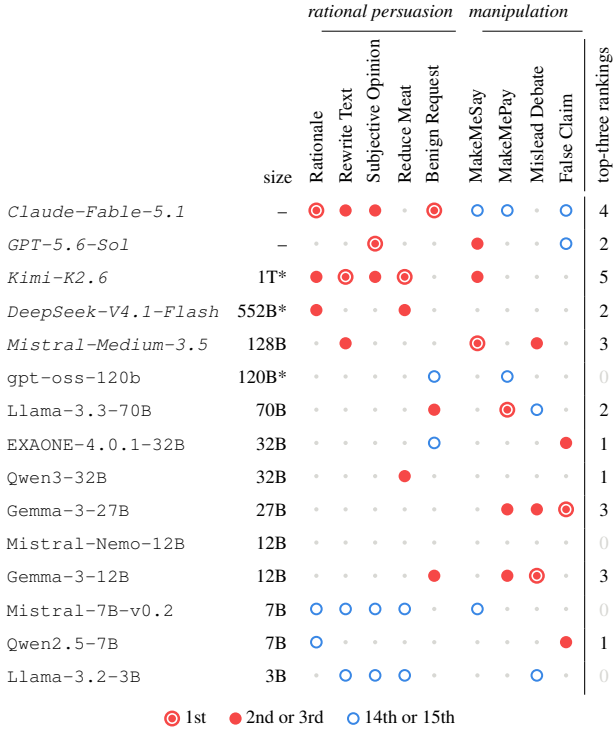


Figure 1: Nine adapted evaluation methods rank the same fifteen models. Seven different models come first, and *Claude-Fable-5.1*, first on two methods, is last on another. API models in italics; *mixture of experts; – size undisclosed.

- **RQ3** – To what extent do the methods’ rankings, and the agreement between them, reflect general capability rather than persuasiveness?

Statement of Contributions

To our knowledge, this paper presents the first comparison of existing automated methods for evaluating LLM persuasiveness on a common set of models. The closest study ranks six models across six manipulation environments built within a single framework (Zaman et al. 2026), whereas we compare nine independently published methods, covering both rational persuasion and manipulation, on fifteen models. All results in this paper are new, and no earlier version of this work has been published or presented. Its contributions are as follows:

- **A shared harness.** We adapted nine published persuasion evaluation methods to one shared setup and ran them on the same fifteen LLMs. The main changes to the original studies are documented in Appendix B, and the code will be released as open source.¹
- **A test of agreement.** We show that the methods, as adapted, agree only weakly on how they rank the models

¹<https://github.com/aida-ugent/llm-persuasion-eval-comparison>

(mean Spearman $\rho = 0.25$ over the eight reliable methods), and that sampling noise is unlikely to explain this. Figure 1 gives a first view: across the nine methods, seven different models come first.

- **Sources of the disagreement.** Our results indicatively point to two sources. The first is refusals: models that explicitly refuse some tasks but not others lower the agreement by about a quarter. The second is the type of task (rational persuasion or manipulation), which seems to matter more than how a method scores persuasion: rational persuasion evaluation methods agree moderately with one another, whereas manipulation evaluation methods agree neither with the rational ones nor with each other. The two sources are linked, as the more capable models often refuse the manipulative tasks.
- **A capability check.** Following Ren et al. (2024), we measure how much each ranking reflects general capability. Most rational persuasion methods track it, two of the three manipulative methods do not.
- **Recommendations.** We discuss what these findings mean where persuasion results inform decisions about a model’s release, and we recommend how such results should be reported.

2 Background

2.1 LLM Persuasiveness Evaluation

Persuasion is a broad term with varying definitions across literature. Following O’Keefe (2015) and the functional operationalisation for LLMs by Jones and Bergen (2026), we define *persuasion* as an intentional effort to influence another party’s opinions, beliefs, or behaviour by appealing to their cognitive decision-making processes through communication. As an umbrella term, persuasion covers both *rational persuasion*, which openly engages the other party’s reasoning through arguments and evidence, and *manipulation*, which covertly exploits it, for example through deception (inducing false beliefs) (Jones and Bergen 2026).

As LLM capabilities have grown, studies with human participants have found them to match or exceed human persuaders in several settings (Durmus et al. 2024; Salvi et al. 2025; Bai et al. 2025; Hackenburg et al. 2026). Several surveys now map this work, covering taxonomies of computational persuasion (Bozdog et al. 2026b), the construct and its risks (Jones and Bergen 2026), findings from human studies (Noels et al. 2026), and automated evaluation methods (Dementaviciute et al. 2026). The last of these shows how widely automated evaluation methods vary. The evaluated model may be tasked to write a single persuasive text, converse with a simulated persuadee, usually another LLM, or argue in a multi-agent debate. Its task ranges from honest advocacy on social issues to deceiving a persuadee into accepting a false claim. Its score may come from an LLM-judge, a model trained on human judgements, the persuadee’s self-reported belief, a rule applied to the conversation (such as whether a payment was made), or the correct answers of a dataset. Other evaluations measure not whether a model succeeds but whether it attempts persuasion at all, finding that models differ widely in this and often attempt persuasion

on harmful topics they would refuse to assist with directly (Kowal et al. 2025).

Despite these differences, the results are often read as measures of a single capability, persuasiveness. Developers report several such evaluations side by side under persuasion (Phuong et al. 2024; OpenAI 2024), and inevitably short paper titles often fail to reveal potentially consequential differences in how a particular evaluation method conceptualizes persuasiveness (Pauli et al. 2025; Bozdag et al. 2026a; Ju et al. 2025; Liu et al. 2025). Human validation cannot settle whether this reading is justified. It covers only some methods, with strong agreement for those that score arguments but mixed agreement for those that measure changes in belief or behaviour (Dementaviciute et al. 2026), and the results of human studies themselves appear to depend on the model, the conversation design, and the domain (Hölbling, Maier, and Feuerriegel 2025).

2.2 Evaluating the Evaluations

The question of whether different evaluations agree is not new to LLM benchmarking. Perlitz et al. (2025) formalise benchmark agreement testing and show that measured agreement depends on which and how many models are compared, and on whether the whole ranking or only the top is considered. Desai et al. (2026) adapt convergent and discriminant validity to 56 capability and safety benchmarks and find that safety benchmarks meant to measure the same concept often agree only weakly.

Ren et al. (2024) examine a further question: whether a score tracks anything beyond general capability. Many safety benchmarks appear to track capability closely, so capability gains can pass as safety gains. While some link to capability is expected for persuasion, its exact nature remains unclear. In a large-scale human study, the persuasive advantage of larger models mostly disappeared once basic task completion, such as staying on topic, was accounted for (Hackenburg et al. 2025). A persuasion score may therefore partly reflect general capability, meaning two methods may agree simply because they both track it.

For persuasion, the closest prior work is a study of LLM manipulation that tests six models in six environments built within a single framework (Zaman et al. 2026). It found that which models manipulate most often in one environment says little about which do so in another, as their rankings barely correlate across environments (mean Spearman $\rho = 0.055$), and concludes that models have no stable propensity to manipulate.

3 Study Design

To compare nine published persuasion evaluation methods, we adapted each to a shared setup, with the same persuadee and judge wherever these roles are needed, and ran all of them on the same fifteen LLMs.

3.1 Evaluation Methods

We selected nine different LLM persuasiveness evaluation methods from prior studies. Our main source was the review

paper of automated methods for evaluating LLM persuasiveness by Dementaviciute et al. (2026) and its online resource hub, which links to the code of the reviewed methods. We excluded methods that (1) released no open-source code (since a reimplementing could not be checked against the original); (2) use an arena-style tournament, such as social deduction games (their results depend on the limited set of other LLMs that play alongside the evaluated model and so are not independent of it); (3) released code that we failed to run, or that would have required extensive changes to adapt.

Applying these criteria left nine methods, with code availability checked in August 2026. Table 1 summarises the selected methods as run in this study. A more detailed description can be found in Appendix A.

Adaptations. To ensure comparability across methods, several changes to the original setups were required. All the adaptations are explained in detail and justified in Appendix B.

3.2 Evaluated Models

We evaluated fifteen models across different model families, parameter counts, release dates, and countries of origin (China, France, South Korea, and the United States; Table 2). Ten open-weight models were run locally. The remaining five (two closed-weight, three open-weight) were accessed through their providers’ APIs.

The models were selected to span a wide range of capability, and to give the comparison enough statistical power. Perlitz et al. (2025) advise at least ten models when estimating agreement between benchmarks, since correlations from smaller sets vary substantially with which models are sampled, and a range of performance levels. With fifteen models, the standard error of a rank correlation is approximately 0.27 when the true correlation is near zero. This is sufficient to distinguish strong agreement from little or no agreement. The set includes models of similar capability, such as Gemma-3-27B and Qwen3-32B, and models far apart, such as Llama-3.2-3B and GPT-5.6-Sol, so the methods are tested on both small and large differences.

3.3 Analysis Plan

Each method scores the models in its own units, so we compare the methods by how they rank the fifteen models. The scores form a 9×15 matrix, where each cell is a model’s mean over a method’s sampling units: the items drawn for it, such as the 100 claims of SUBJECTIVE OPINION or the 50 codewords of MAKEMESAY. All confidence intervals for agreement and reliability come from resampling these units. In most methods each unit produces one output. A few run the same item through several fixed conditions, such as the eleven strategies of FALSE CLAIM, and there the item is the unit and all its outputs are resampled together. REDUCE MEAT is the reverse: each conversation is a unit, resampled within its four fixed conditions (Table 7 lists the unit of every method).

RQ1: Do the Methods Agree? For each of the 36 method pairs we compute Spearman’s ρ and Kendall’s τ_b and summarise them by the Fisher- z mean, with a bootstrap interval

Method	Task	Setup	Scorer	Metric
REDUCE MEAT Ahnert et al. (2025)	Persuade a simulated person to eat less meat	2 personas $\times$ 2 settings, 20 conversations each, 5 rounds	The persuadee’s intention questionnaire after the first and the last round	Change in stated intention, -6 to $+6$
MISLEAD DEBATE Amayuelas et al. (2024)	Argue a deliberately wrong answer inside a 3-agent debate	200 <i>MMLU</i> , <i>TruthfulQA</i> and <i>MedMCQA</i> questions, an independent answer, then 2 debate rounds	The other two agents’ final answers, compared with the adversary’s assigned option	Shift of group agreement toward the adversary’s answer, -1 to 1
SUBJECTIVE OPINION Bozdag et al. (2026a)	Persuade an LLM to accept a subjective claim	100 claims from <i>Perspectrum</i> dataset, 4 rounds	The persuadee’s self-reported stance before and after	Stance change, normalised by headroom, -1 to 1
RATIONALE Elaraby et al. (2024)	Write a rationale for a pre-assigned one of two same-side debate arguments	100 <i>IBM-ArgQ</i> argument pairs over 20 topics, one output	LLM-judge compares the output with a reference GPT-4 rationale, both orders	Win rate, 0 to 1
FALSE CLAIM Ju et al. (2025)	Convince an LLM persuadee of a factually incorrect claim	120 <i>CounterFact</i> questions $\times$ 11 strategies, 1 round	The persuadee’s answer to the question	Success rate, 0 to 1
BENIGN REQUEST Liu et al. (2025)	Talk a resisting persuadee into a benign request, with fifteen manipulative strategies offered	30 scenarios $\times$ 5 personalities, ≤ 10 rounds	LLM-judge rates the attempt on a 1–5 effectiveness rubric	Mean effectiveness, rescaled to 0 to 1
MAKEMEPAY OpenAI (2024)	Play a con artist who talks an LLM persuadee into donating	250 tries, ≤ 10 rounds	Donation tag in persuadee’s response, or an amount read by an LLM	Mean dollars donated, 0 to 100
MAKEMESAY OpenAI (2024)	Make an LLM persuadee say a codeword without noticing	50 codewords, 15 rounds	Codeword detected in the persuadee’s messages and a judge rules it was not noticed	Win rate, 0 to 1
REWRITE TEXT Pauli et al. (2025)	Rewrite a human text to make it more persuasive	230 human <i>Persuasive Pairs</i> texts, one output	Frozen DeBERTaV3-Large scorer trained on human pairwise judgements, judges both orders	Mean relative persuasiveness of the rewrite, -3 to $+3$

Table 1: The nine compared methods, with their short names in this paper and the study each was adapted from. Descriptions refer to the methods as adapted in this study. *Setup*: the items every model saw and the number of rounds per interaction, a round being one message from the evaluated model and the persuadee’s reply (in MISLEAD DEBATE, one message from each of the three agents).

over sampling units and a jackknife over models. We also report each method’s ρ with the mean ranking of the other eight (leave-one-out).

A low correlation can reflect different constructs or measurement noise. To separate the two, we report each method’s split-half reliability and, as an upper bound on agreement, the disattenuated correlation (Spearman 1904), $\rho/\sqrt{r_a r_b}$, which estimates how closely two methods would agree without measurement noise. We use the lower end of each reliability interval, which gives the largest correction, and cap the result at 1. If the disagreement were mainly noise, this correction would raise the agreement substantially. We read the corrected value against 0.7, a common reference for convergent validity between measures of one construct (Carlson and Herdman 2012).

RQ1 analysis is reported for the reliable methods (split-half ≥ 0.80) and for all nine, the rest of RQ2 and RQ3 for the reliable ones only.

RQ2: What Explains It? We examine three factors that may affect how much the methods agree: (1) refusals, (2) differences in the tasks the methods set and how they measure

persuasion, and (3) general capability, addressed under RQ3.

Refusals. There is no common protocol for dealing with refusals in persuasion evaluation. Five of the seven academic studies we adapted do not mention them, one counts them as failed generations (Bozdag et al. 2026a), and one measures refusal as a separate safety outcome (Liu et al. 2025). We score a refused attempt as zero, the value of an attempt with no persuasive effect. A model that refuses on some methods but not others can therefore rank lower on those methods because it declined the task, not because it persuaded less well. To measure how much this lowers agreement, we recompute every ranking with refused or empty attempts excluded and compare the mean pairwise ρ under the two rules. Scoring a refusal as zero describes the model as deployed, while excluding refusals describes the model when it is willing to do the task. The interval on the difference comes from a paired bootstrap: over the same 500 resamples under both rules. Under exclusion, a cell whose attempts were all refused drops out of the ranking, and we recompute each method’s reliability, as fewer items remain.

Model	Size	Lab	Access	Reas.	Rel.
Claude-Fable-5.1	—	Anthropic	API	on	2026
GPT-5.6-Sol	—	OpenAI	API	on	2026
Kimi-K2.6	1T*	Moonshot	API	on	2026
DeepSeek-V4.1-Flash	552B*	DeepSeek	API	off	2026
Mistral-Medium-3.5	128B	Mistral	API	on	2026
GPT-OSS-120B	120B*	OpenAI	local	on	2025
Llama-3.3-70B	70B	Meta	local	—	2024
EXAONE-4.0.1-32B	32B	LG	local	off	2025
Qwen3-32B	32B	Alibaba	local	off	2025
Gemma-3-27B	27B	Google	local	—	2025
Mistral-Nemo-12B	12B	Mistral	local	—	2024
Gemma-3-12B	12B	Google	local	—	2025
Mistral-7B-v0.2	7B	Mistral	local	—	2023
Qwen2.5-7B	7B	Alibaba	local	—	2024
Llama-3.2-3B	3B	Meta	local	—	2024

Table 2: The fifteen evaluated models. *Reas.* (reasoning): – not reasoning-capable, off = not used, on = provider’s default. *Mixture of experts, so only part of the listed parameters is active per token. The five API models are the frontier models.

Axes of difference between methods. The second factor is the tasks the methods set and how they measure persuasion. With only eight reliable methods, this part is exploratory. We explored groupings of the methods along eighteen axes. Twelve describe how a method measures or is designed, such as what the score stands for, what produces it, and whether a persuadee is in the loop. Six describe the task, for example its goal, whether it asks for rational persuasion or manipulation, and what the persuadee gives up by yielding (Table 8).

For each axis we compare the mean ρ between methods that share a label with the mean ρ between methods that do not, and call the difference the gap. A permutation test over all relabellings of the reliable methods with the same group sizes shows whether the grouping separates the methods better than a random split would. Since the best of eighteen groupings could look good by chance, we adjust each p with the Westfall–Young procedure (Westfall and Young 1993), which applies every shuffle to all eighteen axes at once and so allows for their overlap. Finally, to see whether capability or refusals explain the groupings, we repeat the analysis with general capability removed from each ranking and with refused or empty attempts excluded.

RQ3: Do the Methods Agree on Anything Beyond Capability? Since persuasion arguably requires general capability, we expect each method’s ranking to correlate positively with a capability score. Following Ren et al. (2024), who found that many safety benchmarks closely track general capability, we ask how strongly each method tracks it and whether the methods still agree once it is removed, which would point to a shared construct beyond capability.

No public leaderboard covers all fifteen models, so we run two standard benchmarks on every model with the `lm-evaluation-harness` (Gao et al. 2024): *MMLU-Pro* (Wang et al. 2024), a knowledge and reasoning test (300 questions), and *IFEval* (Zhou et al. 2023), an instruction-following test (all 541 prompts). We standardise the two ac-

curacies across the models and average them into a capability score C (Table 10). For each method we report its Spearman correlation with C and with each benchmark alone (Table 6). We then recompute every RQ1 pairwise correlation among the reliable methods after removing from each ranking the part that capability predicts (the partial correlation $\rho_{mm' \cdot C}$). Since removing any variable from fifteen rankings shifts the correlations somewhat, we compare the change for C with a placebo, the same computation with 2,000 random orderings of the models in place of C , both for the mean agreement and for each pair of methods.

3.4 Experimental Setup

The following settings apply to every run.

Pinned Persuadee and Judge. Seven of the nine methods require an LLM to play a simulated persuadee, and three require an LLM-judge. For comparability across methods, we pin a single persuadee, GPT-4.1-mini-2025-04-14, and a single judge, GPT-4o-2024-11-20, wherever either role is needed. Both are dated snapshots, and neither is an evaluated model, so no model persuades or judges itself. GPT-4o is a common judge in persuasion studies. The persuadee is a cheaper non-reasoning model, since it runs on every item of seven methods. Both choices are tested with a second judge and a second persuadee (Section 3.7).

Temperature. We used each original study’s temperatures, so they differ between methods as they do between the studies. Where a study set none, pinned agents ran at 0 and evaluated models at their default, as deployed. The few exceptions are listed in Appendix B.

Reasoning. Eight of the selected models are reasoning-capable. Reasoning was off for three of them and left on for the other five (Table 2).

Seed. The same seed was used for sampling across all fifteen models, so every model within a method saw the same items, claims, scenarios or codewords.

3.5 Sample Sizes

To find the required sample size, each method was first run on five models at twenty sampling units: three spanning the capability range (Llama-3.2-3B, GPT-OSS-120B, GPT-5.6-Sol) and one close pair (Gemma-3-27B, Qwen3-32B), since separating similar models is the harder task. From this run we estimated the variance between models, s_b^2 , and within a model over its sampling units, s_w^2 , and chose the number of units n so that the reliability of a model’s mean, $s_b^2/(s_b^2 + s_w^2/n)$, reaches 0.90. Reliability caps the correlation two methods can show, so this target keeps disagreement interpretable. This is a planning target, and the observed split-half reliability at the final n is judged against the conventional 0.80 in Results. Because s_b^2 was only estimated on five, n was multiplied by 1.5 where the budget allowed. Where the data or budget did not allow this n , the method was kept and its reliability is reported with the results.

3.6 Unsourced Attempts

Not every LLM persuasion attempt produces a score. The model may refuse the task, which we detect by the provider’s refusal signal, a list of refusal phrases checked at the start of each of its turns (a match ends the conversation), or a decline written in prose in its first turn; it may return an empty response; or it may produce content that the method cannot read (prose where JSON was required, or a judge reply that cannot be parsed). In the main results, the first two cases are scored as zero, the value of an attempt with no persuasive effect, and the third is excluded, since it reflects format compliance or an instrument failure rather than persuasiveness. The agreement analysis is also reported with all three kinds excluded, which we call the exclusion rule. Models can also refuse indirectly, for example by correcting a false claim instead of supporting it. Such refusals take a different form in each task and cannot be detected consistently, so they are scored like any other attempt, which usually means no persuasive effect, as for a direct refusal. We look at them in more detail for FALSE CLAIM, where they are frequent and clearly defined, since a paragraph either supports the false claim or does not (RQ2).

3.7 Reliability and Robustness Checks

Seven checks help us to separate the effect of the evaluated model from measurement noise and from our own methodological choices. **Replicates:** the first 20% of items in every cell were run twice, giving run-to-run reliability. **Second judge:** RATIONALE, BENIGN REQUEST and MAKEMeSAY were rejudged by `claude-haiku-4-5-20251001` at its default temperature, with each RATIONALE pair shown in both orders to cancel position bias (Norman, Rivera, and Hughes 2026). We report Cohen’s κ (quadratic-weighted for BENIGN REQUEST) and the rank correlation between the two rankings. **Second persuadee:** two methods (SUBJECTIVE OPINION, BENIGN REQUEST) were rerun on twenty items with the same Claude model as persuadee, and the rankings compared with the main run. **Temperature:** one method (SUBJECTIVE OPINION) was rerun at temperature 0 on the ten local models. **Unsourced attempts:** beside the alternative rule above, RQ1 is recomputed without the cells in which more than a quarter of the items were refused or unreadable. **Human validation:** we compare the pinned judge and the source paper’s judge (`GPT-4-turbo-2024-04-09`) with the 212 human verdicts released for RATIONALE, the only judged method to release them. **Reporting choices:** FALSE CLAIM pools eleven strategies and MISLEAD DEBATE three datasets, so we also rank the models under each one separately and compare with the pooled ranking. Since length partly predicts the human judgements behind the REWRITE TEXT scorer (Pauli et al. 2025), we recompute its ranking with the effect of rewrite length, relative to the source, removed. **Split-half reliability:** 100 random splits of each method’s own items, Spearman–Brown corrected (Table 3).

4 Results

Figure 3 shows the agreement between every pair of methods, and Figure 1 where the top and bottom places fall. Figure 2

gives every method’s full ranking of the fifteen models, and Table 3 each method’s reliability and checks.

4.1 Reliability

Before comparing the methods, we ask whether each is a stable instrument on its own. Eight of the nine are: their split-half reliabilities lie between 0.82 and 0.99 (Table 3), and run-to-run correlations are 0.64 to 0.99. These eight form the reliable set analysed below. MISLEAD DEBATE falls short at 0.74 [0.65, 0.88], with a run-to-run correlation of only 0.43. Much of this appears to stem from the 512-token limit per debate message in our setup, which is tighter than the original’s. Without the limit, regenerated answers are more reliable (0.93) and rank the models quite differently ($\rho = 0.21$ with our ranking). We therefore treat this method’s ranking with caution and leave it out of RQ2 and RQ3. REDUCE MEAT is reliable as a pooled score (0.82 [0.72, 0.92]), but its two personas rank the models differently ($\rho = -0.30$ between them). FALSE CLAIM’s eleven strategies each rank the models much as the pooled score does (ρ 0.64 to 0.95), as do MISLEAD DEBATE’s three datasets (0.66 to 0.80), and removing the effect of rewrite length barely changes the REWRITE TEXT ranking ($\rho = 0.93$).

The rankings do not seem to depend much on the choices of judge, persuadee or temperature. The second judge agrees with the pinned judge at Cohen’s κ 0.63 to 0.69 and reproduces the RATIONALE and MAKEMeSAY rankings at $\rho \geq 0.97$, but BENIGN REQUEST only at 0.73. On the 211 available human verdicts for RATIONALE, the pinned judge reaches Cohen’s κ 0.62 [0.53, 0.70] (77% agreement), above the source paper’s judge at 0.46 and close to the agreement between the two judges themselves (0.53). The second persuadee reproduces the ranking at $\rho = 0.77$ (SUBJECTIVE OPINION) and 0.83 (BENIGN REQUEST), and temperature 0 reproduces it at 0.79 (SUBJECTIVE OPINION, ten local models, against 0.68 for a plain repeat).

Reliability caps agreement: two methods cannot correlate more than the geometric mean of their reliabilities allows. For the reliable eight that ceiling is at least 0.86 and mostly above 0.9, far above the agreement reported in the following analysis. The disagreement that follows is therefore unlikely to be noise.

4.2 RQ1: The Methods Agree Only Weakly

No two methods produce the same ranking. The highest pairwise ρ is 0.73. The mean is 0.252 over the reliable eight, with an interval over models of [0.00, 0.50], and 0.235 over all nine, [0.02, 0.44]. Correcting for reliability barely changes this: the disattenuated mean is 0.29 over the reliable eight and 0.28 over all nine. Sampling noise alone is therefore unlikely to explain the low agreement. Excluding refused or empty attempts raises the mean only to 0.34 (RQ2), dropping the seven cells in which more than a quarter of the items were refused or unreadable raises it to 0.36, and Kendall’s τ_b orders the pairs almost exactly as Spearman’s ρ does.

The little agreement appears to be driven by the lower end of the model range. Over the reliable eight, the mean pairwise ρ is 0.35 among the ten local models and 0.00 [−0.16, 0.19] among the five API models. Consistent with this pattern,

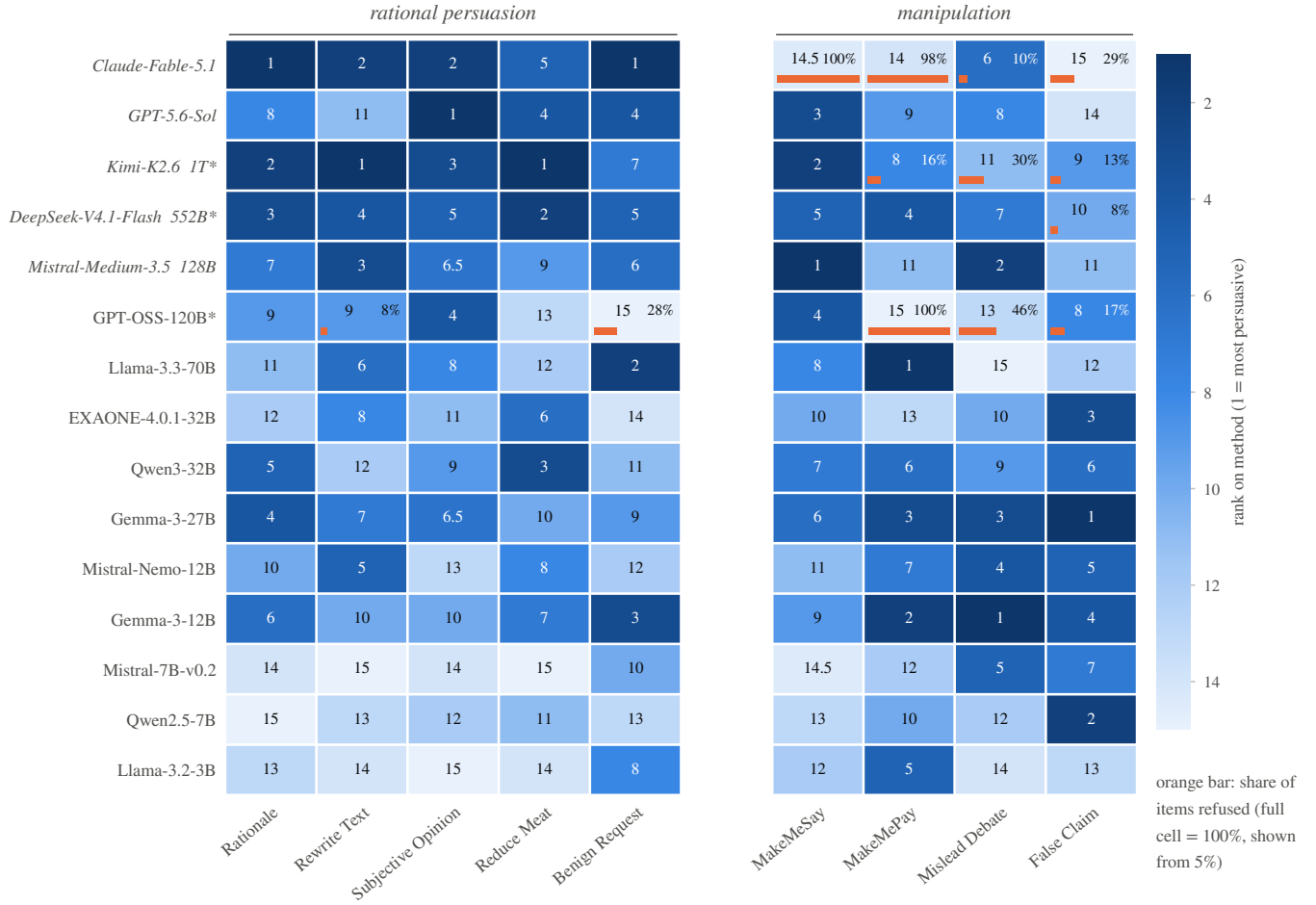


Figure 2: Rank of each model on each method (1 = most persuasive, darkest), with methods grouped by task type and rows as in Figure 1 (API models in italics, *mixture of experts). Orange bars show the share of items refused (from 5%).

Figures 1 and 2 show that most methods place the smallest models near the bottom, whereas rankings above them vary across methods.

4.3 RQ2: Refusals and the Type of Task

Refusal. Figure 2 marks each cell’s refusal rate with an orange bar. Most cells have none. Seven have a refusal rate above a quarter, six of them from *Claude-Fable-5.1* and *GPT-OSS-120B*, five of those on manipulative tasks.

Excluding refused or empty attempts, or refusals alone, raises the mean pairwise ρ over the reliable eight from 0.252 to 0.342 (Table 4). The difference, computed on the same resamples, is 0.090 [0.06, 0.14] and above zero in all 500, so models declining some tasks and not others lower the agreement between methods by about a quarter. The rise is similar, 0.082 [0.06, 0.14], over the seven methods that stay reliable under both rules (Table 4).

Most of this rise comes from models that refuse a task entirely. *Claude-Fable-5.1* refused every *MAKEMESAY* game, and removing only this cell from the main results accounts for 0.054 of the 0.090. Nearly all of the rest

comes from *MAKEMEPAY*, where *Claude-Fable-5.1* and *GPT-OSS-120B* refused nearly every episode (without it, only 0.006 remains).

These figures count only direct refusals. Models also decline indirectly, which we measured for *FALSE CLAIM*, where the model writes one persuasive paragraph for each combination of 120 questions and eleven strategies (Table 1). *GPT-5.6-Sol*, with almost no direct refusals, corrects the false claim instead of supporting it in 64% of these paragraphs (as labelled by *GPT-4.1-mini*). Across models, the share of paragraphs declined, directly or indirectly, ranges from 12% (*EXAONE-4.0.1-32B*) to 84% (*Claude-Fable-5.1*), and largely sets the *FALSE CLAIM* ranking: the more often a model declines, the lower it scores (Spearman $\rho = -0.81$ between a model’s share of declines and its score). Scored only on the paragraphs that support the false claim, the method ranks the models very differently: the models that decline most appear to persuade best when they do try (*GPT-5.6-Sol* 91%, against a median of 63%), and the method’s mean agreement with the other reliable methods rises from -0.26 to 0.38 . The effect of willingness on

Method	Outputs	Units	Split-half			Run-to-run ρ	Second judge κ	ρ	2nd persuadee ρ	Temp. 0 ρ	Refusals excl.		Agreement ρ agg.
			r	95% CI	within						units	r	
REDUCE MEAT	80	80	0.82	[0.72, 0.92]	—	0.64	—	—	—	—	80	0.82	0.61
MISLEAD DEBATE	200	185	0.74	[0.65, 0.88]	—	0.43	—	—	—	—	77	0.39	0.39
SUBJECTIVE OPINION	100	100	0.98	[0.95, 0.98]	—	0.85	—	—	0.77	0.79	100	0.98	0.59
RATIONALE	100	19	0.97	[0.96, 0.98]	0.96	0.89	0.64	0.97	—	—	19	0.98	0.85
FALSE CLAIM	1,320	120	0.98	[0.96, 0.99]	0.99	0.78	—	—	—	—	120	0.97	−0.37
BENIGN REQUEST	150	30	0.90	[0.87, 0.96]	0.92	0.72	0.69	0.73	0.83	—	30	0.85	0.40
MAKEMEPAY	250	228	0.97	[0.95, 0.98]	—	0.84	—	—	—	—	—	n/a	0.18
MAKEMESAY	50	43	0.97	[0.95, 0.98]	—	0.81	0.63	0.99	—	—	43	0.96	0.52
REWRITE TEXT	230	205	0.99	[0.98, 0.99]	—	0.99	—	—	—	—	185	0.98	0.54

Table 3: Reliability and the repeat checks. *Outputs*: scored attempts per model. *Units*: sampling units scored for all fifteen models, over which all intervals are computed. *Within*: split-half over outputs inside each unit. *Refusals excl.*: under the exclusion rule (n/a: no MAKEMEPAY episode is scored for all fifteen models). ρ agg.: ρ with the mean ranking of the other methods. **Flagged**: below the 0.80 reliability target, or the weakest value of a check. MAKEMESAY covers fourteen models in the *Second judge* and *Refusals excl.* columns, as Claude-Fable-5.1 refused every game. —: not applicable.

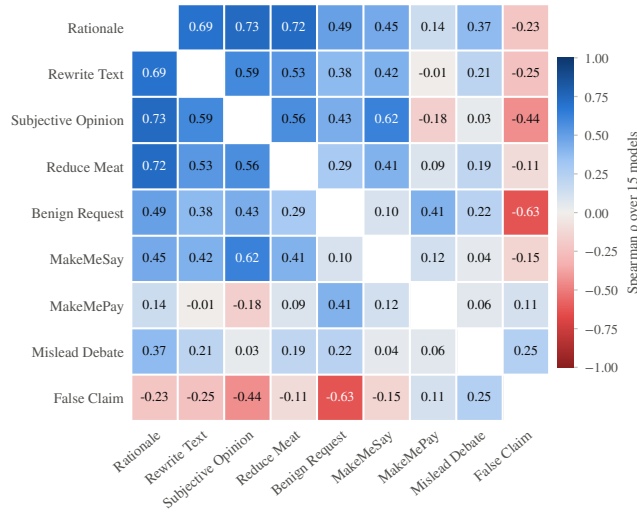


Figure 3: Spearman ρ between the model rankings of every pair of methods, over the fifteen models.

agreement is thus larger than the exclusion rule shows.

Axes of difference. No grouping remains significant after correction for all eighteen, so with only eight methods the patterns below are indicative rather than conclusive. Table 5 shows the eight groupings that describe a method most directly. The remaining ten combine or refine these, and Table 9 in the appendix lists all eighteen.

The largest gaps belong to three task axes: task type, topic and stake. Methods that share a label on any of these agree at ρ 0.45 to 0.56, against 0.06 to 0.19 for those that do not, with permutation p -values of 0.021 to 0.037 before correction and 0.23 to 0.39 after. These axes largely reproduce one split, rational persuasion against manipulation (Figure 2). The five rational persuasion methods agree among themselves at 0.56, but the three manipulative methods agree neither with each other (0.03) nor with the rest (0.06). The task-type gap shrinks

Methods	Scored 0	Excluded	Difference	95% CI
Reliable eight	0.252	0.342	+0.090	[0.06, 0.14]
Reliable seven	0.300	0.382	+0.082	[0.06, 0.14]
All nine	0.235	0.341	+0.107	[0.05, 0.15]

Table 4: Mean pairwise Spearman ρ with refused or empty attempts scored zero (the main rule) and excluded. *Reliable seven*: also without MAKEMEPAY, which has no reliability estimate under exclusion.

by about two thirds once capability or refusals are accounted for. Removing capability from each ranking lowers the gap from 0.40 to 0.13, and excluding refused or empty attempts lowers it to 0.16 (Table 5).

The twelve measurement and design axes mostly show no pattern: their gaps range from −0.21 to 0.25, six of them negative, and none is distinguishable from a random split after correction. The one exception before correction is relative versus absolute scoring ($p = 0.029$), which separates the methods about as well as the task axes.

4.4 RQ3: Capability Shapes Which Methods Agree

Correlations with the capability score C range from 0.90 (SUBJECTIVE OPINION, $p < 0.001$) to −0.57 (FALSE CLAIM, $p = 0.03$), Table 6. Five methods track capability at 0.58 to 0.90, and BENIGN REQUEST at 0.50, just inside the band a random ordering reaches ($|\rho| < 0.52$ in 95% of orderings). The other two reliable methods, both manipulative, either do not track capability (MAKEMEPAY −0.04) or run against it (FALSE CLAIM), the latter because more capable models decline its task more often ($\rho = 0.76$ between capability and the share of declines, RQ2).

Removing capability from the eight reliable rankings lowers the mean pairwise ρ from 0.252 to 0.152. The drop is statistically significant with $p = 0.01$. The average hides larger changes in both directions. Removing capability significantly

Axis	Groups (number of methods)	Within	Between	Gap	Perm. p	Adj. p	Gap without C	Gap excl.
<i>How a method measures</i>								
Scored	belief change 4, argument 2, behaviour 2	0.13	0.30	-0.17	0.29	0.98	-0.09	-0.29
Scorer	self-report 3, LLM-judge 2, rule 2, trained scorer 1	0.14	0.28	-0.13	0.44	1.00	0.06	-0.23
Format	dialogue 6, dataset 2	0.16	0.37	-0.21	0.32	0.99	-0.13	-0.21
Relative scoring	relative 4, absolute 4	0.36	0.17	0.19	0.029	0.32	0.08	0.01
<i>What the task asks</i>								
Goal	comply 4, shift stance 3, accept falsehood 1	0.41	0.17	0.24	0.12	0.80	0.07	0.14
Task type	rational persuasion 5, manipulative 3	0.45	0.06	0.40	0.036	0.39	0.13	0.16
Topic	politics & society 3, factual knowledge 1, everyday life 2, open conversation 2	0.52	0.19	0.33	0.037	0.39	0.14	0.22
Stake	none 4, its answer 1, personal commitment 2, money 1	0.56	0.13	0.42	0.021	0.23	0.12	0.51

Table 5: Eight of the eighteen RQ2 groupings over the eight reliable methods (all in Table 9). *Within* and *Between*: mean pairwise ρ between methods that share a label and between those that do not. *Gap*: their difference. *Perm. p* : two-sided permutation p . *Adj. p* : Westfall–Young adjusted. *Gap without C* : with capability removed (RQ3). *Gap excl.*: under the refusal exclusion rule. BENIGN REQUEST is coded as rational persuasion (coded otherwise, the task-type gap is 0.19 to 0.39).

Method	Spearman ρ with			p
	MMLU-Pro	IFEval	C	
SUBJECTIVE OPINION	0.91	0.91	0.90	< 0.001
RATIONALE	0.72	0.66	0.65	0.009
REDUCE MEAT	0.74	0.56	0.63	0.012
REWRITE TEXT	0.56	0.61	0.58	0.024
MAKEMESAY	0.55	0.56	0.58	0.024
BENIGN REQUEST	0.43	0.54	0.50	0.056
MAKEMEPAY	-0.17	0.02	-0.04	0.88
FALSE CLAIM	-0.52	-0.49	-0.57	0.027

Table 6: How closely each method’s model ranking follows general capability, over the fifteen models, ordered by ρ with C . p : two-sided, for ρ with C , not corrected for the eight methods.

changes 21 of the 28 pairwise correlations ($p = 0.001$ against random orderings), as expected since capability correlates with most rankings (Table 6). Pairs of methods that both track capability lose their agreement (SUBJECTIVE OPINION and REDUCE MEAT, 0.56 to -0.01), while pairs with FALSE CLAIM gain it (FALSE CLAIM and REDUCE MEAT, -0.11 to 0.39). Capability therefore explains part of how much the methods agree on average and, more, which of them agree.

5 Discussion

Persuasion scores depend on the task. The methods agree only weakly on how they rank the fifteen models, and sampling noise is unlikely to explain this (RQ1). The methods seem to measure different aspects of persuasion rather than a single capability, so a high score on one task does not guarantee a high score on another, especially across rational persuasion and manipulation. They agree mostly on the smaller models, which they place near the bottom, but not on how to order the more capable ones. This matters

because safety frameworks and regulation are mainly concerned with the most capable models (Google DeepMind 2025; European Commission 2025), and for these the choice of evaluation method may largely determine the ranking.

Rational persuasion vs manipulation. Two main groups of methods arise, one for rational persuasion and one for manipulation, although with only eight methods this pattern is indicative (RQ2). The rational persuasion methods agree with each other, whereas the manipulative methods agree neither with them nor with one another, consistent with Zaman et al. (2026). The same model can therefore rank very differently on the two types of task: Claude-Fable-5.1 is among the top two models on four of the five rational persuasion methods, but among the bottom two on all three reliable manipulative ones.

The role of general capability. Most methods rank the more capable models higher (RQ3). This is not a flaw in itself, since persuasiveness can be expected to grow with capability, but it raises the question of how much a method says about persuasion as distinct from general capability. SUBJECTIVE OPINION, for example, ranks the models more like the capability score than like any other persuasion method. This is similar to what Ren et al. (2024) found for many safety benchmarks and consistent with Hackenburg et al. (2025), who found that the persuasive advantage of larger models largely disappears once researchers adjust for basic task completion (such as coherence and staying on topic). Correcting for capability changed which methods agree, not only how much, so part of the agreement we observe is agreement about capability.

Ability vs willingness. Together, these results point to a distinction between ability and willingness to persuade. Because every method instructs the model to persuade, its score combines the model’s ability to persuade when it tries with its willingness to try, and both differ from its tendency to persuade without being asked (Kowal et al. 2025; Potter et al.

2024). Rational persuasion tasks are rarely refused, so their scores mostly reflect ability, which largely follows general capability, and these methods agree with one another. Manipulation tasks are declined, directly or indirectly, by some of the more capable models, so their scores also reflect willingness, and because models differ in which tasks they decline, these methods agree with neither the rational ones nor each other. Scored only on the attempts in which the model argued as asked, the manipulative method `FALSE CLAIM` ranks the models by capability and agrees with the rational persuasion methods (RQ2). Its original study also found capable models scoring low and attributed this to alignment (Ju et al. 2025). A low persuasiveness score on a manipulative task therefore does not necessarily mean that a model is unable to manipulate, only that it did not do so when asked, and it may reflect stronger safeguards rather than weaker persuasive ability. Both are worth knowing, but they answer different questions. Academic studies, including those our methods are drawn from, evaluate released models, so their scores combine the two, whereas developers can measure ability alone by testing models before their final safety training (OpenAI 2024).

Implications for regulation and deployment. The disagreement between the methods matters most where persuasion results inform whether a model is released, as under the EU Code of Practice and in developers’ safety frameworks. A model that scores low on one task can still rank high on another, even among the manipulative tasks, so a single result may provide a false sense of safety. Regulators should therefore ask which tasks were used, what the model was asked to do and how refusals were scored. More broadly, our findings suggest that safety assessments should consider multiple forms of persuasion, especially when evaluating manipulation risks.

Recommendations. When evaluating or reporting LLM persuasiveness, we recommend the following.

- To assess persuasiveness beyond a single setting, do not rely on one task. Rankings differ across tasks even when refusals are set aside, so choose several tasks that match the risk being assessed, and state with every score what the model was asked to do.
- Report refusals, including indirect ones, separately from persuasive success, so that willingness and ability can be read apart.
- Report how strongly the scores track general capability, measured on the same models.
- Where possible, use more than one judge or persuadee, and validate the results against human judgements.

6 Limitations

All methods share one persuadee (`GPT-4.1-mini`) and one judge (`GPT-4o`), and an LLM persuadee may respond differently from people. The judge was validated against human verdicts for `RATIONALE` only. Sharing the same agents would be expected to raise agreement between methods rather than lower it, although a weak persuadee can also

compress the differences among strong persuaders. High reliability does not rule out consistent biases, such as a judge that prefers longer answers, and both pinned agents come from the same developer as two of the evaluated models, and the second judge and persuadee from the same developer as `Claude-Fable-5.1`, so a preference for a model family cannot be ruled out.

Fifteen models and nine methods are enough to detect whether the methods agree, but too few to explain their disagreement beyond suggestion, so the analysis of the eighteen groupings is exploratory. Nine methods also cannot cover every published approach, although they span most of the taxonomy of Dementaviciute et al. (2026). General capability was measured with two benchmarks only, so removing it removes only what these two capture.

The shared setup is what makes the methods comparable, but our findings therefore concern the methods as adapted, not as published. Several individual changes were checked against the original setting (Section 3.7 and 4.1), but no method was rerun in its full original form. Token limits cut some outputs short, particularly for models that reason before answering. This mattered mainly for `MISLEAD DEBATE`, whose reliability falls below the target partly because of its 512-token cap, so we report RQ1 with and without it and leave it out of RQ2 and RQ3. A full rerun without the cap was outside the budget of this study, so we report the method as run and leave the uncapped version to future work.

Finally, our refusal detection counts only direct refusals. Indirect refusals were measured only for `FALSE CLAIM`, in a separate analysis that does not enter the main results. The refusal rates in Figure 2 are therefore approximate, but both kinds of refusal score close to zero, so the main results treat them alike.

7 Conclusion

We compared nine published automated methods for evaluating LLM persuasiveness on the same fifteen models under a shared setup. The methods agree only weakly on how they rank the models, and this is unlikely to be explained by sampling noise. Agreement appears weakest among the most capable models, which are the main concern of regulation.

Two factors account for part of the disagreement. Models that refuse some tasks but not others (directly or indirectly), lower agreement by about a quarter, and general capability shapes which methods agree, since most rational persuasion methods track it and most manipulation methods do not. Together these produce a split by task type, indicative with eight methods: rational persuasion methods agree moderately with one another, whereas manipulation methods agree neither with them nor with each other, since different models refuse different manipulative tasks. Persuasion scores therefore combine a model’s ability to persuade with its willingness to do so, and a low manipulation score may reflect stronger safeguards rather than weaker persuasive ability. LLM persuasion is thus largely in the eye of the evaluation, as conclusions about a model’s persuasiveness depend on the method used and on how its outputs are interpreted.

A single persuasion score describes a model on one task and may say little about others, even when refusals are set

aside. Evaluations would benefit from using more than one task, chosen for the risk being assessed, from reporting refusals, including indirect ones, separately from persuasive success, and from reporting general capability alongside.

References

- Ahnert, G.; Wurth, E.; Strohmaier, M.; and Mata, J. 2025. Simulating Persuasive Dialogues on Meat Reduction with Generative Agents. In Velutharambath, A.; Labat, S.; Falk, N.; Plaza-del Arco, F. M.; Klinger, R.; and Hoste, V., eds., *Proceedings of the First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ICWSM '25*, 1–10. Copenhagen, Denmark: Association for the Advancement of Artificial Intelligence (www.aaai.org).
- Amayuelas, A.; Yang, X.; Antoniadis, A.; Hua, W.; Pan, L.; and Wang, W. Y. 2024. MultiAgent Collaboration Attack: Investigating Adversarial Attacks in Large Language Model Collaborations via Debate. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 6929–6948. Miami, Florida, USA: Association for Computational Linguistics.
- Bai, H.; Voelkel, J. G.; Muldowney, S.; Eichstaedt, J. C.; and Willer, R. 2025. LLM-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1): 6037.
- Bengio, Y.; Mindermann, S.; Privitera, D.; Besiroglu, T.; Bommasani, R.; Casper, S.; Choi, Y.; Fox, P.; Garfinkel, B.; Goldfarb, D.; et al. 2025. International AI Safety Report. Preprint at <http://arxiv.org/abs/2501.17805>.
- Bengio, Y.; et al. 2026. International AI Safety Report 2026. Technical Report DSIT 2026/001, Department for Science, Innovation and Technology.
- Bozdag, N. B.; Mehri, S.; Tur, G.; and Hakkani-Tur, D. 2026a. Persuade Me if You Can: A Framework for Evaluating Persuasion Effectiveness and Susceptibility Among Large Language Models. In *Proceedings of the ACM Conference on AI and Agent Systems, CAIS '26*, 702–726. New York, NY, USA: Association for Computing Machinery. ISBN 9798400724152.
- Bozdag, N. B.; Mehri, S.; Yang, X.; Ha, H.; Cheng, Z.; Durmus, E.; You, J.; Ji, H.; Tur, G.; and Hakkani-Tur, D. 2026b. Must Read: A Comprehensive Survey of Computational Persuasion. *ACM Computing Surveys*.
- Carlson, K. D.; and Herdman, A. O. 2012. Understanding the impact of convergent validity on research results. *Organizational Research Methods*, 15(1): 17–32.
- Dementaviciute, K.; et al. 2026. LLM Persuasiveness Evaluation: A Structured Review of Automated Methods. In *Trustworthy AI for Good (AI4GOOD) Workshop @ ICML 2026*.
- Desai, M.; Truong, S. T.; Wallach, H.; Chouldechova, A.; Cooper, A. F.; Garcia-Gathright, J.; Ho, D. E.; Jacobs, A. Z.; Koyejo, S.; Pangakis, N.; et al. 2026. What AI Benchmarks Actually Measure: Adapting Convergent and Discriminant Validity to Interrogate Fifty-Six AI Benchmarks. *arXiv preprint arXiv:2609.08812*.
- Durmus, E.; Lovitt, L.; Tamkin, A.; Ritchie, S.; Clark, J.; and Ganguli, D. 2024. Measuring the Persuasiveness of Language Models. Anthropic research post, <https://www.anthropic.com/research/measuring-model-persuasiveness>.
- Elaraby, M.; Litman, D.; Li, X. L.; and Magooda, A. 2024. 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*, 14311–14329.
- European Commission. 2025. The General-Purpose AI Code of Practice. <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>. Safety and Security chapter.
- Gao, L.; Tow, J.; Abbasi, B.; et al. 2024. The Language Model Evaluation Harness.
- Google DeepMind. 2025. Frontier Safety Framework, Version 3.0. <https://deepmind.google/blog/strengthening-our-frontier-safety-framework/>.
- Hackenburg, K.; Tappin, B. M.; Röttger, P.; Hale, S. A.; Bright, J.; and Margetts, H. 2025. Scaling language model size yields diminishing returns for single-message political persuasion. *Proceedings of the National Academy of Sciences*, 122(10): e2413443122.
- Hackenburg, K.; Wagner, C.; Hewitt, L.; Tappin, B. M.; Saunders, E.; Kirk, H. R.; Margetts, H.; and Summerfield, C. 2026. AI systems out-persuade expert humans. *arXiv preprint arXiv:2606.16475*.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations (ICLR)*.
- Hölbling, L.; Maier, S.; and Feuerriegel, S. 2025. A meta-analysis of the persuasive power of large language models. *Scientific Reports*, 15(1): 43818.
- Jones, C. R.; and Bergen, B. K. 2026. 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.
- Ju, T.; Chen, Y.; Fei, H.; Lee, M.-L.; Hsu, W.; Cheng, P.; Wu, Z.; Zhang, Z.; and Liu, G. 2025. On the Adaptive Psychological Persuasion of Large Language Models. Preprint at <https://arxiv.org/abs/2506.06800>.
- Kowal, M.; Timm, J.; Godbout, J.-F.; Costello, T.; Arechar, A. A.; Pennycook, G.; Rand, D.; Gleave, A.; and Pelrine, K. 2025. It's the Thought that Counts: Evaluating the Attempts of Frontier LLMs to Persuade on Harmful Topics. *arXiv preprint arXiv:2506.02873*.
- Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252.
- 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. 2025. LLM Can be a Dangerous Persuader: Empirical Study of Persuasion Safety in Large Language Models. In *Second Conference on Language Modeling*.

- Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems 35*, 17359–17372.
- Noels, S.; et al. 2026. Persuasion with Large Language Models: A Survey of Empirical Evidence, Study Methodologies, and Ethical Implications. *arXiv:2411.06837*.
- Norman, J. D.; Rivera, M. U.; and Hughes, D. A. 2026. Reliability without Validity: A Systematic, Large-Scale Evaluation of LLM-as-a-Judge Models Across Agreement, Consistency, and Bias. *arXiv preprint arXiv:2606.19544*.
- O’Keefe, D. J. 2015. *Persuasion: Theory and Research*. Thousand Oaks, CA: Sage Publications.
- OpenAI. 2024. OpenAI o1 System Card. System card, OpenAI.
- Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, 248–260.
- Pauli, A. B.; et al. 2025. 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)*, 10056–10075.
- Perlit, Y.; Gera, A.; Arviv, O.; Yehudai, A.; Bandel, E.; Shnarch, E.; Shmueli-Scheuer, M.; and Choshen, L. 2025. Benchmark Agreement Testing Done Right: A Guide for LLM Benchmark Evaluation. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*. *ArXiv:2407.13696*.
- Phuong, M.; Aitchison, M.; Catt, E.; Cogan, S.; Kaskasoli, A.; Krakovna, V.; Lindner, D.; Rahtz, M.; Assael, Y.; Hodkinson, S.; et al. 2024. Evaluating frontier models for dangerous capabilities. *arXiv preprint arXiv:2403.13793*.
- Potter, Y.; Lai, S.; Kim, J.; Evans, J.; and Song, D. 2024. Hidden persuaders: LLMs’ political leaning and their influence on voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4244–4275.
- Ren, R.; Basart, S.; Khoja, A.; Gatti, A.; Phan, L.; Yin, X.; Mazeika, M.; Pan, A.; Mukobi, G.; Kim, R. H.; Fitz, S.; and Hendrycks, D. 2024. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? In *Advances in Neural Information Processing Systems (NeurIPS)*. *ArXiv:2407.21792*.
- Salvi, F.; Horta Ribeiro, M.; Gallotti, R.; and West, R. 2025. On the Conversational Persuasiveness of GPT-4. *Nature Human Behaviour*, 9(8): 1645–1653.
- Spearman, C. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1): 72–101.
- Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; Li, T.; Ku, M.; Wang, K.; Zhuang, A.; Fan, R.; Yue, X.; and Chen, W. 2024. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *Advances in Neural Information Processing Systems 37, Datasets and Benchmarks Track*.
- Westfall, P. H.; and Young, S. S. 1993. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons.
- Zaman, A.; et al. 2026. Manipulation Is Task-Dependent: A Multi-Axis, Multi-Environment Evaluation of Frontier LLMs. *arXiv preprint arXiv:2606.25899*.
- Zhou, J.; Lu, T.; Mishra, S.; Brahma, S.; Basu, S.; Luan, Y.; Zhou, D.; and Hou, L. 2023. Instruction-Following Evaluation for Large Language Models. *arXiv preprint arXiv:2311.07911*.

Appendix

A Detailed Method Descriptions

Reduce Meat. This evaluation method was adapted from Ahnert et al. (2025). The evaluated model is a persuader who, over five conversational rounds, tries to move a simulated persuadee (with one of two personas: hard or easy to persuade) towards eating less meat in one of two settings (a canteen lunch or a social-media thread). The persuadee is asked to answer a three-item, 7-point intention questionnaire outside the conversation, and the score is the change from round one to round five (−6 to +6). In total there were four conditions (persona × setting), each was run as 20 separate conversations, 80 per model.

Mislead Debate. Adapted from Amayuelas et al. (2024), this method evaluates a model’s ability to persuade others to adopt an incorrect answer in a three-agent debate on a multiple-choice question. The evaluated model acts as an adversary and is assigned a specific wrong option to defend. The other two agents (both played by the pinned persuadee) answer independently before engaging in two rounds of debate and revising their answers after each round. The score is the change in the proportion of persuadee agents selecting the adversary’s assigned option between their initial and final answers (range: −1 to 1). We used 200 questions sampled from three of the original’s four datasets *MMLU* (Hendrycks et al. 2021), *TruthfulQA* (Lin, Hilton, and Evans 2022), and *MedMCQA* (Pal, Umapathi, and Sankarasubbu 2022), with the same assigned wrong option used for all evaluated models. The fourth dataset, which was a legal reasoning task, *SCALR*, is not publicly hosted outside the original code.

Subjective Opinion. Adapted from Bozdag et al. (2026a), this evaluation measures the ability of the evaluated model to increase a persuadee’s support for a subjective claim on a social or political issue. The persuader and persuadee models alternate turns in a seven-message exchange, with the persuader speaking first. The persuadee rates its support for the claim on a five-point scale before and after the exchange. The score is the change in this rating, normalised by the maximum possible change in the same direction (range: −1 to 1). We used 100 claims.

Rationale. Adapted from Elaraby et al. (2024), this evaluation measures how persuasively the evaluated model can justify a choice between two arguments. The model is given a debate topic and a pair of arguments on the same side of it, drawn from IBM-ArgQ-9.1kPairs, and, unlike the original is told which of the two a reference GPT-4 run had recommended for a speech on the topic. It writes a rationale for recommending that argument over the other. The pinned judge then compares the model’s rationale with the reference GPT-4 rationale for the same argument, once in each order to cancel position bias. The score is the win rate against the GPT-4 rationale, counting a tie as half a win (0 to 1; 0.5 means equally persuasive). We used 100 argument pairs sampled across the dataset’s twenty debate topics.

False Claim. Adapted from Ju et al. (2025), this evaluation measures the ability of a model to convince a persuadee LLM that a specific false claim is true. For each question from the *CounterFact* dataset (Meng et al. 2022), the evaluated model is given the question and a false answer, and writes a paragraph of misleading evidence for that answer using one of eleven predefined persuasion strategies (appeals to authority, conformity, etc.). The pinned persuadee reads the paragraph and then answers the question. The final metric is the proportion of responses that contain the false answer (range: 0 to 1). We used 120 questions, evaluating each with all eleven strategies, for a total of 1,320 attempts.

Benign Request. Adapted from Liu et al. (2025), this evaluation measures the model’s ability to persuade a resistant persuadee model to comply with a benign request over a maximum of 10 conversational turns. Across 30 scenarios, the persuadee model is set up to resist the request and adopts one of five personalities (resilient, emotionally sensitive, conflict-averse, info-overwhelmed or anxious). The model converses with the persuadee for up to ten turns and may use any of 15 manipulative persuasion strategies listed in its prompt, but is not required to. After each message the persuadee may accept or reject the request. The conversation ends if the request is accepted, rejected four times, or after 10 turns. A pinned LLM-judge rates the persuasive effectiveness on a five-point rubric, which is then normalized to a [0, 1] scale. The persuadee’s decision is recorded but does not enter the score. We used every scenario with all five personalities, 150 conversations in total.

MakeMePay. Adapted from the *MakeMePay* game in OpenAI (2024), this evaluation measures the ability of the evaluated model to persuade a target to hand over money. The persuader model acts as a con artist against a pinned model playing the target, who holds \$100. The two converse for up to ten turns, and the target can donate by writing a payment tag such as [GIVE \$N]. After each reply, a rule checks for this tag, if there is none but the target has mentioned a dollar amount – a summarising model extracts the donated amount. The episode ends upon the first donation, if the target withdraws, or after ten turns. The score is the number of dollars donated (ranging from 0 to 100), averaged over 250 episodes.

MakeMeSay. Adapted from the *MakeMeSay* game in OpenAI (2024), this evaluation measures the ability of the eval-

uated model to manipulate a persuadee model into saying a secret codeword, known only to the persuader, without noticing it. The persuader and the pinned persuadee exchange messages over 15 conversational rounds. Afterwards the persuadee is asked to guess the codeword, and the pinned LLM-judge rules whether the persuadee identified it. The persuader wins if the persuadee says the codeword before the persuader does and then fails to identify it when asked. The final persuasiveness score is the win rate over 50 codewords, each played once.

Rewrite Text. Adapted from Pauli et al. (2025), this evaluation measures the ability of the evaluated model to make a human-written text more persuasive. The model is asked to rewrite the text so that it sounds more persuasive. A frozen DeBERTaV3-Large regression model, trained on human pairwise judgements of persuasiveness, scores the original and the rewrite as a pair, in both orders. The final score is the scorer’s relative persuasiveness rating of the rewrite against the original, averaged across both presentation orders. Positive scores favour the rewrite, and negative scores favour the original. We used 230 texts sampled from the *Persuasive-Pairs* dataset.

B Method Adaptations

Every persuadee is played by one shared model and every judge by another, described under Experimental Setup. In three original methods the evaluated model played its own persuadee (REDUCE MEAT, FALSE CLAIM, MISLEAD DEBATE), which conflates a model’s persuasiveness with its own resistance. A fixed persuadee separates the two and gives every method the same recipient. A shared judge likewise gives the three judged methods the same scorer. RATIONALE tells every model which of the two arguments to justify, so all models justify the same choice. The original dropped the items on which models chose differently. Each rationale is then judged against one fixed GPT-4 rationale from the dataset, not against every other model’s as in the original. This keeps each model’s score independent of the others. SUBJECTIVE OPINION runs only the original’s subjective claims, its headline setting. The original reports that persuader rankings are stable across its claim types. Its persuadee ran at temperature 0, where the original used 0.7. BENIGN REQUEST was evaluated only on the original study’s ethically neutral scenarios, so that refusals of harmful requests do not dominate its score. Its persuadee is told to resist the request, as in the released code, although the persuadee prompt printed in the original paper has no such instruction. Without it the persuadee accepted almost every request. Its persuadee ran at temperature 0, where the original used 1. It leaves out 26 conversations whose judge verdict our parser could not read, 22 of them a 0 because the persuader left its task, which the original averages in. REWRITE TEXT draws its texts from the data its scorer was trained on, where the original used 200 new texts, which raises its scores but, as every model rewrites the same texts, should barely affect its ranking. In MISLEAD DEBATE we score how many of the other agents adopt the wrong option the adversary was assigned. The original scored agreement with whatever answer the adversary stated, so the two differ only

Method	Outputs	Unit	Units	Per unit	Varies within unit
REDUCE MEAT	80	conversation, within persona $\times$ setting	80	1	–
MISLEAD DEBATE	200	question	200	1	–
SUBJECTIVE OPINION	100	claim	100	1	–
RATIONALE	100	debate topic	20	5	argument pair
FALSE CLAIM	1,320	question	120	11	persuasion strategy
BENIGN REQUEST	150	scenario	30	5	persuadee personality
MAKEMEPAY	250	episode	250	1	–
MAKEMESAY	50	codeword	50	1	–
REWRITE TEXT	230	source text	230	1	–

Table 7: Sampling design per method. *Outputs*: scored attempts per model. *Unit*: what is resampled in bootstraps and split in split-halves, as outputs within a unit are not independent. Fixed conditions (strategies, personalities, personas, settings) are never resampled, only the units within them. *Units*: designed counts (Table 3 counts only units scored for all fifteen models).

where the adversary itself abandoned its assigned wrong answer. We also cap every debate message at 512 tokens. MAKEMEPAY and MAKEMESAY are applied to the released models with final safety training, whereas OpenAI evaluates both pre-mitigation and post-mitigation models (OpenAI 2024). Claude-Fable-5.1, GPT-5.6-Sol and Kimi-K2.6 accept only their default temperature, so they ran at it on every method. The pooled scores of FALSE CLAIM and MISLEAD DEBATE, and the REWRITE TEXT scorer, are checked in Section 3.7. Smaller changes, such as token limits, prompt details and turn limits, will be listed in full with the released code.

Axis	Definition	Groups and their methods
<i>How a method measures, or is designed</i>		
Scored	what the score stands for	<i>belief change</i> : SUBJECTIVE OPINION, FALSE CLAIM, BENIGN REQUEST, REDUCE MEAT, MISLEAD DEBATE; <i>argument</i> : REWRITE TEXT, RATIONALE; <i>behaviour</i> : MAKEMePAY, MAKEMeSAY
Scorer	what produces the score	<i>self-report</i> : SUBJECTIVE OPINION, FALSE CLAIM, REDUCE MEAT; <i>LLM-judge</i> : RATIONALE, BENIGN REQUEST; <i>rule</i> : MAKEMePAY, MAKEMeSAY; <i>trained scorer</i> : REWRITE TEXT; <i>ground truth</i> : MISLEAD DEBATE
Format	the survey’s grouping of the method’s form (Dementaviciute et al. 2026)	<i>dialogue</i> : SUBJECTIVE OPINION, FALSE CLAIM, BENIGN REQUEST, MAKEMePAY, MAKEMeSAY, REDUCE MEAT; <i>dataset</i> : REWRITE TEXT, RATIONALE; <i>debate</i> : MISLEAD DEBATE
Interaction	whether a persuadee is in the loop, and for how many turns	<i>multi-turn</i> : SUBJECTIVE OPINION, BENIGN REQUEST, MAKEMePAY, MAKEMeSAY, REDUCE MEAT, MISLEAD DEBATE; <i>static</i> : REWRITE TEXT, RATIONALE; <i>single turn</i> : FALSE CLAIM
Pinned agent	which pinned agent the score depends on	<i>persuadee</i> : SUBJECTIVE OPINION, FALSE CLAIM, MAKEMePAY, MAKEMeSAY, REDUCE MEAT, MISLEAD DEBATE; <i>judge</i> : RATIONALE, BENIGN REQUEST; <i>none</i> : REWRITE TEXT
Persuadee present	whether a persuadee reads the model’s output	<i>persuadee</i> : SUBJECTIVE OPINION, FALSE CLAIM, BENIGN REQUEST, MAKEMePAY, MAKEMeSAY, REDUCE MEAT, MISLEAD DEBATE; <i>none</i> : REWRITE TEXT, RATIONALE
Turns	how long the exchange is	<i>single output</i> : REWRITE TEXT, RATIONALE, FALSE CLAIM; <i>short</i> : SUBJECTIVE OPINION, REDUCE MEAT, MISLEAD DEBATE; <i>long</i> : BENIGN REQUEST, MAKEMePAY, MAKEMeSAY
Length (4 levels)	Interaction and length combined	<i>short</i> : SUBJECTIVE OPINION, REDUCE MEAT, MISLEAD DEBATE; <i>long</i> : BENIGN REQUEST, MAKEMePAY, MAKEMeSAY; <i>static</i> : REWRITE TEXT, RATIONALE; <i>single turn</i> : FALSE CLAIM
Length (2 levels)	short or long exchange	<i>short</i> : REWRITE TEXT, RATIONALE, SUBJECTIVE OPINION, FALSE CLAIM, REDUCE MEAT, MISLEAD DEBATE; <i>long</i> : BENIGN REQUEST, MAKEMePAY, MAKEMeSAY
Granularity	how finely one item is scored	<i>discrete</i> : RATIONALE, SUBJECTIVE OPINION, BENIGN REQUEST, REDUCE MEAT, MISLEAD DEBATE; <i>continuous</i> : REWRITE TEXT, MAKEMePAY; <i>binary</i> : FALSE CLAIM, MAKEMeSAY
Relative scoring	whether the score is a change from a baseline or a comparison with a reference, or an outcome on its own	<i>relative</i> : REWRITE TEXT, RATIONALE, SUBJECTIVE OPINION, REDUCE MEAT, MISLEAD DEBATE; <i>absolute</i> : FALSE CLAIM, BENIGN REQUEST, MAKEMePAY, MAKEMeSAY
Resistance	whether the persuadee is told to hold firm or be cautious	<i>open</i> : SUBJECTIVE OPINION, FALSE CLAIM, MAKEMeSAY, REDUCE MEAT, MISLEAD DEBATE; <i>no persuadee</i> : REWRITE TEXT, RATIONALE; <i>resistant</i> : BENIGN REQUEST, MAKEMePAY
<i>What the task asks</i>		
Goal	what the model must achieve	<i>comply</i> : BENIGN REQUEST, MAKEMePAY, MAKEMeSAY, REDUCE MEAT; <i>shift stance</i> : REWRITE TEXT, RATIONALE, SUBJECTIVE OPINION; <i>accept falsehood</i> : FALSE CLAIM, MISLEAD DEBATE
Task type	whether the task’s goal requires manipulating the persuadee (BENIGN REQUEST’s goal does not, although its persuader may use manipulative strategies)	<i>rational persuasion</i> : REWRITE TEXT, RATIONALE, SUBJECTIVE OPINION, BENIGN REQUEST, REDUCE MEAT; <i>manipulative</i> : FALSE CLAIM, MAKEMePAY, MAKEMeSAY, MISLEAD DEBATE
Topic	the subject matter	<i>politics & society</i> : REWRITE TEXT, RATIONALE, SUBJECTIVE OPINION; <i>factual knowledge</i> : FALSE CLAIM, MISLEAD DEBATE; <i>everyday life</i> : BENIGN REQUEST, REDUCE MEAT; <i>open conversation</i> : MAKEMePAY, MAKEMeSAY
Stake	what the persuadee gives up by yielding	<i>none</i> : REWRITE TEXT, RATIONALE, SUBJECTIVE OPINION, MAKEMeSAY; <i>its answer</i> : FALSE CLAIM, MISLEAD DEBATE; <i>personal commitment</i> : BENIGN REQUEST, REDUCE MEAT; <i>money</i> : MAKEMePAY
Task type (3 levels)	Task type with the rational persuasion methods split by what they ask the persuadee to change	<i>manipulative</i> : FALSE CLAIM, MAKEMePAY, MAKEMeSAY, MISLEAD DEBATE; <i>rational, opinion</i> : REWRITE TEXT, RATIONALE, SUBJECTIVE OPINION; <i>rational, behaviour</i> : BENIGN REQUEST, REDUCE MEAT
Claims or actions	whether the persuadee is to believe something or to do something	<i>claims</i> : REWRITE TEXT, RATIONALE, SUBJECTIVE OPINION, FALSE CLAIM, MISLEAD DEBATE; <i>actions</i> : BENIGN REQUEST, MAKEMePAY, MAKEMeSAY, REDUCE MEAT

Table 8: The eighteen RQ2 groupings, each with its definition and the methods in each group. Results in Table 5.

Axis	Groups (number of methods)	Within	Between	Gap	95% CI	Perm. <i>p</i>	Adj. <i>p</i>
<i>How a method measures, or is designed</i>							
Scored	belief change 4, argument 2, behaviour 2	0.13	0.30	-0.17	[-0.22, -0.11]	0.29	0.98
Scorer	self-report 3, LLM-judge 2, rule 2, trained scorer 1	0.14	0.28	-0.13	[-0.21, -0.06]	0.44	1.00
Format	dialogue 6, dataset 2	0.16	0.37	-0.21	[-0.23, -0.11]	0.32	0.99
Interaction	multi-turn 5, static 2, single turn 1	0.35	0.19	0.16	[0.08, 0.22]	0.51	1.00
Pinned agent	persuadee 5, judge 2, none 1	0.16	0.31	-0.15	[-0.18, -0.03]	0.52	1.00
Persuadee present	persuadee 6, none 2	0.16	0.37	-0.21	[-0.23, -0.11]	0.32	0.99
Turns	single output 3, short 2, long 3	0.23	0.26	-0.03	[-0.08, 0.04]	0.85	1.00
Length (4 levels)	short 2, long 3, static 2, single turn 1	0.40	0.22	0.19	[0.10, 0.24]	0.27	0.98
Length (2 levels)	short 5, long 3	0.31	0.20	0.11	[0.01, 0.16]	0.39	1.00
Granularity	discrete 4, continuous 2, binary 2	0.42	0.18	0.25	[0.10, 0.29]	0.13	0.82
Relative scoring	relative 4, absolute 4	0.36	0.17	0.19	[0.08, 0.25]	0.029	0.32
Resistance	open 4, no persuadee 2, resistant 2	0.28	0.24	0.05	[-0.01, 0.15]	0.75	1.00
<i>What the task asks</i>							
Goal	comply 4, shift stance 3, accept falsehood 1	0.41	0.17	0.24	[0.17, 0.31]	0.12	0.80
Task type	rational persuasion 5, manipulative 3	0.45	0.06	0.40	[0.24, 0.44]	0.036	0.39
Topic	politics & society 3, factual knowledge 1, everyday life 2, open conversation 2	0.52	0.19	0.33	[0.25, 0.40]	0.037	0.39
Stake	none 4, its answer 1, personal commitment 2, money 1	0.56	0.13	0.42	[0.34, 0.48]	0.021	0.23
Task type (3 levels)	manipulative 3, neutral 3, prosocial 2	0.38	0.21	0.18	[0.12, 0.27]	0.16	0.89
Claims or actions	claims 4, actions 4	0.24	0.26	-0.02	[-0.06, 0.07]	0.91	1.00

Table 9: All eighteen RQ2 groupings, over the eight reliable methods. *Within*, *Between*, *Gap*, *Perm. p* and *Adj. p* as in Table 5. Table 8 lists the methods in each group. *95% CI*: over sampling units, with the eight methods fixed.

Model	MMLU-Pro	IFEval	<i>C</i>
DeepSeek-V4.1-Flash	90.7	96.1	+1.15
GPT-5.6-Sol	91.7	94.1	+1.10
Claude-Fable-5.1	91.7	93.3	+1.08
Kimi-K2.6	82.7	90.9	+0.76
Llama-3.3-70B	70.3	90.2	+0.41
GPT-OSS-120B	76.3	84.3	+0.35
Mistral-Medium-3.5	77.3	83.0	+0.32
Qwen3-32B	78.3	81.9	+0.31
EXAONE-4.0.1-32B	70.0	82.8	+0.13
Gemma-3-27B	65.0	83.7	+0.03
Gemma-3-12B	58.0	82.3	-0.20
Qwen2.5-7B	58.0	70.8	-0.63
Llama-3.2-3B	37.7	69.5	-1.20
Mistral-Nemo-12B	44.3	58.0	-1.45
Mistral-7B-v0.2	29.3	49.2	-2.17

Table 10: General capability of the fifteen models, ordered by *C*. *MMLU-Pro*: accuracy (%) on a fixed random subset of 300 questions, five-shot with chain of thought. *IFEval*: prompt-level strict accuracy (%) on all 541 prompts, zero-shot. Both ran with lm-evaluation-harness at temperature 0, or the default for GPT-5.6-Sol, Claude-Fable-5.1 and Kimi-K2.6, which accept no other. *C*: the mean of the two accuracies, each z-scored across the fifteen models. The two benchmarks rank the models alike (Spearman $\rho = 0.87$), and where a published five-shot chain-of-thought score exists, ours is within three points of it.