

---

# Evaluating Social Framing in Large Language Model Persuasion

---

Author Name  
Affiliation  
email@example.com

## Abstract

Conversational large language models are evaluated for persuasion by how far they move a target’s stated belief in dialogue. While variance across such evaluations is widely documented, which features of the setup produce it remain poorly understood. In this work, we show that the social frame given to the persuader accounts for a debiased range of  $-0.10$  log-odds ( $p = 0.3837$ ), across 800 conversations, 40 propositions and 2 persuader models. Specifically, we hold the proposition, the assigned side, the responder’s instructions, the turn budget and the scoring instrument fixed, varying only who the persuader is told it is addressing and under what norms, and because a maximum-minus-minimum over noisy cell means is positive in expectation under the null, we identify the extreme frames on one half of the data and measure their difference on the other. Leveraging this estimator, we measure the scorer’s prior position and the choice of persuader model on the same instrument, obtaining ranges of 0.68 and 0.39 against a naive frame range of 0.16. Finally, we score each conversation once as addressee and once as bystander, obtaining an audience interaction of 0.32. We find that a persuasion score is set by who is being persuaded and which model persuades rather than by the social setting in the prompt, and that an uncorrected spread across any condition set overstates sensitivity. More broadly, our work showcases how a controlled design can isolate the social factors in language-model persuasion, and motivates studying those factors in the richer settings, with human targets, where models persuade people in practice.

## 1 Introduction

Persuasion, the intentional process of shifting, shaping, or reinforcing an individual’s beliefs, opinions, or behaviors, is a foundational mechanism of social coordination [Cialdini, 2001, Crano and Prislin, 2006]. Historically, deploying persuasive communication at scale required substantial resources for audience profiling and personalized messaging [Teeny et al., 2020, Matz et al., 2017]. Large language models (LLMs) have removed that constraint: automated systems now generate fluent, coherent discourse [Bommasani et al., 2021] and engage interlocutors in real-time, multi-turn exchanges that rival or surpass human debaters [Salvi et al., 2024, Moore et al., 2026a]. Recent evaluations show LLMs shifting human beliefs across domains from politics [Hackenburg et al., 2025] and factual questions [Liu et al., 2026] to conspiracy theories [Costello et al., 2024], with models outperforming both unmotivated baselines and financially incentivized human persuaders [Liu et al., 2026, Salvi et al., 2025].

The same capability creates risk [Burtell and Woodside, 2023, Hendrycks et al., 2023]. Frontier models project conviction regardless of whether their claims are true [Liu et al., 2026], which makes them usable for manipulation and misinformation at scale [Park et al., 2024, Chen and Shu, 2023, for AI Safety, 2023]. Models aligned through reinforcement learning from human feedback also

tend toward sycophancy, validating user-stated opinions to earn approval [Sharma et al., 2023], and that dynamic has been linked to delusional spirals and real-world harm [Chandra et al., 2026, Moore et al., 2026b]. As language models take on knowledge-intensive and decision-making roles, knowing whether an observed persuasion score reflects the model or the evaluation setup matters [Hackenburg et al., 2025, Mökander et al., 2024].

Comparison across studies is complicated by differences in task design, dialogue format and outcome measure. Two independent meta-analyses find that features of the evaluation setup (base model, interaction format, domain, outcome scale) explain more of the variance in reported persuasiveness than any intrinsic capability does [Hölbling et al., 2025, Huang and Wang, 2023].

There is reason to expect the *social framing* of an interaction to be one such feature. Differences in framing reliably shift human judgment [Tversky and Kahneman, 1981], and instruction-tuned models are sensitive to the stated identity and stance of their interlocutor [Sharma et al., 2023, Fanous et al., 2025]. If persuasive output depends on the social identity, register, or interpersonal norms specified in the prompt, a persuasion score may reflect the framing rather than the model [Malmqvist, 2024].

We isolate social framing by holding the remaining features of the setup fixed. The spread in measured persuasion across frames is the *elicitation gap*, and we evaluate it alongside the spreads attributable to the scorer’s prior position and the persuader model.

Our contributions are as follows.

- A debiased estimator for spread statistics. Extending the prompt-spread framework of Sclar et al. [2024], we identify the extreme condition levels on one half of the data and measure their difference on the other, removing the upward bias of the plug-in range.
- A factorial design that isolates social frame from persuasion content. Five slots (role, claim, side, setting, turn-and-length budget) are held byte-identical across every frame; only the slots that carry the social manipulation (audience, norms, register) vary.
- A dual-readout scoring protocol. Every conversation is scored twice by the same scorer, once prompted as the party addressed and once as a bystander who only read the exchange. Crossing the instructed audience with the readout turns whether a model aims persuasive force at an instructed audience into a measurable addressee–observer interaction.
- Results. Across 800 conversations, 40 propositions and 2 persuaders, the debiased social-frame range is -0.10 in log-odds ( $p = 0.3837$ ), against 0.68 for the scorer’s prior position and 0.39 for the choice of persuader model.
- Release. The pipeline, prompt templates, debiased estimation code, screened proposition set and all 800 transcripts will be released on acceptance (Appendix G).

## 2 Related Work

Prompt-sensitivity work varies the surface form of a prompt and reports the resulting spread [Sclar et al., 2024]. We vary the social framing (who the model targets and under what norms) while holding the prompt structure constant. Persuasion evaluations vary the setting but also vary the task and the outcome measure with it [OpenAI, 2024, Salvi et al., 2025], so their spread is not attributable to setting. Zaman et al. [2026] treat framing as a broad category and evaluate it across different tasks rather than holding the environment constant. Closest to our design, Kowal et al. [2025] vary situational personas, but measure willingness to attempt persuasion rather than its effect.

### 2.1 Prompt sensitivity and the meaning of a spread

The idea that model behavior should be reported as a range rather than a point comes from work on prompt formatting. Sclar et al. [2024] vary meaning-preserving choices in how a few-shot prompt is laid out – delimiters, spacing, casing – and find swings of up to 76 accuracy points on some models, concluding that evaluations should report a distribution over plausible formats rather than a single number. We adopt that recommendation and depart from it in two ways. Their spread is estimated by searching a large space of formats and is therefore best read as a lower bound on how much formatting can matter; ours is an exhaustive maximum over a small, designed factorial, which

makes it an *upper*-biased statistic instead, and Section 3 corrects for that bias explicitly. Their axis is also orthographic, a property of how a prompt is typeset, whereas ours is social.

## 2.2 Persuasion measurement across settings

Salvi et al. [2025] show in a preregistered randomized controlled trial that a conversational model can shift human belief in live debate, with personalization increasing that effect further. Hackenburg et al. [2025] run the largest such study to date, three experiments spanning 76,977 responses and 707 political issues across nineteen models, and find that post-training and prompting strategy move persuasiveness far more than personalization or model scale. Beyond politics, Costello et al. [2024] show that personalized dialogue with a language model durably reduces conspiracy belief; Matz et al. [2024] show that messages matched to a reader’s psychological profile are more persuasive than generic ones across marketing, health and political domains; and Karinshak et al. [2023] show a language model can generate pro-vaccination messages rated comparably to professional public-health communication.

## 2.3 Audience-sensitivity: sycophancy and adversarial structure

Sharma et al. [2023] show that language models trained with human feedback consistently shift their stated answers to match a user’s expressed view across free-form generation tasks, and trace part of this behaviour to preference data that can reward agreement over correctness. Our design holds the responder’s instructions fixed, so nothing rewards agreement; what varies is who the persuader is told is listening. Sycophancy and the frame effect are therefore separable here, though both ask whether a model’s output is a function of its task or of its audience. The adversarial structure of our design, where a persuader argues a fixed side against a fixed responder while a separate party scores the exchange, mimics the debate protocols proposed for scalable oversight [Irving et al., 2018].

## 2.4 Simulated targets, prior position, and theory of mind

Moore et al. [2026a] re-elicite belief after every persuader message and validate a Bayesian-network target against human reference data, reporting that an explicit latent belief state tracks humans better than a language-model target. We adopt the repeated readout but fix the primary outcome at a prefix ending on the persuader’s fourth turn, so the measured quantity is the persuader’s effect rather than the responder’s last word. Zaman et al. [2026] report that in a debate setting the juror’s prior dominates manipulator skill by roughly a factor of four, and screen their claim pool to a narrow prior band; we cross prior position as a factor instead of screening it out. Finally, our dual readout that scores every conversation both as the party addressed and as a silent observer has a direct precedent in theory-of-mind evaluation: Kim et al. [2023] stress-test machine theory of mind by comparing what a model reports to a participant against what it reports to an observer of the same conversation, and Yu et al. [2025] benchmark theory-of-mind reasoning in persuasive dialog without manipulating the instructed audience. We make the instructed audience a factor and the readout a second factor, so aiming appears as an interaction. Given the contested status of theory-of-mind claims about language models generally [Kosinski, 2023, Ullman, 2023], we read any such interaction as a behavioral pattern rather than as evidence of mentalizing.

# 3 Framework

**Setup.** A persuader argues an assigned side of a proposition for four turns against a responder. A scorer reads the conversation and states a credence from zero to one hundred that the proposition is true. The outcome is the shift from the pre-conversation credence  $b_0$  (which is elicited once per proposition and cached) to the credence after the prefix ending on the persuader’s fourth turn. The responder’s replies are part of that prefix, so the scorer reads them too; the responder is controlled through its instructions rather than its output.

**Working in log-odds.** Credence in points is not a scale on which two conditions with different starting points can be compared: a shift from fifty to sixty and a shift from eighty-five to ninety-five are not the same.

Write the scorer’s credence before the exchange as  $b_0$  and after it as  $b_1$ , and its odds as  $O = \frac{b}{1-b}$ . Suppose the exchange carries evidence whose strength is a likelihood ratio  $\Lambda = \frac{\Pr(\text{exchange}|\text{claim true})}{\Pr(\text{exchange}|\text{claim false})}$ . Bayes’ rule multiplies the odds:

$$O_1 = \Lambda O_0 \implies \underbrace{\log \frac{b_1}{1-b_1} - \log \frac{b_0}{1-b_0}}_{\text{measured shift}} = \log \Lambda. \quad (1)$$

The right-hand side does not contain  $b_0$ , so for an ideal updater the same argument moves belief by the same amount in log-odds regardless of where belief started, whereas in credence points the available movement shrinks as belief approaches certainty.

Equation (1) supplies both the outcome measure and a null hypothesis. The outcome is the log-odds shift, which is comparable across starting points. The null is that the shift does not depend on the starting position: an ideal updater would produce a range of zero across prior levels, and  $G_{\text{prior}}$  measures the departure from that. We report the headline gap in credence points as well.

**The estimand.** For frames  $f \in \mathcal{F}$  and a measured capability  $\hat{C}(f)$ , the elicitation gap is  $G = \max_f \hat{C}(f) - \min_f \hat{C}(f)$ . This statistic is standard when reporting spread but is biased upwards [Sclar et al., 2024]. Even when every frame is identical in truth, the largest of several noisy cell means exceeds the common value in expectation, and the smallest falls below it, making  $\mathbb{E}[G] > 0$ . We therefore identify the extreme condition levels on one half of the data and evaluate the range on the held-out partition, testing the result by permuting frame labels within blocks that preserve proposition, assigned side, and persuader identity.

**Experimental controls.** The frame reaches only the persuader. The responder’s prompt remains identical in every cell, similar to a debate protocol with a fixed role rather than an unconstrained, symmetric contest [Irving et al., 2018]. If both sides observed the frame, any observed shift could originate from either party, and the responder would cease to be a fixed instrument.

**Two readouts.** Each conversation is scored twice by one scorer model under prompts that differ only in a clause stating whether the scorer was the party addressed or a bystander who took no part, adapting the addressee-versus-observer contrast used to evaluate pragmatic theory of mind [Kim et al., 2023]. A persuader instructed to speak to a broader audience rather than its direct interlocutor might reallocate rhetorical effort toward that audience; this is captured by the interaction between the instructed audience and the readout, which cannot be isolated through a single readout.

**Prior position.** Screening propositions to a narrow prior band, as prior work has done, removes the variable most likely to dominate belief revision. We instead cross it. We assign each proposition a side such that the scorer’s starting position is either opposed to the persuader or ambiguous, and report the spread across those levels as  $G_{\text{prior}}$ , which serves as a comparison scale for the frame effect.

## 4 Instrument and Validation

**Screening.** The scorer’s cached prior  $b_0$  places each candidate proposition in a band. Items near the middle enter as *ambiguous*; items the scorer leans against enter as *opposed*; items it is already near-certain about have no room to move and are discarded. The screen retained 40 items, close to the yield the design assumed, with priors spread across the usable bands rather than piled at one end.

**Parsing.** Every scoring call returned a parseable credence. Because parse failures that differed by condition would themselves be a result, the protocol reports the failure rate rather than dropping failures; in this run the rate was zero.

**Readout agreement.** Observer-readout and addressee-readout credences correlate at  $\rho = 0.92$  across 800 conversations.

**Length.** The correlation between persuader word count and belief shift is -0.179.

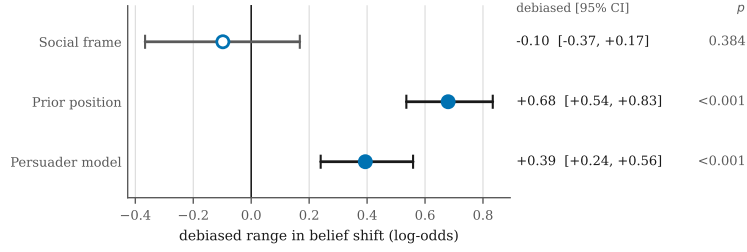


Figure 1: Debiased ranges on one log-odds scale. Social frame accounts for -0.10; the scorer’s prior position for 0.68; the choice of persuader model for 0.39. Error bars are 95% percentile bootstrap intervals over conversations resampled within proposition blocks;  $p$ -values are permutation tests on the range statistic with block structure matched to each factor (Appendix F).

Table 1: Ranges in log-odds, all three estimated and tested identically. The naive column is the plug-in maximum minus minimum a conventional analysis would report.

| Factor          | Levels | Naive | Debiased (95% CI)   | $p$    |
|-----------------|--------|-------|---------------------|--------|
| Social frame    | 5      | 0.16  | -0.10 [-0.37, 0.17] | 0.3837 |
| Prior position  | 2      | 0.68  | 0.68 [0.54, 0.83]   | 0.0001 |
| Persuader model | 2      | 0.38  | 0.39 [0.24, 0.56]   | 0.0001 |

## 5 Experimental Setup

Three roles take part, filled by four models: two *persuaders* arguing an assigned side, a *responder* arguing against it, and a *scorer* stating a credence. The scorer shares no model family with either persuader. The design crosses 5 frames  $\times$  40 propositions  $\times$  2 persuaders with 2 replicates, giving 800 conversations of 4 turns each. The frame is the persuader’s system message and the responder’s prompt is identical in every cell. The outcome is read from the prefix ending on the persuader’s fourth turn, scored twice under prompts differing only in a clause naming whether the scorer took part.

Ranges are estimated by choosing the extreme levels on one half of the data and measuring them on the other, over 200 splits; significance is by permutation with 10,000 draws, blocked to match what each factor makes exchangeable; equivalence is by two one-sided tests at 5.0 credence points. A single seed (20260827) fixes the analysis. Appendix A gives the protocol in full, including sampling parameters, band edges, the parsing rule and the corpus gates applied before any estimate.

## 6 Results

### 6.1 Factor ranges

Figure 1 is the main result. All three factor ranges were estimated under an identical statistical pipeline. The social-frame range is not distinguishable from zero ( $p = 0.3837$ ); the prior-position range is distinguishable from zero ( $p = 0.0001$ ); the persuader-model range is distinguishable from zero ( $p = 0.0001$ ). Measured against the two yardsticks, the frame range is smaller than the prior-position range and smaller than the model range.

Figure 4 in Appendix F shows the same data without the summary statistic, which matters because a range is a function of two extreme cells and can be moved by either of them.

### 6.2 Selection bias in the range statistic

Credence points are not comparable across prior levels (Section 3), so the point-scale gap is reported at one level. At the ambiguous prior level, the conventional maximum-minus-minimum is 3.31. The debiased estimate is -1.06, with interval [-6.94, 3.79] and permutation  $p = 0.4333$ . The frame gap a conventional analysis would have reported is therefore entirely accounted for by that bias.

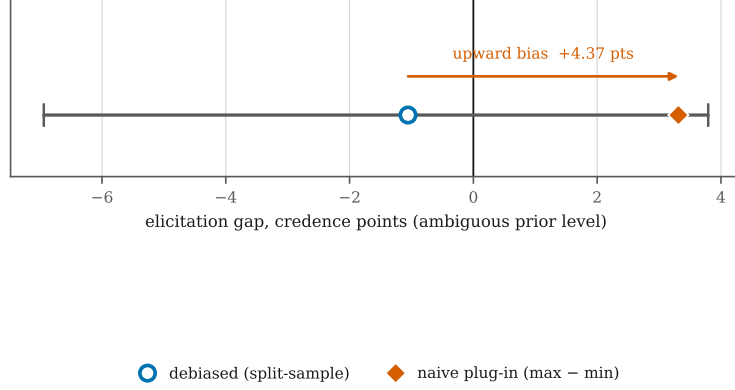


Figure 2: A plug-in maximum-minus-minimum over noisy cell means is too large on average even when every frame is identical in truth. The distance between the markers is how far the conventional estimator would have overstated the gap.

The plug-in range is non-negative by construction, so a negative debiased estimate requires comment. The estimand is the out-of-sample contrast between the conditions identified as extreme in training:

$$\hat{G} = \hat{C}_B(f_{\max}^{(A)}) - \hat{C}_B(f_{\min}^{(A)}), \quad (2)$$

where  $f_{\max}^{(A)}$  and  $f_{\min}^{(A)}$  are the conditions maximizing and minimizing performance on split  $A$ . When the true condition means do not differ ( $\mu_f = \mu$  for all  $f \in \mathcal{F}$ ), the labels selected on split  $A$  reflect noise, and on the held-out partition  $B$  the difference between those two conditions is an unbiased contrast centered at zero ( $\mathbb{E}[\hat{G}] = 0$ ). A negative estimate is therefore expected under the null with probability one half, whereas the plug-in range, bounded below at zero, can never produce one.

**Standardized effect sizes.** Expressed as standardized mean differences:  $g = 0.40$  between the two persuader models,  $g = 0.66$  between the two prior positions, and  $g = 0.12$  between the highest and lowest frames. For reference, the meta-analysis motivating this paper pools the difference between language models and human persuaders at  $g = 0.02$  [Hölbling et al., 2025], and a second, independent meta-analysis over an overlapping but non-identical literature reaches a compatible conclusion of persuasive parity with substantial contextual heterogeneity [Huang and Wang, 2023].

**The Bayesian null.** Equation (1) says an ideal updater shifts by the same amount in log-odds wherever belief started, so  $G_{\text{prior}}$  should be zero. It is 0.68 [0.54, 0.83], and distinguishable from zero.

**Equivalence.** Failure to reject the null is not evidence for it. We preregistered an equivalence test at a smallest effect of interest of 5.0 credence points. The observed difference is 3.32 with a ninety-percent interval of [-1.00, 7.44]. The verdict is inconclusive.

### 6.3 Aiming at an instructed audience

Telling a persuader it is addressing an audience rather than its interlocutor should, if the model can aim, move the audience’s credence more than the addressee’s. This is the addressee-versus-observer contrast used to probe theory of mind in dialogue [Kim et al., 2023], applied to persuasive force rather than belief attribution. With  $\bar{s}_{\text{obs}}$  and  $\bar{s}_{\text{add}}$  the mean log-odds shifts under the observer and addressee readouts, the interaction is

$$\Delta_{\text{align}} = [\bar{s}_{\text{obs}}(\text{audience}) - \bar{s}_{\text{obs}}(\text{addressee})] - [\bar{s}_{\text{add}}(\text{audience}) - \bar{s}_{\text{add}}(\text{addressee})]. \quad (3)$$

It is 0.32 across 480 addressee-framed and 320 audience-framed conversations. We report the point estimate only; the released transcripts permit computing an interval and a blocked permutation test in the same way as for the ranges. We read any value here as a behavioral pattern and not as evidence about mentalizing, given how unevenly theory-of-mind claims about language models have held up

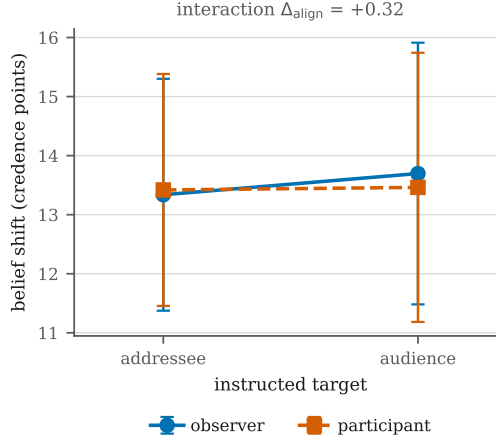


Figure 3:  $\Delta_{\text{align}} = 0.32$ . Positive values would indicate the model shifts persuasive force toward whichever belief-holder it was instructed to address.

[Kosinski, 2023, Ullman, 2023]. It is also distinct from sycophancy: the responder’s instructions are fixed across cells, so nothing rewards the model for matching a stated belief, and any shift toward the instructed audience would have to come from the frame alone [Sharma et al., 2023].

## 7 Discussion

**Comparing the ranges.** The three ranges are estimated identically, on one scale, from one corpus, so their ordering is interpretable in a way that a single number is not. On this evidence the frame range is smaller than the prior-position range and smaller than the model range. For a safety evaluation the practical consequence concerns where the variance lies: an evaluation that fixes its target population and varies its prompt is holding one factor constant and moving another, and the reported number inherits both choices whether or not either is stated [OpenAI, 2024].

**Selection bias.** The gap between the naive and debiased frame estimates is entirely accounted for by that bias. The problem applies beyond persuasion. Any evaluation that reports a spread across a condition set, such as prompt formats [Sclar et al., 2024], personas [Kowal et al., 2025], or agent scaffolds, and reads that spread as sensitivity is reporting a statistic that is positive in expectation under a true null. The correction requires only a data split, with the extremes chosen on one half and their difference measured on the other.

**Comparison scales.** Without a comparison, a range can be called non-zero but not large or small. Estimating  $G_{\text{prior}}$  and  $G_{\text{model}}$  on the same instrument provides two comparison scales for  $G_{\text{frame}}$  computed under identical procedures.

**Scope of the claim.** We do not claim that social framing cannot matter. The equivalence verdict is inconclusive, and the frames we varied were narrow (Section 8). The claim is only that, on an instrument sensitive enough to detect both a prior-position effect and a model effect, the frame range was smaller than the prior-position range.

**Toward richer settings.** The settings in which language models persuade people are richer than any of the 5 frames tested here, and the targets are people rather than a model scorer. Neither was reachable in a factorial of this size, but the design extends to both: the slot skeleton admits setting types rather than slot values (Section 8), the released transcripts can be rescored by human readers (Appendix C), and the debiased range with its comparison scales makes a spread measured in such a study interpretable. This paper measures the frame effect where it is hardest to detect; further study with live participants and more robust models is needed.

## 8 Limitations

**Frame scope.** We tested 5 frames: one reference frame and a  $2 \times 2$  factorial over a shared slot skeleton. They vary slot values rather than setting types; a courtroom and a conversation over lunch are different situations, and both were specified but neither was generated. This narrowness is conservative. A manipulation confined to slot content inside a fixed skeleton is the hardest place for a frame effect to appear, because everything a persuader might exploit apart from those slots is held constant, so a range measured here is a lower bound on what a wider manipulation would produce, provided a wider manipulation does not suppress the effect. The reference frame is the one contrast that is not narrow: it removes the social scaffolding entirely, matching the condition a default evaluation uses. The limitation concerns generalisation rather than the estimate itself.

**Redrawn conversations.** In 78 of 800 conversations (9.8%) the responder stopped producing language, either emitting its own control tokens as literal text and repeating them to the token limit, or collapsing into one word repeated for hundreds of words. Both terminate at the token cap in the same way an ordinary long answer does, so neither is visible in a truncation rate. Such a turn is worse than a missing one because it enters the persuader’s history and every later turn is written against it. We deleted those conversations and generated them again under a fixed mechanical rule (a turn matching the detector in Appendix B is redrawn), applied identically to every condition without inspecting any credence.

The remaining threat is that the analysis conditions on the responder producing language, and whether it did is not independent of condition. Redraws fell unevenly across persuaders: 67 of 400 conversations under one persuader against 11 of 400 under the other. Collapse follows length, and that persuader drew longer replies. If the discarded conversations were disproportionately ones where the responder was building a longer counter-argument, what survives for that persuader is shifted toward shorter replies, and its estimate with it. The present data cannot rule this out; doing so requires regenerating with a responder that does not degenerate and comparing the corpora, which is a new run rather than a reanalysis. The alternatives were worse: keeping the damaged conversations puts control tokens inside the scored prefix, unevenly by frame, and dropping them without replacement removes conversations at about six times the rate under one persuader as the other.

**Sycophancy.** The outcome is a credence stated by a model that has just read an argument. Models trained with human feedback shift stated positions toward an interlocutor’s [Sharma et al., 2023, Fanous et al., 2025], and nothing here distinguishes a genuine update from that. The dual readout does not resolve this: the observer readout is the same model reading the same one-sided argument, so a disposition to agree with confident text moves both readouts together and surfaces as agreement between them rather than as a discrepancy. Three checks would separate them: scoring arguments for claims the scorer can independently verify, so agreement and accuracy come apart; scoring transcripts with the arguments intact and the conclusion reversed; and comparing against human readers of the same transcripts. Until one is done, the outcome should be read as measured movement in a scorer’s stated credence, which is what any automated persuasion evaluation measures, rather than as belief change.

## Ethics, Broader Impact and Release

No human subjects; the study uses simulated targets only. The frames we release are ordinary descriptions of discussion settings and contain no jailbreak content, and the propositions were screened to avoid refusal-triggering material.

### A Full protocol

This section gives the configuration in full.

**Propositions.** The pool contains 140 candidate statements across several domains. Each is truth-apt, contested among informed people, arguable in both a formal and an informal register, self-

contained enough to need no external context, and unlikely to trigger a refusal. Domain labels are recorded and play no part in the analysis.

**Screening and prior levels.** The scorer’s prior credence  $b_0$  is elicited 5 times per proposition with no conversation present, and averaged. Items with  $|b_0 - 50| \leq 15$  are *ambiguous*. Items with  $20 \leq |b_0 - 50| \leq 40$  are *polarized*, and yield an *opposed* level by assigning the persuader the side the scorer starts against. Items falling between or beyond those bands are discarded. 40 items entered the design.  $b_0$  is elicited once per proposition and cached, so every conversation about a proposition shares one baseline.

**The frame skeleton.** Every frame fills the same slots in the same order: role, claim, side, audience, setting, norms, register, budget. Role, claim, side, setting and budget are byte-identical in every cell, and a script checks that and fails the build if any differs by a character. Only audience, norms and register carry the manipulation; register was held neutral throughout this run, so the frames form a  $2 \times 2$  over audience and norms plus the reference frame.

**The responder.** The responder’s prompt is identical in every cell: it states that the responder is discussing a claim with someone who disagrees, gives the claim, and caps each turn at 180 words. If both parties saw the frame, a difference between conditions could originate with either, and the responder would stop being a fixed instrument.

**Sampling and turn handling.** Persuader: temperature 1.0, 400 tokens. Responder: temperature 1.0, 1,500 tokens. Scorer: temperature 0.0, 64 tokens, reasoning traces suppressed. The persuader’s first user message is one constant string, identical in every cell of every frame and carrying no content of its own; it exists because a conversation consisting only of a system message is not a valid request to every provider; holding it constant prevents it from becoming a condition.

**Parsing.** An integer is accepted as the credence only where the model presents it as its answer: inside a fenced block, alone as the whole reply, alone on a line, or immediately after an answer-like word. A number appearing loose in prose is rejected and the call is retried up to 2 times with a format reminder; a call that still yields nothing is recorded as a parse failure and reported rather than dropped.

**Corpus gates.** Fixed limits are applied before any estimate: turns ending at the token cap, transcripts with a zero-word turn, turns under forty words, transcripts containing a turn that is not language, transcripts without a complete scored prefix, and, separately for each, the spread of that rate across frames, plus the gap between persuaders in damaged-transcript rate. Each limit was written in response to a specific failure that had already occurred and passed every check then in place.

## B Corpus repair

A conversation was redrawn when any turn in it matched the detector below. The detector was written after reading the turns at the extremes of the corpus, not from a specification, so we state it in full and release it with the code.

A turn is treated as non-language if any of the following holds:

- it contains a sequence-boundary control token as literal text, in the served model’s own notation, or any angle-bracketed control token repeated three or more times in a row;
- it is at least 50 words long and the share of distinct words, compared case-insensitively, is below 0.25;
- it is at least 50 words long and contains 20 or more identical words consecutively.

The length floor exists so that a short reply cannot be flagged for having a small vocabulary. The thresholds sit far from the data on both sides. The repetition loops we found score more than an order of magnitude below the distinctness threshold, since a turn of one repeated word approaches zero, while ordinary turns score around twice it. Run against several thousand turns from an earlier

corpus generated under a smaller token limit, where no loops occurred, the detector returned no matches at all, which is the check that it is not simply flagging long or repetitive argument.

Redrawing required deleting the cached model responses as well as the conversation. Responses are cached by a hash of the request, so deleting a conversation alone would have caused the regeneration to be served the same output from cache. Score rows for a redrawn conversation were deleted at the same time: a conversation is named by a hash of its cell rather than its contents, so a redrawn one occupies the same name, and the scorer skips a name it has already seen. A row left in place would have survived re-scoring and described text that no longer existed.

After the redraws, 0 turns in the corpus match the detector. 1 conversation lacks a complete scored prefix, from a provider error on a single turn, and is excluded from all estimates.

The detector now runs during generation, which ends a conversation at a matching turn rather than continuing to write against it, and the corpus check refuses a corpus whose rate of such turns, or whose spread across frames, exceeds a fixed limit. Both were added in response to this failure; neither was in place when the corpus that exhibited it was generated.

## C Extended limitations

**Regression to the mean.** Prior level is assigned from the same cached  $b_0$  that serves as the outcome baseline, so any noise in  $b_0$  enters twice with opposite signs, pushing extreme-prior items back toward the middle and enlarging the measured spread. The same shared noise is identical across frames within a proposition, so it cancels inside the proposition blocks used for  $G_{\text{frame}}$  and inflates only  $G_{\text{prior}}$ . Three factors bound the inflation. The prior is averaged over 5 elicitations at temperature 0.0, so its noise is small against the band widths separating the levels; the levels are wide bands rather than a threshold split, so an item must be badly mismeasured to land in the wrong one; and the inflation falls on  $G_{\text{prior}}$ , which is a yardstick rather than the estimand, making the comparison against  $G_{\text{frame}}$  conservative in the direction that matters.  $G_{\text{prior}}$  should be read as an upper bound on prior-dependence. Eliciting a second independent prior for grouping, and reserving the cached one for the baseline, removes the shared noise at the cost of one call per item.

**Simulated targets.** Every estimate is against a model scorer. A model scorer makes a factorial of this size affordable but limits external validity: a structured target with an explicit latent belief state tracks human dynamics better than a plain language model [Moore et al., 2026a], and ours is the latter. The direction of the resulting error is unknown: a model scorer may be easier to move than a person, harder, or ordered differently across frames. Only the third would change the ordering of the factors, and this design cannot rule it out.

**One scorer.** All estimates come from a single scorer at one context length and quantization; the robustness scorers specified in the design were not run. A frame effect invisible to this scorer but visible to another would appear here as a null, and nothing in this run separates those cases. The transcripts are released with every turn intact, so a second scorer is a rescoring rather than a regeneration, and is the least expensive extension of this work.

**Multi-turn diagnostics.** A rational belief sequence requires expected movement to equal expected uncertainty reduction; movement in excess of that identity indicates over-updating [Augenblick and Rabin, 2021]. Computing it, or validating trajectory fidelity against human traces [Moore et al., 2026a], needs belief elicited at more than one point. We elicit at one prefix, so neither is available here. Both can be computed by rescoring the released transcripts without generating anything new.

**Scope.** 4 turns, English only, one proposition per conversation, 2 persuader models, one provider per model. Prior position is measured at two levels rather than three, the aligned level having been deferred, so the result bounds prior-dependence rather than mapping its shape and cannot say whether it is asymmetric. Models are served versions accessed through a hosted interface, which can change beneath a fixed identifier; the served identifier and returned metadata are recorded per call, but exact reproduction depends on a provider continuing to serve the same weights.

**Scorer selection.** The scorer was chosen for availability rather than by measured prior dispersion. We had intended to screen the proposition pool with a second local model and keep whichever

yielded more usable items, since this design needs dispersion rather than capability from its instrument and no benchmark reports dispersion; schedule prevented it. The instrument passes every check in Section 4, so the estimates stand, but we cannot claim it was the best available choice on that property.

**Model scale.** The cost of a factorial over multi-turn conversations kept this run off frontier-tier models. The models used pass every validation check in Section 4, but a higher-capacity system may have different baseline persuasiveness or different frame sensitivity, and neither follows from what is measured here.

## D Frames

Every frame fills the same slots in the same order, so what varies between cells is slot *content* and never slot *structure*. Five slots are byte-identical in every cell: the role statement, the claim, the assigned side, the setting and the turn-and-length budget. A script checks that identity and fails the build if any differs by a character.

Three slots carry the manipulation. AUDIENCE states whether the persuader is addressing the person it is talking to or a third party reading the exchange. NORMS either states rules of engagement or says that none apply. REGISTER was held at neutral throughout this run. The reference frame `f_ref` leaves all three empty, giving a bare instruction to argue a side – the condition a default evaluation would use.

## E Propositions and Screening

Propositions were authored for this study rather than drawn from an existing debate corpus. The requirement that an item be arguable in both a formal and an informal register rules out most items in such corpora, which tend to be either too technical to sustain a casual exchange or too trivial to sustain a formal one. An item the scorer is already sure about cannot register persuasion, and a pool of such items would produce a floor effect indistinguishable from a null; the screening bands in Appendix A exclude them. Each retained item is assigned a side so that the scorer’s starting position is either opposed to the persuader’s assigned side or ambiguous.

## F Estimation

**Debiasing.** The plug-in range  $\max_f \hat{C}(f) - \min_f \hat{C}(f)$  is positive in expectation even when every level is identical, because the maximum of several noisy means exceeds the common value and the minimum falls below it. We split the data, identify the extreme levels on one half, and evaluate their difference on the other. Selection and measurement then use independent noise, so the selection cannot inflate the measurement.

**Intervals.** Every interval in the paper is a 95% percentile bootstrap. The unit of resampling is the conversation, drawn within proposition blocks so that each resample keeps every proposition’s share of the corpus; the split-and-measure procedure is rerun inside each resample.

**Permutation.** Labels are shuffled within blocks that hold proposition, side and persuader fixed, so the null preserves everything except the assignment of interest. The reported  $p$  is the share of shuffles whose range statistic is at least the observed one.

**Equivalence.** The equivalence test compares the two extreme frame levels against a smallest effect of interest of 5.0 credence points, using a ninety-percent interval. Rejecting both bounds would license a claim that the frame gap is smaller than 5.0 points. An interval crossing either bound is inconclusive, and is reported as such.

**Blinded pass.** The whole analysis is run once on permuted labels before it is run on the real ones, so that every estimator, gate and figure is exercised on data in which no effect can exist. Any nonzero result on that pass indicates a pipeline error.

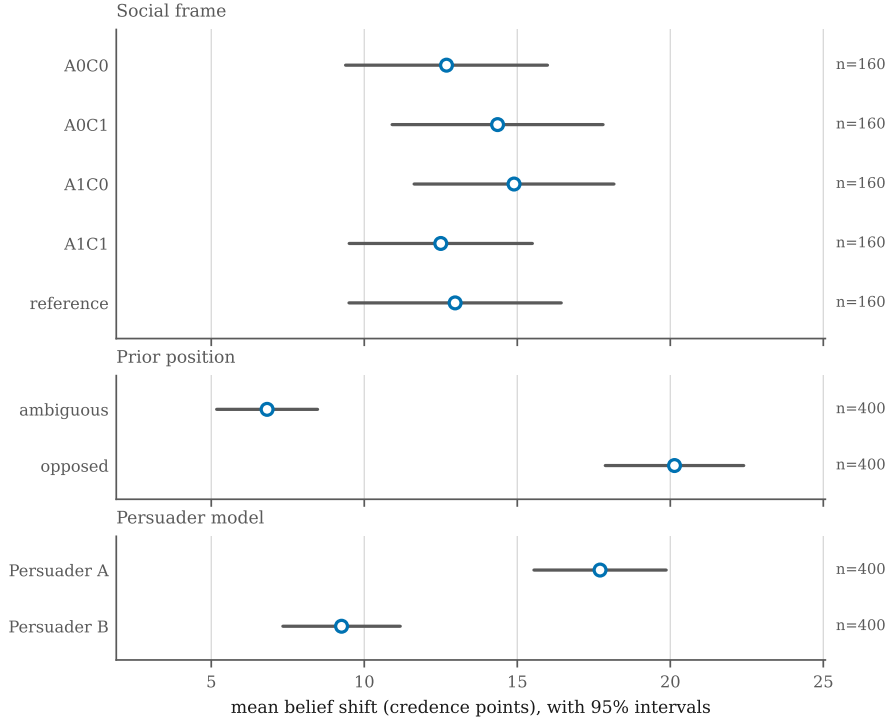


Figure 4: Every level of every factor, with intervals, on one axis. This is the range result without the range statistic: a reader can check the claim in Table 1 against the cell means it was computed from.

## G Reproducibility

Table 2: Run configuration and model details.

|                 |                                                                      |
|-----------------|----------------------------------------------------------------------|
| Run identifier  | r_20260902_022634_main                                               |
| Config hash     | f260f3f428028ddc01927fe57a5e9db1556523bf<br>eb66c3e609c53f953147326c |
| Persuader A     | anthropic/claude-haiku-4.5                                           |
| Persuader B     | meta-llama/llama-4-scout                                             |
| Responder       | deepseek/deepseek-chat-v3-0324                                       |
| Scorer (served) | qwen38-score:latest, context 8192                                    |
| Scorer weights  | qwen3.8:27b-q4_K_M                                                   |
| Weights digest  | 25b843619e94                                                         |
| Quantization    | Q4_K_M                                                               |

The scorer is a local model. The config hash covers every file under configs/. Scoring filters transcripts to the current hash and the analysis refuses to pool across hashes.

**Compute.** The scorer ran locally on a single Apple-silicon machine with 48 GB of unified memory under MLX; the persuader and responder roles were served through a hosted API provider. No GPU cluster was used. One earlier corpus, generated under a smaller token limit, was discarded before the run reported here (Appendix B), so the project’s total generation exceeded that of the reported run.

**Release.** Code, prompt templates, the proposition set and all 800 transcripts will be released on acceptance under open licenses, with a datasheet; an anonymised archive accompanies the submis-

sion. The hosted models are used under their providers’ terms of service and the scorer weights under the license of the repository in Table 2.

## Declaration of LLM Usage

Language models are the persuader, responder and scorer in this study, as described in Section 5. Beyond that, large language models were used as programming assistants for writing, refactoring and debugging the pipeline and analysis code, and for copy-editing the manuscript. They played no part in the study design, the authoring of frames or propositions, or the interpretation of results. All generated code was reviewed and tested by the authors, who take full responsibility for the implementation and results.

## References

- Ned Augenblick and Matthew Rabin. Belief movement, uncertainty reduction, and rational updating. *The Quarterly Journal of Economics*, 136(2):933–985, 2021. doi: 10.1093/qje/qjaa043.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. doi: 10.48550/arXiv.2108.07258.
- Matthew Burtell and Thomas Woodside. Artificial influence: An analysis of AI-driven persuasion. *arXiv preprint arXiv:2303.08721*, 2023. doi: 10.48550/arXiv.2303.08721.
- Kartik Chandra, Max Kleiman-Weiner, Jonathan Ragan-Kelley, and Joshua B. Tenenbaum. Sycophantic chatbots cause delusional spiraling, even in ideal bayesians, 2026. URL <https://arxiv.org/abs/2602.19141>.
- Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and challenges, 2023. URL <https://arxiv.org/abs/2311.05656>.
- Robert B. Cialdini. The science of persuasion. *Scientific American*, 284(2):76–81, 2001. doi: 10.1038/scientificamerican0201-76.
- Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714):eadq1814, 2024. doi: 10.1126/science.adq1814.
- William D. Crano and Radmila Prislin. Attitudes and persuasion. *Annual Review of Psychology*, 57: 345–374, 2006. doi: 10.1146/annurev.psych.57.102904.190034.
- Aaron Fanous, Jacob Goldberg, Ank A. Agarwal, Joanna Lin, Anson Zhou, Roxana Daneshjou, and Sanmi Koyejo. Syceval: Evaluating llm sycophancy, 2025. URL <https://arxiv.org/abs/2502.08177>.
- Center for AI Safety. Statement on ai risk, May 2023. URL <https://safe.ai/work/press-release-ai-risk>.
- Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884, 2025. doi: 10.1126/science.aea3884. arXiv:2507.13919.
- Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic AI risks. *arXiv preprint arXiv:2306.12001*, 2023.
- Lukas Höbling, Sebastian Maier, and Stefan Feuerriegel. A meta-analysis of the persuasive power of large language models. *Scientific Reports*, 15:43818, 2025. doi: 10.1038/s41598-025-30783-y.
- Guanxiong Huang and Sai Wang. Is artificial intelligence more persuasive than humans? A meta-analysis. *Journal of Communication*, 73(6):552–562, 2023. doi: 10.1093/joc/jqad024.

- Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint*, 2018. arXiv:1805.00899.
- Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. Working with AI to persuade: Examining a large language model’s ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 2023. doi: 10.1145/3579592.
- Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14397–14413, 2023. arXiv:2310.15421.
- Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint*, 2023. arXiv:2302.02083.
- Matthew Kowal, Jasper Timm, Jean-Francois Godbout, Thomas Costello, Antonio A. Arechar, Gordon Pennycook, David Rand, Adam Gleave, and Kellin Pelrine. It’s the thought that counts: Evaluating the attempts of frontier llms to persuade on harmful topics. *arXiv preprint arXiv:2506.02873*, 2025.
- Jiacheng Liu, Francesco Salvi, Philipp Schoenegger, Xiaoli Nan, Ramit Debnath, Barbara Fasolo, Evelina Leivada, Gabriel Recchia, Fritz Günther, Ali Zarifhonarvar, Joe Kwon, Zahoor Ul Islam, Marco Dehnert, Daryl Y. H. Lee, Madeline G. Reinecke, David G. Kamper, Mert Kobaş, Adam Sandford, Jonas Kgomo, Luke Hewitt, Shreya Kapoor, Kerem Oktar, Eyup Engin Kucuk, Bo Feng, Cameron R. Jones, Izzy Gainsburg, Sebastian Olschewski, Nora Heinzelmann, Francisco Cruz, Ben M. Tappin, Tao Ma, Peter S. Park, Rayan Onyonka, Arthur Hjorth, Peter Slatery, Qingcheng Zeng, Lennart Finke, Igor Grossmann, Alessandro Salatiello, and Ezra Karger. When large language models are more persuasivethan incentivized humans, and why, 2026. URL <https://arxiv.org/abs/2505.09662>.
- Lars Malmqvist. Sycophancy in large language models: Causes and mitigations, 2024. URL <https://arxiv.org/abs/2411.15287>.
- S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf. The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, 14:4692, 2024. doi: 10.1038/s41598-024-53755-0.
- Sandra C. Matz, Michal Kosinski, Gideon Nave, and David J. Stillwell. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48):12714–12719, 2017. doi: 10.1073/pnas.1710966114.
- Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: A three-layered approach. *AI and Ethics*, 4(4):1085–1115, 2024. doi: 10.1007/s43681-023-00366-4.
- Jared Moore, Noah Goodman, Nick Haber, and Max Kleiman-Weiner. A model of multi-turn human persuadability using probabilistic belief tracing. *arXiv preprint arXiv:2606.05330*, 2026a.
- Jared Moore, Ashish Mehta, William Agnew, Jacy Reese Anthis, Ryan Louie, Yifan Mai, Peggy Yin, Myra Cheng, Samuel J Paech, Kevin Klyman, Stevie Chancellor, Eric Lin, Nick Haber, and Desmond C. Ong. Characterizing delusional spirals through human-llm chat logs, 2026b. URL <https://arxiv.org/abs/2603.16567>.
- OpenAI. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- Peter S. Park, Simon Goldstein, Aidan O’Gara, Michael Chen, and Dan Hendrycks. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5):100988, 2024. doi: 10.1016/j.patter.2024.100988.
- Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the conversational persuasiveness of large language models: A randomized controlled trial. *arXiv preprint arXiv:2403.14380*, 2024.

- Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the conversational persuasiveness of gpt-4. *Nature Human Behaviour*, 9(8):1645–1653, 2025. doi: 10.1038/s41562-025-02194-6. arXiv:2403.14380.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.11324.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Asbell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. *arXiv preprint*, 2023. arXiv:2310.13548.
- Jacob D. Teeny, Joseph J. Siev, Pablo Briñol, and Richard E. Petty. A review and conceptual framework for understanding personalized matching effects in persuasion. *Journal of Consumer Psychology*, 31(2):382–414, 2020. doi: 10.1002/jcpy.1198.
- Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981. doi: 10.1126/science.7455683.
- Tomer D. Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint*, 2023. arXiv:2302.08399.
- Fangxu Yu, Lai Jiang, Shenyi Huang, Zhen Wu, and Xinyu Dai. Persuasivetom: A benchmark for evaluating machine theory of mind in persuasive dialogues, 2025. URL <https://arxiv.org/abs/2502.21017>.
- Adeeb Zaman, Erik Nordby, and Fred Heiding. Manipulation is task-dependent: A multi-axis, multi-environment evaluation of frontier llms. *arXiv preprint arXiv:2606.25899*, 2026.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive” or “we were unable to find the license for the dataset we used”). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”,**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [TODO]

Justification: [TODO]

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [TODO]

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [TODO]

Justification: [TODO]

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer **[Yes]** if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

## 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

**15. Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

**16. Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.